Machine Learning-1

PROJECT REPORT

Table of Contents

Data Dictionary of Problem 1

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

Data Dictionary of Problem 2

Name	Description
State	State Code
District	District Code
Name	Name
TRU1	Area Name
No_HH	No of Household
TOT_M	Total population Male
TOT_F	Total population Female
M_06	Population in the age group 0-6 Male
F_06	Population in the age group 0-6 Female
M_SC	Scheduled Castes population Male
F_SC	Scheduled Castes population Female
M_ST	Scheduled Tribes population Male
F_ST	Scheduled Tribes population Female
M_LIT	Literates' population Male
F_LIT	Literates' population Female
M_ILL	Illiterate Male
F_ILL	Illiterate Female
TOT_WORK_M	Total Worker Population Male
TOT_WORK_F	Total Worker Population Female
MAINWORK_M	Main Working Population Male
MAINWORK_F	Main Working Population Female
MAIN_CL_M	Main Cultivator Population Male
MAIN_CL_F	Main Cultivator Population Female
MAIN_AL_M	Main Agricultural Labourers Population Male
MAIN_AL_F	Main Agricultural Labourers Population Female
MAIN_HH_M	Main Household Industries Population Male
MAIN_HH_F	Main Household Industries Population Female
MAIN_OT_M	Main Other Workers Population Male
MAIN_OT_F	Main Other Workers Population Female
MARGWORK_M	Marginal Worker Population Male
MARGWORK_F	Marginal Worker Population Female
MARG_CL_M	Marginal Cultivator Population Male
MARG_CL_F	Marginal Cultivator Population Female
MARG_AL_M	Marginal Agriculture Labourers Population Male
MARG_AL_F	Marginal Agriculture Labourers Population Female
MARG_HH_M	Marginal Household Industries Population Male

MARG_HH_F	Marginal Household Industries Population Female
MARG_OT_M	Marginal Other Workers Population Male
MARG_OT_F	Marginal Other Workers Population Female
MARGWORK_3_6_M	Marginal Worker Population 3-6 Male
MARGWORK_3_6_F	Marginal Worker Population 3-6 Female
MARG_CL_3_6_M	Marginal Cultivator Population 3-6 Male
MARG_CL_3_6_F	Marginal Cultivator Population 3-6 Female
MARG_AL_3_6_M	Marginal Agriculture Labourers Population 3-6 Male
MARG_AL_3_6_F	Marginal Agriculture Labourers Population 3-6 Female
MARG_HH_3_6_M	Marginal Household Industries Population 3-6 Male
MARG_HH_3_6_F	Marginal Household Industries Population 3-6 Female
MARG_OT_3_6_M	Marginal Other Workers Population Person 3-6 Male
MARG_OT_3_6_F	Marginal Other Workers Population Person 3-6 Female
MARGWORK_0_3_M	Marginal Worker Population 0-3 Male
MARGWORK_0_3_F	Marginal Worker Population 0-3 Female
MARG_CL_0_3_M	Marginal Cultivator Population 0-3 Male
MARG_CL_0_3_F	Marginal Cultivator Population 0-3 Female
MARG_AL_0_3_M	Marginal Agriculture Labourers Population 0-3 Male
MARG_AL_0_3_F	Marginal Agriculture Labourers Population 0-3 Female
MARG_HH_0_3_M	Marginal Household Industries Population 0-3 Male
MARG_HH_0_3_F	Marginal Household Industries Population 0-3 Female
MARG_OT_0_3_M	Marginal Other Workers Population 0-3 Male
MARG_OT_0_3_F	Marginal Other Workers Population 0-3 Female
NON_WORK_M	Non-Working Population Male
NON_WORK_F	Non-Working Population Female

Problem 1

Context

The digital advertising landscape is characterized by vast amounts of data generated from numerous ad campaigns across various platforms. In this complex environment, marketers face the challenge of efficiently utilizing budgets while enhancing the effectiveness of their advertising efforts. Traditional approaches often rely on broad targeting strategies, which may not efficiently utilize data insights to optimize ad performance.

Objective

This project aims to leverage clustering techniques to dissect a comprehensive dataset of online advertisements, aiming to:

Strategic Segmentation: Systematically segment advertisements based on a range of performance metrics and inherent characteristics to uncover distinct behavioral patterns. This segmentation aims to identify which features contribute most significantly to high-performing ads.

Resource Optimization: Use insights from data-driven clusters to optimize resource allocation. By understanding which ad characteristics correlate with success, marketing budgets can be more accurately targeted towards the most effective strategies.

Enhanced Campaign Strategies: Develop tailored advertising strategies for each identified cluster. This approach ensures that specific, effective tactics are applied to maximize engagement and conversion rates across different audience segments.

Problem 1 - Data Overview

The dataset comprises key performance indicators from online advertisements, including total spend, impressions, clicks, and derived metrics such as Cost per Mille (CPM), Cost per Click (CPC), and Click Through Rate (CTR). Initial actions included:

Solution of Data Overview:

1. We have imported all the required libraries such as,
   - Numpy
   - Pandas
   - Matplotlib
   - Seaborn
   - Scipy and
   - Warnings - to ignore the warning messages

2. After importing all the necessary modules, we then proceed with reading the dataset. The dataset is provided in the form of .xlsx file. So we've used pandas **read_excel ()** method to read the dataset and assigned it to the variable 'data'.
3. We used **shape** attribute of pandas library [data.shape] to determine the number of rows and columns of the dataframe df.

Shape	Description
(23066, 19)	We have 23066 rows and 19 columns in the dataset

Table 1 Shape of data

4. To check the types of data, we used **info()** method of pandas library to list the basic information of

the data such as rows count, columns count and datatype of the columns.

```
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     int64
Matched_Queries           int64
Impressions               int64
Clicks                    int64
Spend                   float64
Fee                     float64
Revenue                 float64
CTR                     float64
CPM                     float64
CPC                     float64
dtype: object
```

Table 2 Basic Information of data

- From the above table, We have **6** object, **7** int and **6** float data types in the dataset. Also we can see that there are **53** and **106** null values in Gender and Partner_salary columns respectively.

Columns	Data Type	No of Missing Values
CTR	float	4736
CPM	float	4736
CTC	float	4736

Table 3 Null values count

5. The data type of CTR, CPM, CTC column is float. So we will impute the missing values using the formula given.

    - CTR = Total Measured Clicks / Total Measured Ad Impressions x 100. Note that the Total Measured Clicks refers to the 'Clicks' Column and the Total Measured Ad Impressions refers to the 'Impressions' Column.

    - CPM = (Total Campaign Spend / Number of Impressions) * 1,000. Note that the Total Campaign Spend refers to the 'Spend' Column and the Number of Impressions refers to the 'Impressions' Column.

    - CPC = Total Cost (spend) / Number of Clicks. Note that the Total Cost (spend) refers to the 'Spend' Column and the Number of Clicks refers to the 'Clicks' Column.

```
Unique values in each column:
 Timestamp                2018
InventoryType                7
Ad - Length                  6
Ad- Width                    5
Ad Size                      7
Ad Type                     14
Platform                     3
Device Type                  2
Format                       2
Available_Impressions    21560
Matched_Queries          20919
Impressions              20405
Clicks                   12752
Spend                    20467
Fee                          7
Revenue                  20578
CTR                       2066
CPM                       2084
CPC                        194
dtype: int64
```

Table 4 Unique values of Categorical Columns

Observations and Insights:

- Demographic Distribution and Variability: The dataset reveals a wide variance in household numbers, with an average of 51,223 households per district but a high standard deviation, suggesting pronounced disparities between districts in terms of population density.

- Females outnumber males on average, with total female populations considerably higher than males in the districts surveyed. This could imply a demographic skew that might impact social services and resource allocation.

- Child Population: There is a slight male bias in the child population aged 0-6 years, which might have implications for early childhood education and health services targeting young children.

- Socio-economic Status of Scheduled Castes and Tribes: The data shows that more females are categorized as belonging to the Scheduled Castes than males, which might indicate socio-economic challenges particular to women in these communities.

- The relatively lower figures for the Scheduled Tribes suggest that certain districts may have smaller indigenous populations, which could affect the focus and funding of tribal development programs.

- Labor and Employment: Marginal workers, particularly female casual laborers, are notably higher in number than their male counterparts, highlighting a gender disparity in economic vulnerability. This suggests that women in these districts may predominantly engage in less stable and lower-paid jobs.

- Non-working population figures further underscore a gender gap in employment, with significantly more females categorized as non-working compared to males, pointing towards potential barriers to female participation in the workforce or cultural norms affecting women's employment.

- Economic Implications for Policy:Given the high counts of marginal and non-working individuals, especially among women, there's a clear indication of economic underutilization and vulnerability. This scenario calls for targeted economic and educational interventions aimed at enhancing employment opportunities for these groups.The variability in demographic and economic data across districts necessitates localized approaches to policy-making, where interventions are tailored to meet specific regional needs based on the unique socio-economic profiles of each district. These insights provide a foundation for further

Problem 1 - Univariate Analysis

Explore all the variables (categorical and numerical) in the data - Check for and treat (if needed) outliers - Observations and Insights.

Solution of Univariate Analysis:

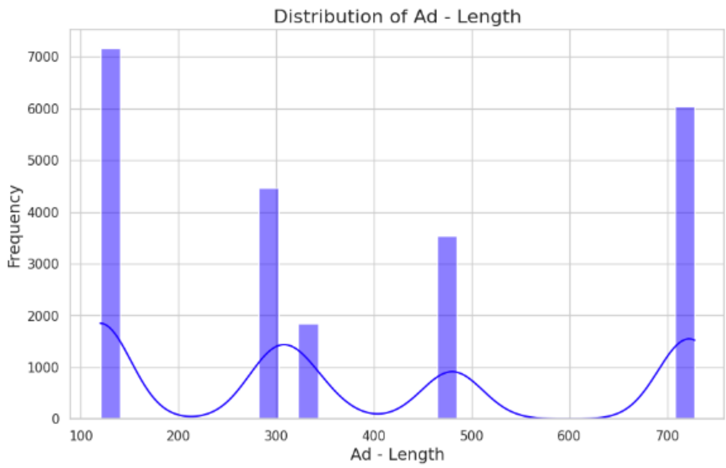
1. First, we selected the numerical and categorical columns from the dataframe and assigned them to the varables '**numerical_cols**' and '**categorical_cols**' using select_dtypes().

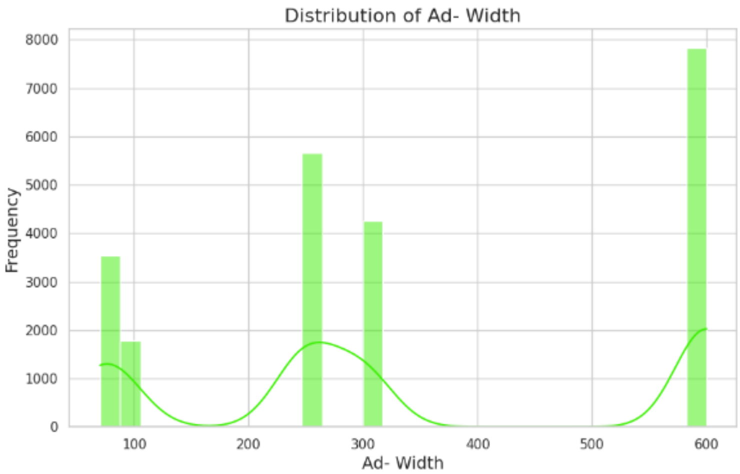
Variable name	Columns
numerical_cols	'Ad - Length', 'Ad- Width', 'Ad Size', 'Available_Impressions', 'Matched_Queries', 'Impressions', 'Clicks', 'Spend', 'Fee', 'Revenue','CTR', 'CPM', 'CPC'.
categorical_cols	'Timestamp', 'InventoryType', 'Ad Type', 'Platform', 'Device Type','Format'

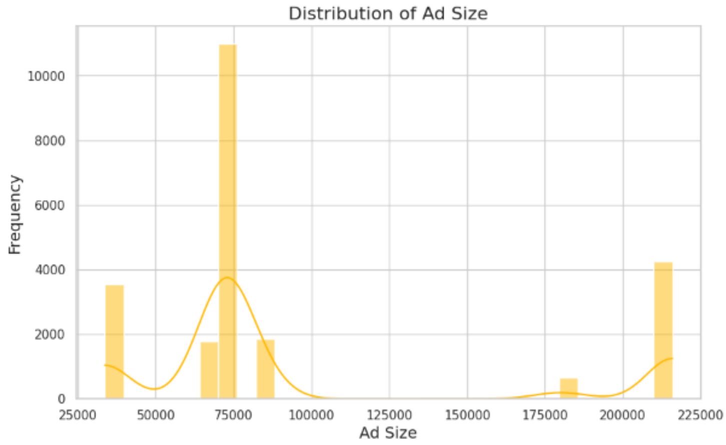
Table 5 Numerical & Categorical Columns

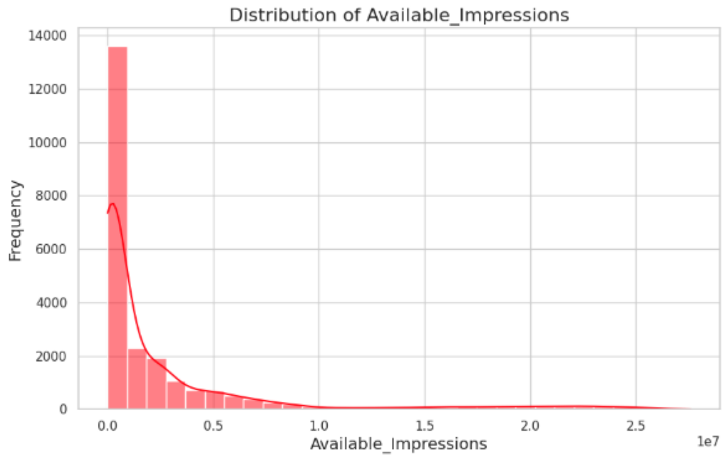
2. After separating the numerical and categorical columns, we created Histplot for Numeric and Countplot for categoricalcolumns with **KDE** value as True using searborn library.

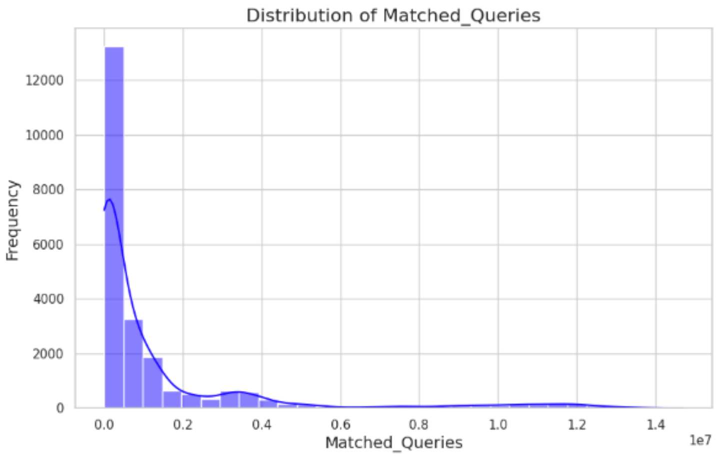
Observation on numerical columns :

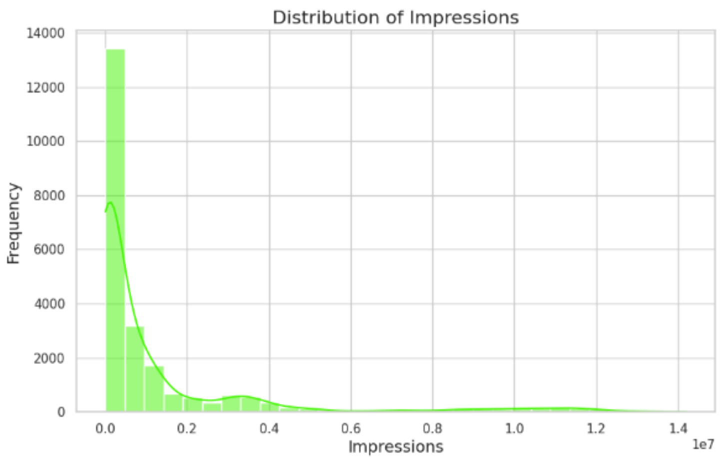


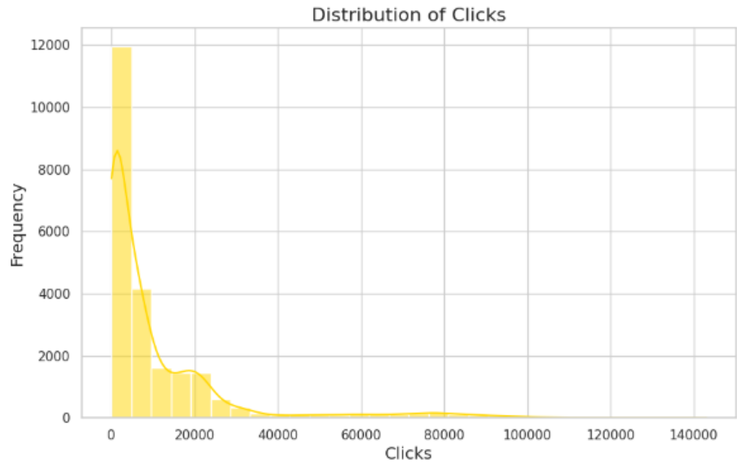


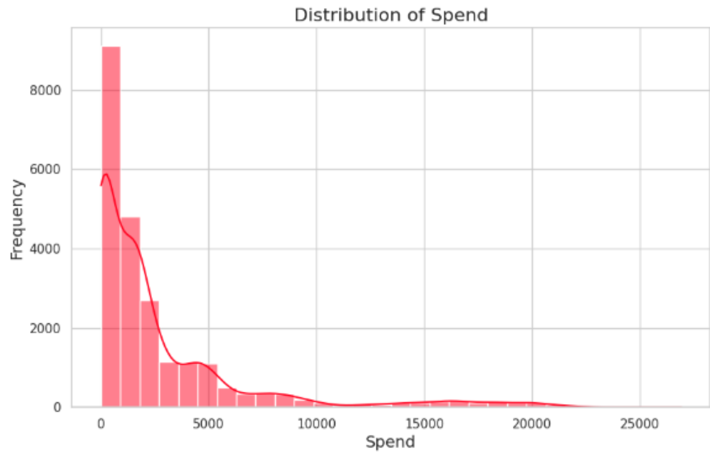

Distribution of Ad Size
Frequency
Ad Size
0
2000
4000
6000
8000
10000
25000
50000
75000
100000
125000
150000
175000
200000
225000

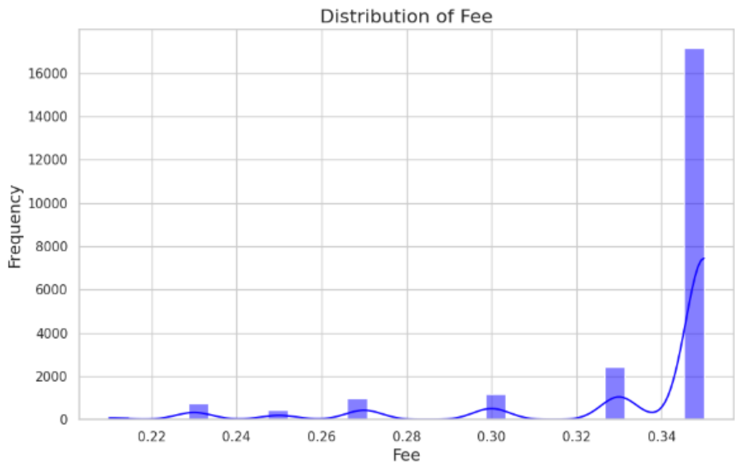

Distribution of Available_Impressions
Frequency
Available_Impressions
1e7
0
2000
4000
6000
8000
10000
12000
14000
0.0
0.5
1.0
1.5
2.0
2.5

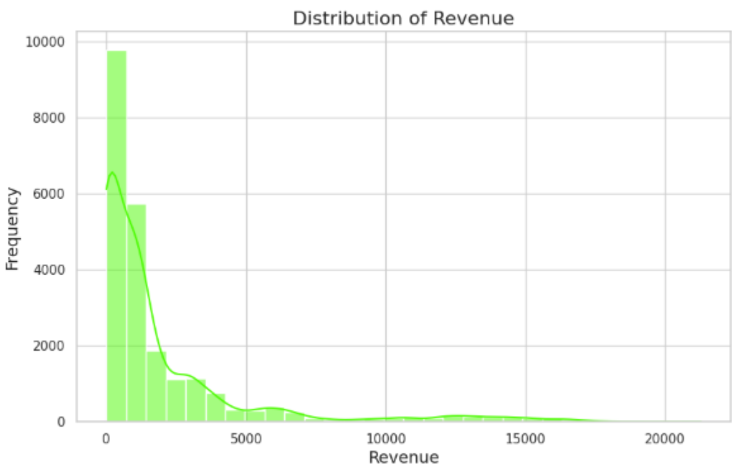

Distribution of Matched_Queries
Frequency
Matched_Queries
1e7
0
2000
4000
6000
8000
10000
12000
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4


Distribution of Impressions
Frequency
Impressions
1e7
0
2000
4000
6000
8000
10000
12000
14000
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4


Distribution of Clicks
Frequency
Clicks
0
2000
4000
6000
8000
10000
12000
20000
40000
60000
80000
100000
120000
140000


Distribution of Spend
Frequency
Spend
0
2000
4000
6000
8000
5000
10000
15000
20000
25000


Distribution of Fee
Frequency
Fee
0
2000
4000
6000
8000
10000
12000
14000
16000
0.22
0.24
0.26
0.28
0.30
0.32
0.34


Distribution of Revenue
Frequency
Revenue
0
2000
4000
6000
8000
10000
5000
10000
15000
20000

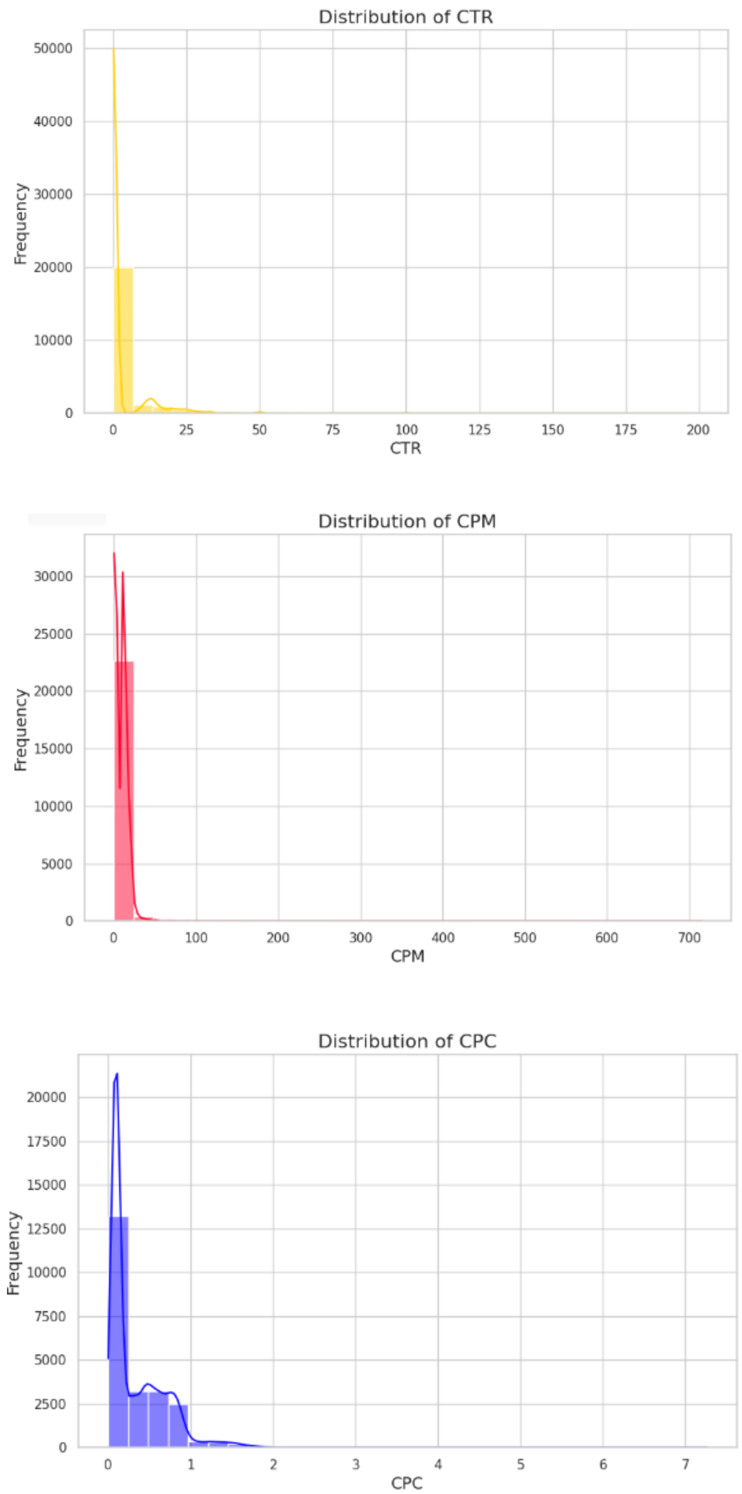
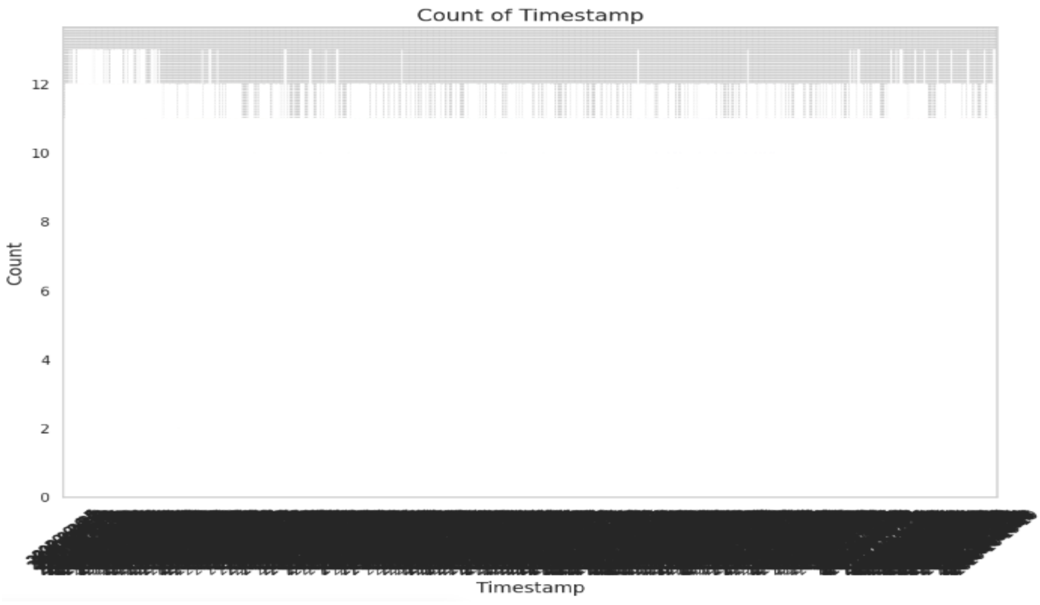


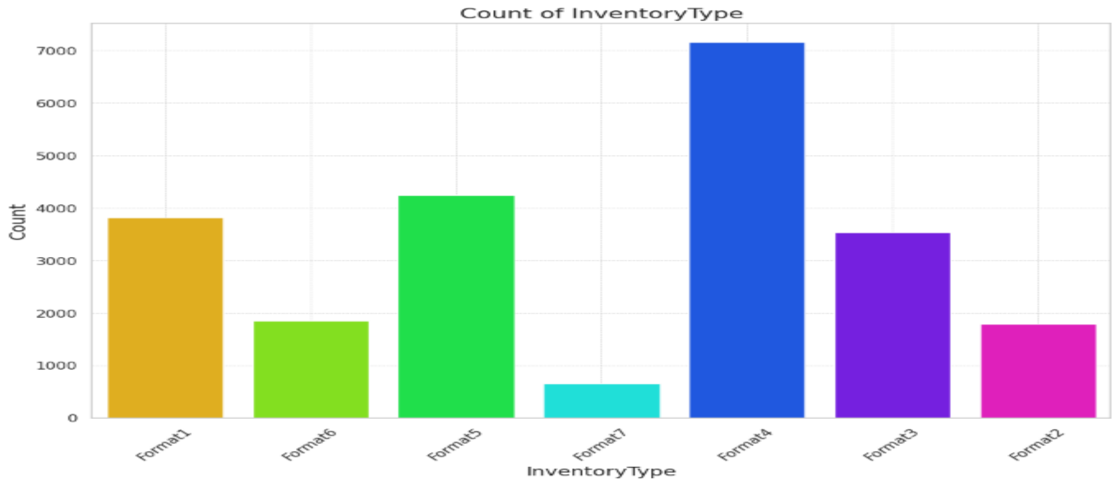
Figure 1 HistPlot of Numerical Columns

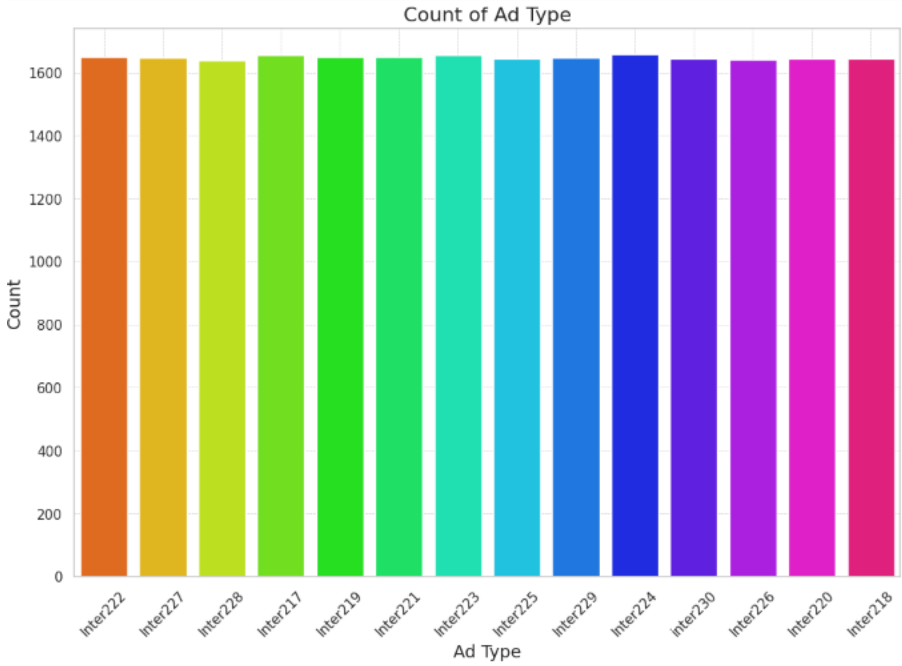
Observation:

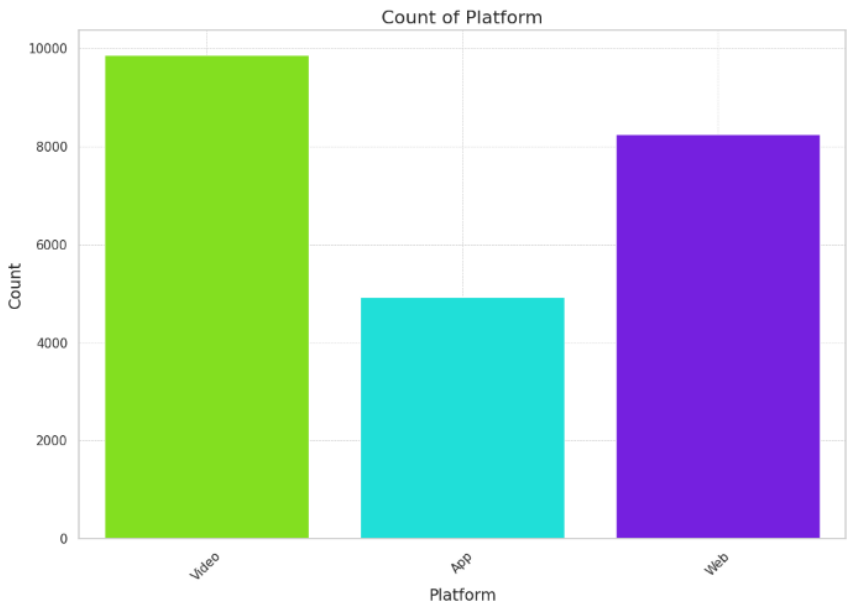
- Ad Length and Ad Width: Both distributions are multimodal, suggesting preferred standard ad sizes that are commonly used in the industry.
- Ad Size: Shows peaks at certain sizes, indicating common ad size configurations.
- Available Impressions: Highly right-skewed, suggesting that most ads receive a low number of impressions, with a few outliers receiving very high impressions.
- Matched Queries: Also right-skewed, indicating that while most ads match a smaller number of queries, a few are highly matched.
- Impressions: Similar to Available Impressions, most ads get fewer impressions, with a few achieving high visibility.
- Clicks: Rapidly declining frequency as clicks increase, typical for ad campaigns where many ads receive few clicks.
- Spend: Shows a decline similar to clicks, indicating varying budget levels with most ads having lower spend.
- Fee: Bimodal distribution suggesting two common fee structures or rates within the platform.
- Revenue: Right-skewed like many other metrics, where most ads generate lower revenue and a few are highly profitable.
- CTR (Click Through Rate): The distribution of CTR shows a heavy concentration at lower values, indicating that most ads have a very low click-through rate, with few ads achieving higher rates. This is typical in digital advertising, where high CTRs are hard to achieve.
- CPM (Cost Per Mille): The CPM graph also shows a strong skew towards the lower end, suggesting that most ads are associated with lower costs per thousand impressions. There's a steep drop-off after the initial peak, indicating that very high CPMs are rare.
- CPC (Cost Per Click): The distribution of CPC highlights that the majority of ads cost very little per click, with a long tail extending to higher costs. This suggests that while it's common to have low CPCs, certain ads or targeted campaigns can result in significantly higher costs per click.

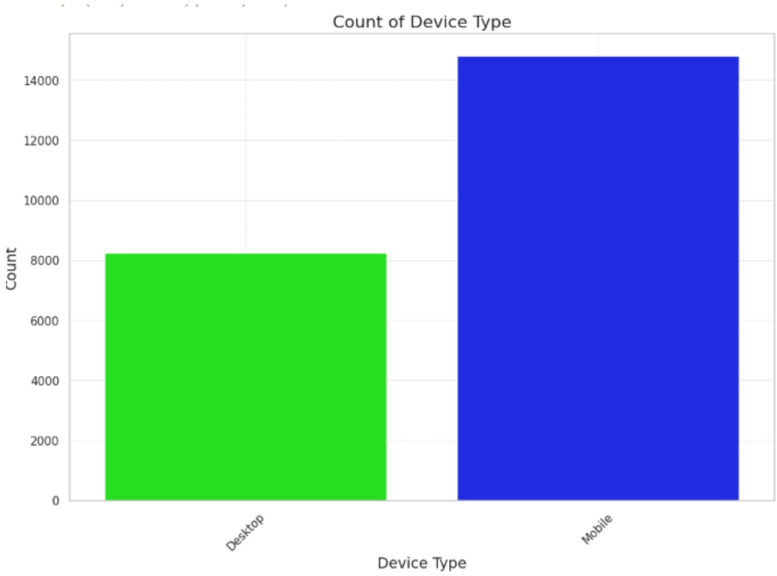
Observations On categorical columns:




Count of InventoryType
Count
InventoryType
Format1
Format6
Format5
Format7
Format4
Format3
Format2
0
1000
2000
3000
4000
5000
6000
7000


Count of Ad Type
Count
Ad Type
Inter222
Inter227
Inter228
Inter217
Inter219
Inter221
Inter223
Inter225
Inter229
Inter224
inter230
Inter226
Inter220
Inter218
0
200
400
600
800
1000
1200
1400
1600


Count of Platform
Count
Platform
Video
App
Web
0
2000
4000
6000
8000
10000



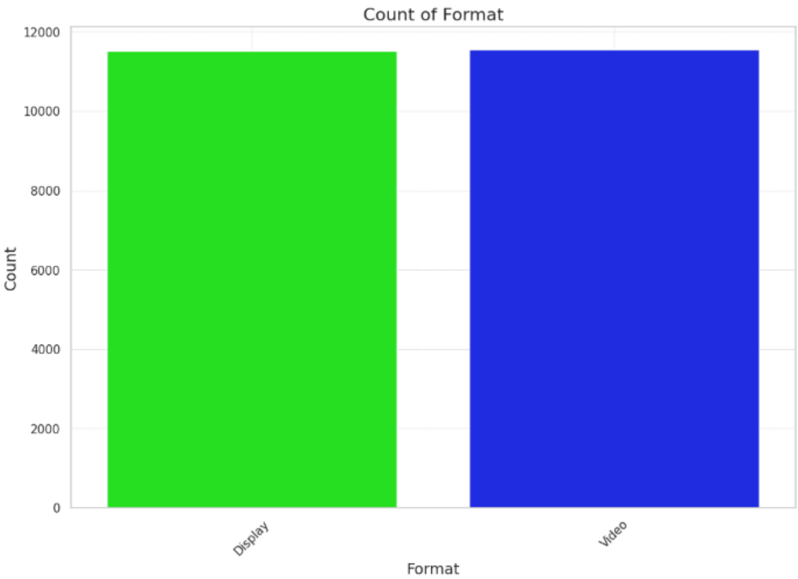


Figure 2 Count Plot of Categorical Columns

Observations:

3. Timestamp Distribution:The histogram shows that the frequency of timestamps is heavily concentrated at certain intervals, suggesting that some specific times are significantly more common for ads to be posted or active.
4. Inventory Type: The inventory type chart shows a significant variance in the counts of different types, with one category (Format7) being much more prevalent than others. This could indicate a preferred or more effective inventory type within the ad campaigns.
5. Ad Type: The distribution of ad types is relatively uniform, suggesting a diverse strategy where various ad types are used almost equally. This might be part of a broader approach to test different ad types or to cater to varied audience preferences.
6. Platform: The platform chart shows a higher count for web and video platforms compared to apps. This indicates that these platforms might be more popular or more effective for advertising purposes.
7. Device Type: Mobile devices dominate over desktops, which reflects the ongoing trend in advertising that targets mobile users due to the increasing use of smartphones for internet access.
8. Ad Format: Video formats outnumber display formats, suggesting that video might be the more engaging and preferred format for advertisements aiming to capture audience attention more effectively.

Observations and Insights:

Spend vs. Impressions:

There is a positive relationship between Spend and Impressions, indicating that as spend increases, the number of impressions generally increases. However, there's significant variability, suggesting not all high spends yield high impressions efficiently.

CTR vs. Clicks:

The relationship between Clicks and CTR is not straightforward. Generally, ads with fewer clicks can have a very high or very low CTR, indicating the variability in effectiveness. As clicks increase, the range of CTR tends to narrow, possibly due to averaging effects over larger impression bases.

Spend vs. CPC:

Spend and CPC show a weak relationship, with many ads having low spend and low CPC, but as spend increases, CPC can vary widely. This suggests that the cost-efficiency of clicks isn't consistently related to the amount spent.

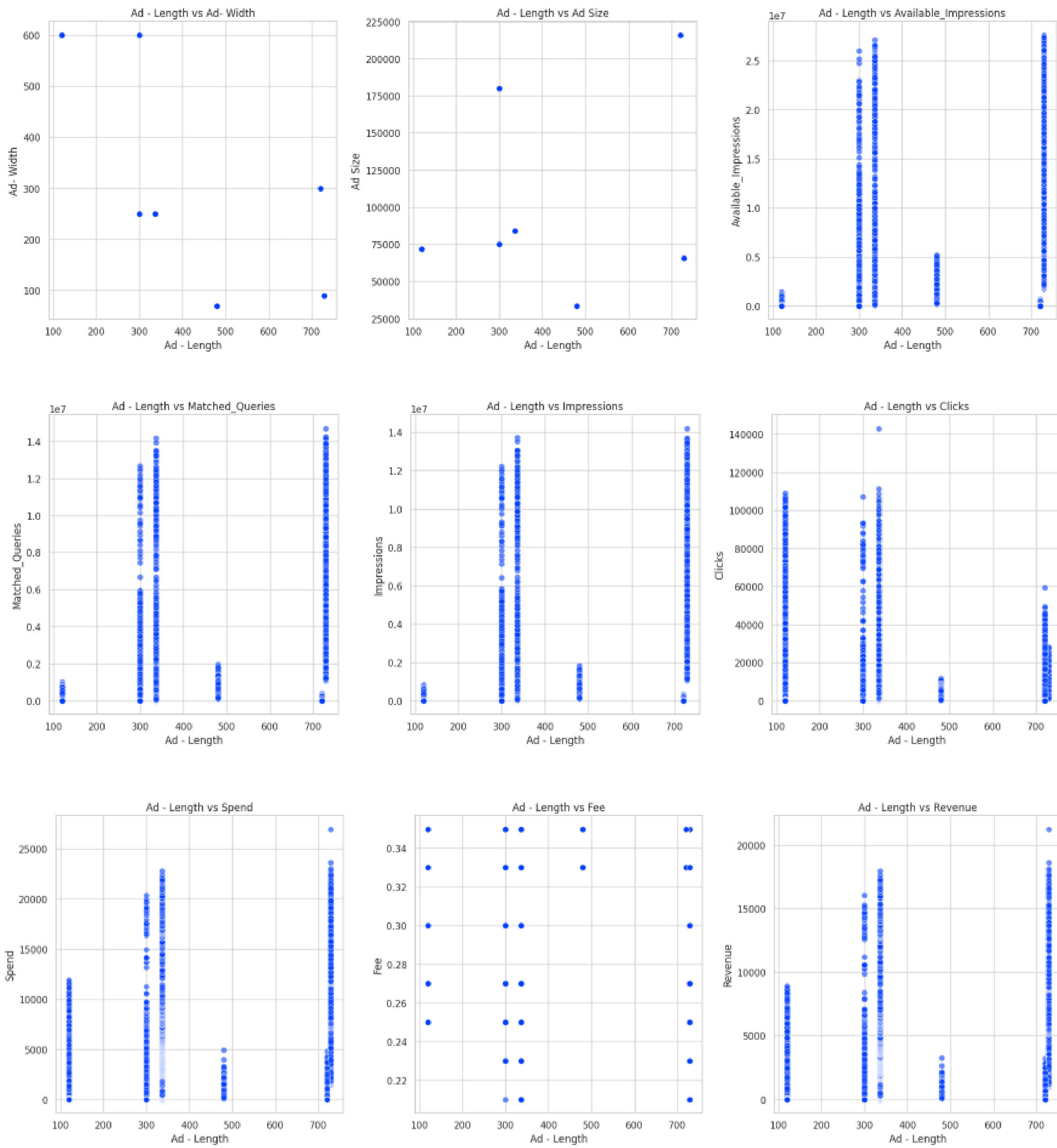
Problem 1 - Bivariate Analysis

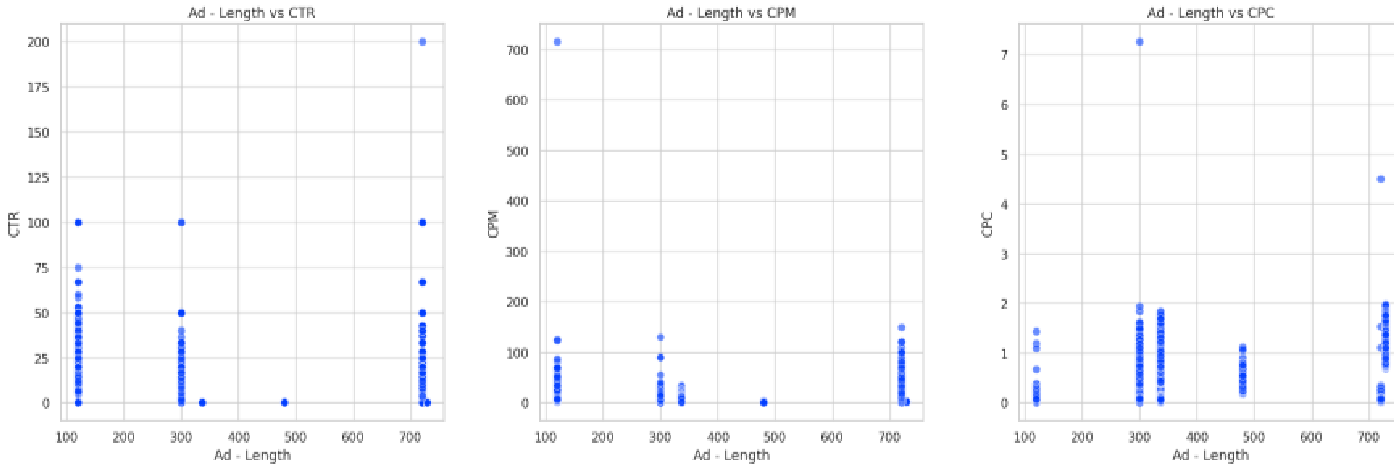
Explore the relationship between all numerical variables - Explore the correlation between all numerical variables - Explore the relationship between categorical vs numerical variables

Solution of Bivariate Analysis:

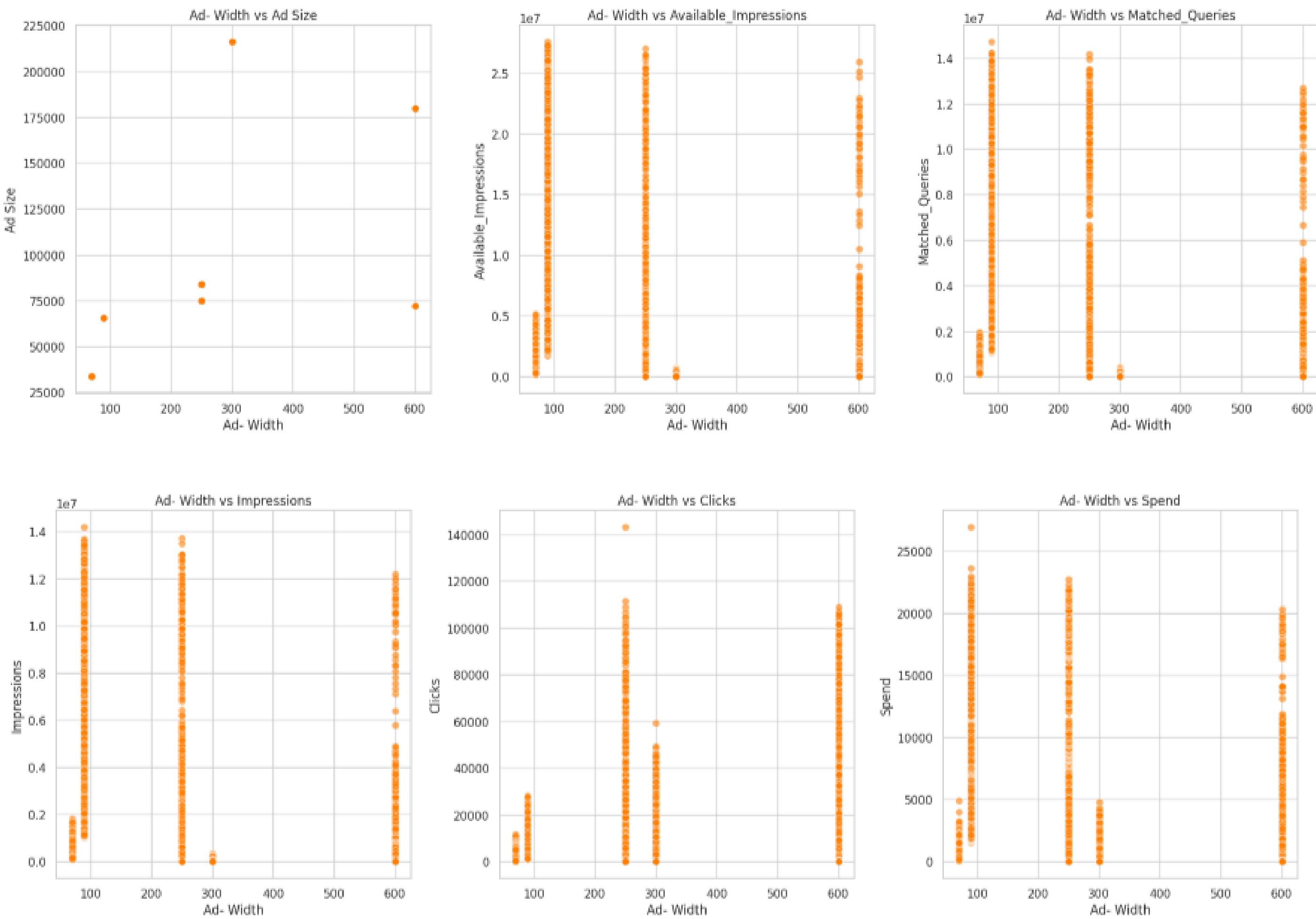
Numerical VS Numerical

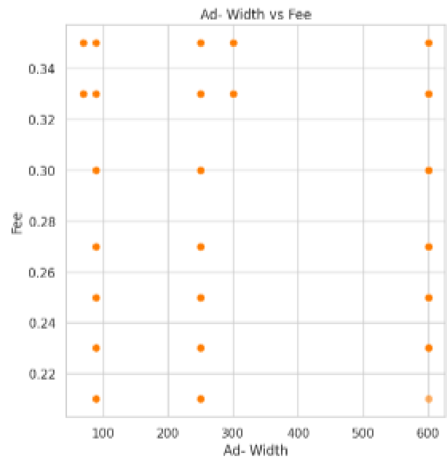
1. **Ad-Length:**

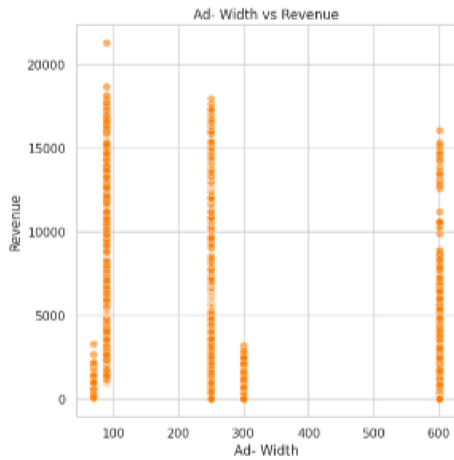


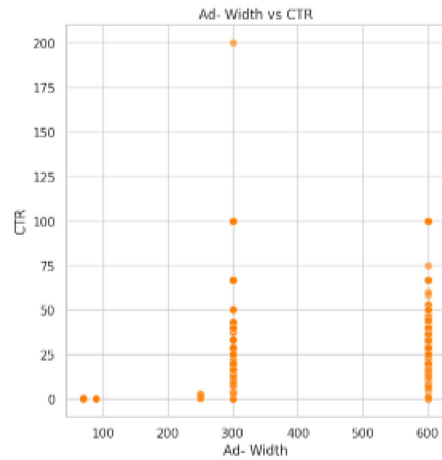

Ad - Length vs CTR
CTR
Ad - Length
Ad - Length vs CPM
CPM
Ad - Length
Ad - Length vs CPC
CPC
Ad - Length

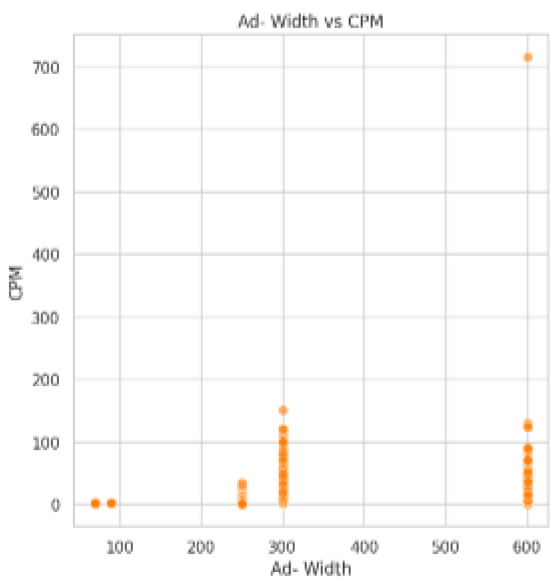
2. Ad-Width:

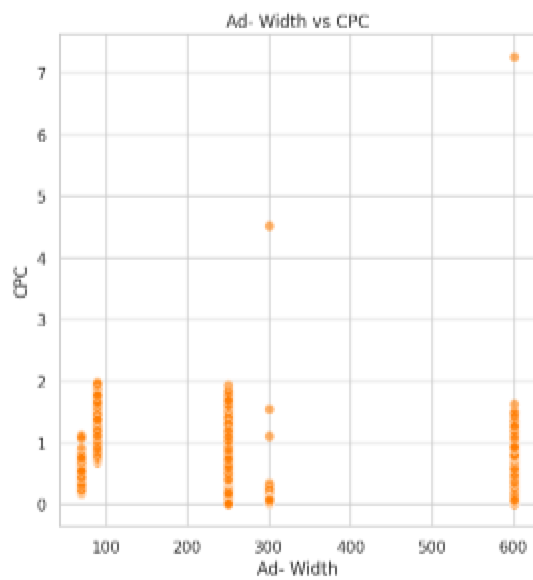

Ad- Width vs Ad Size
Ad Size
Ad- Width
Ad- Width vs Available_Impressions
1e7
Available_Impressions
Ad- Width
Ad- Width vs Matched_Queries
1e7
Matched_Queries
Ad- Width
Ad- Width vs Impressions
1e7
Impressions
Ad- Width
Ad- Width vs Clicks
Clicks
Ad- Width
Ad- Width vs Spend
Spend
Ad- Width


Ad- Width vs Fee
Fee
Ad- Width

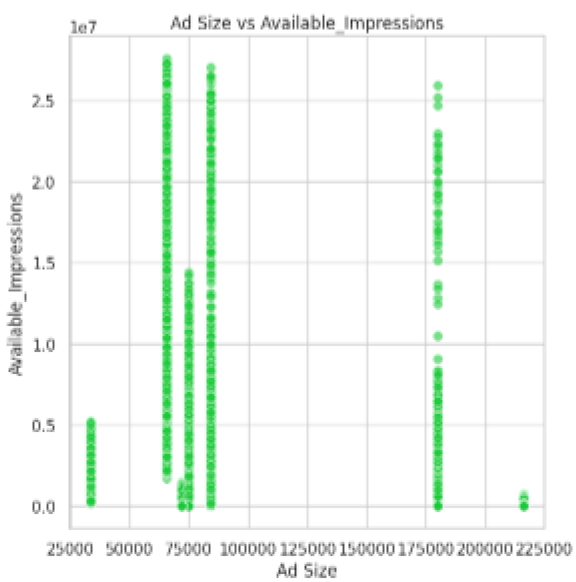

Ad- Width vs Revenue
Revenue
Ad- Width

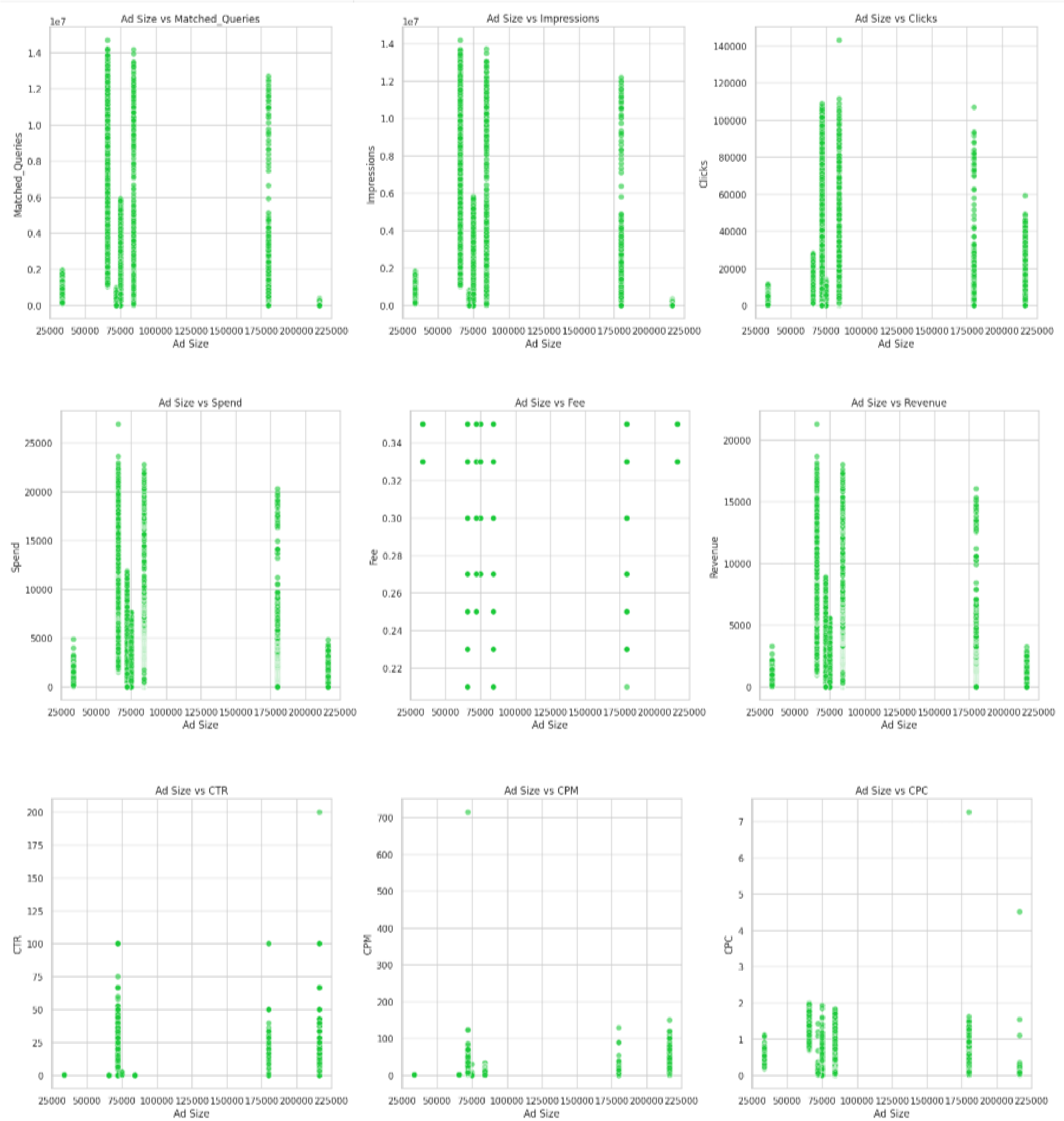

Ad- Width vs CTR
CTR
Ad- Width


Ad- Width vs CPM
CPM
Ad- Width

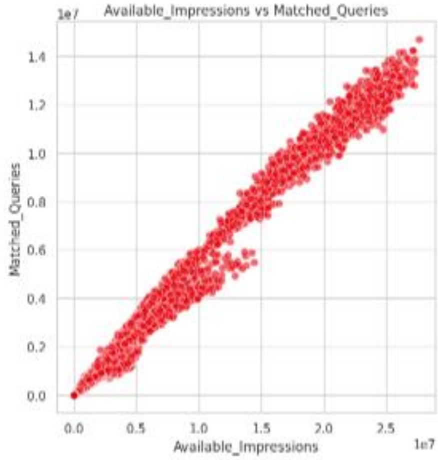

Ad- Width vs CPC
CPC
Ad- Width

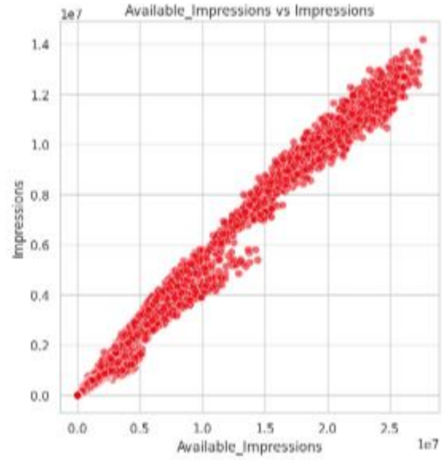
3. Ad-Size:

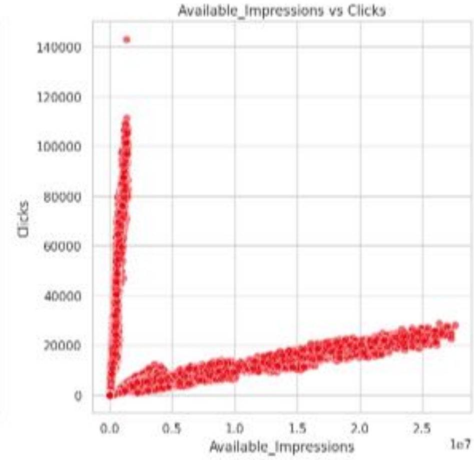

Ad Size vs Available_Impressions
1e7
Available_Impressions
Ad Size

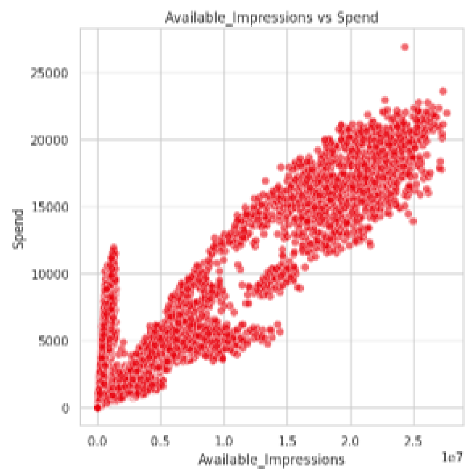


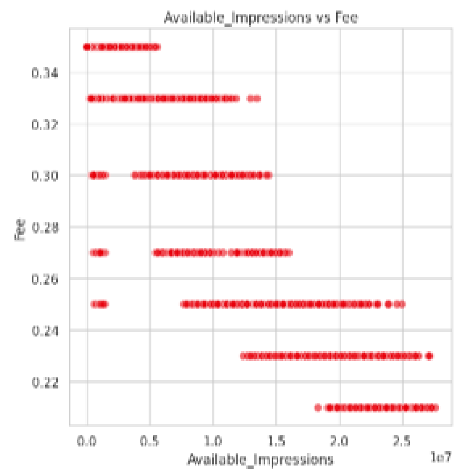
4. Available Impression:

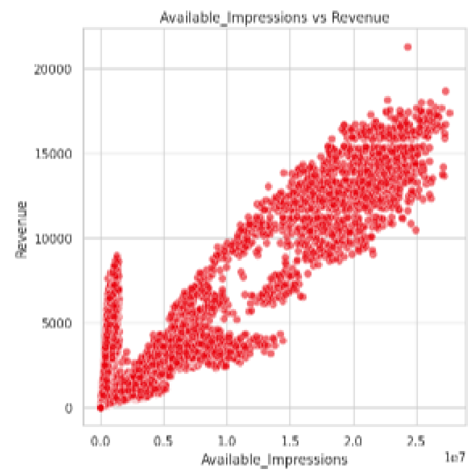


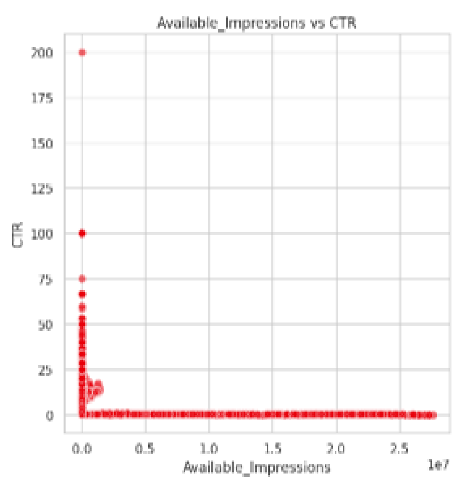


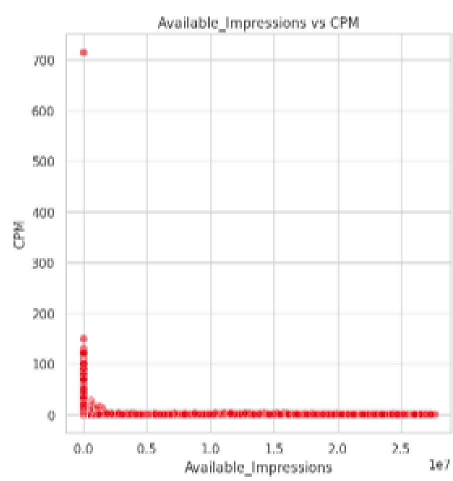


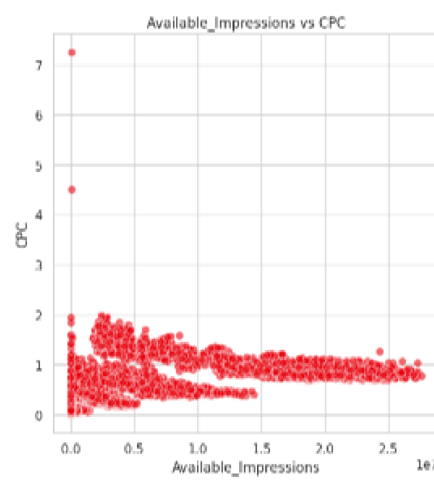




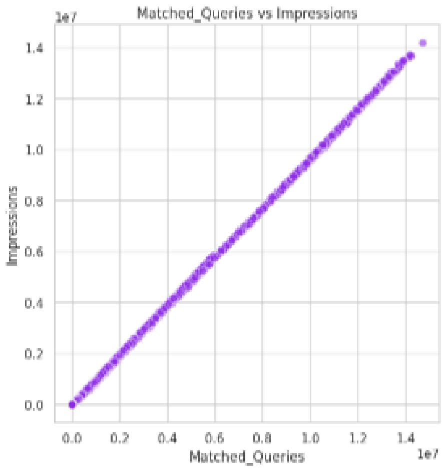


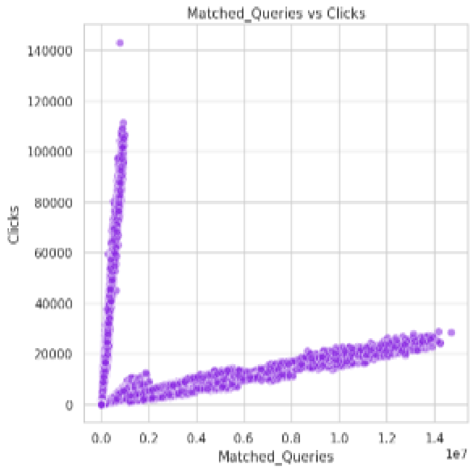


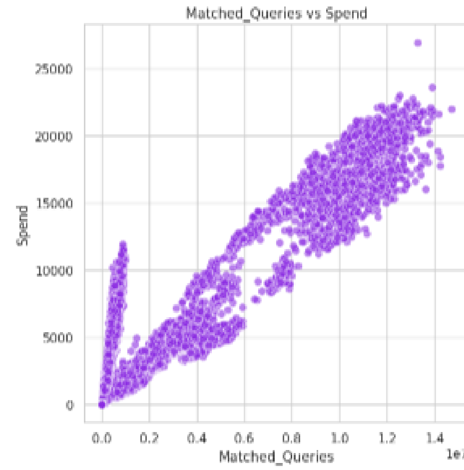


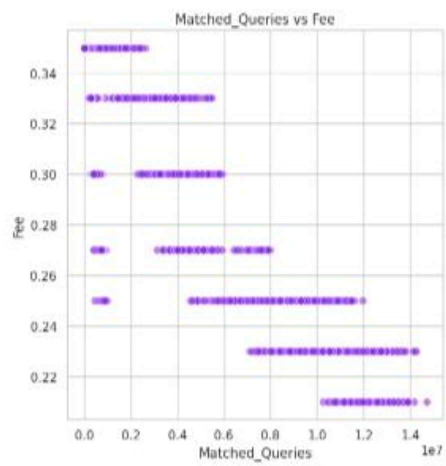


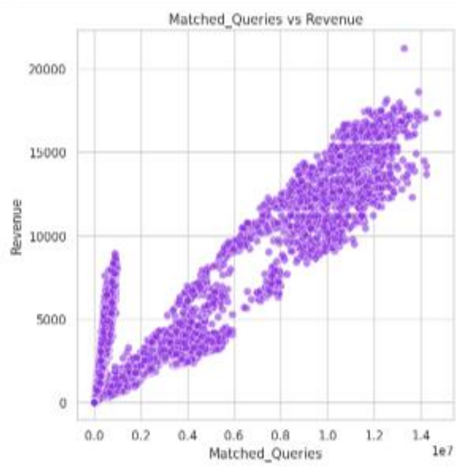
5. Matched_Queries:

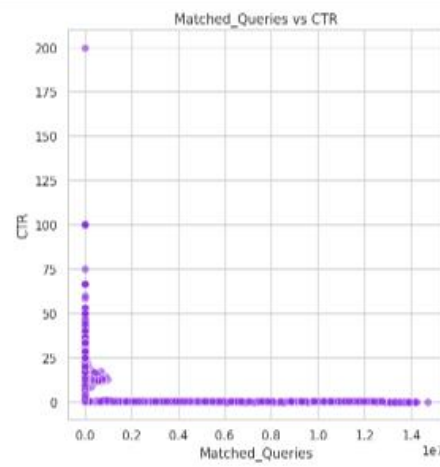


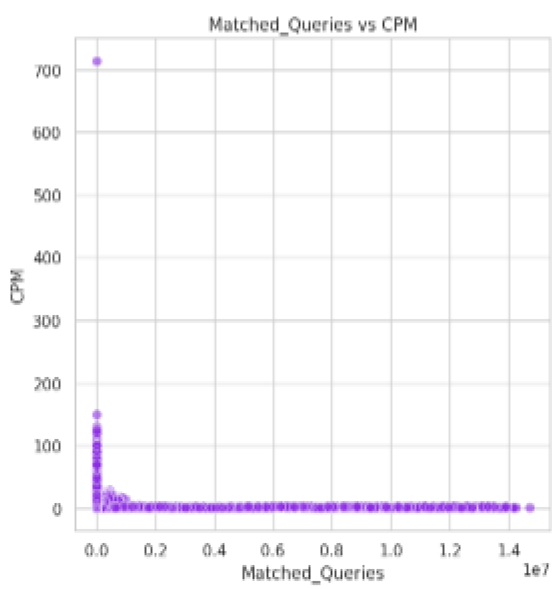


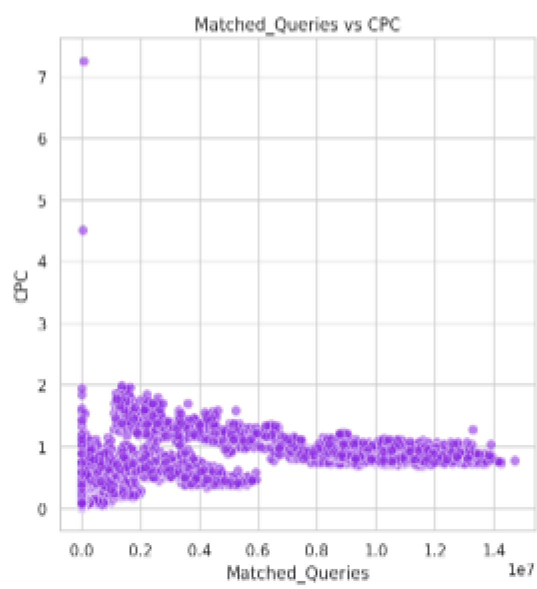




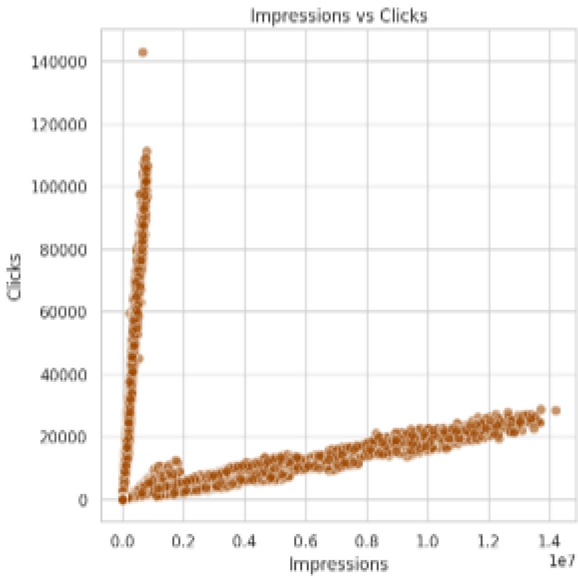


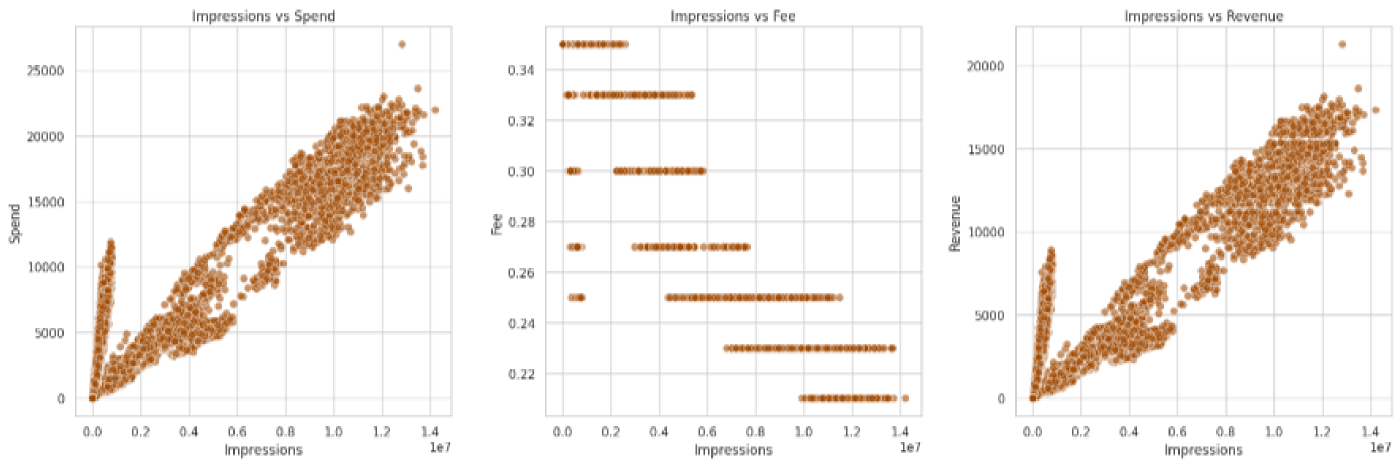


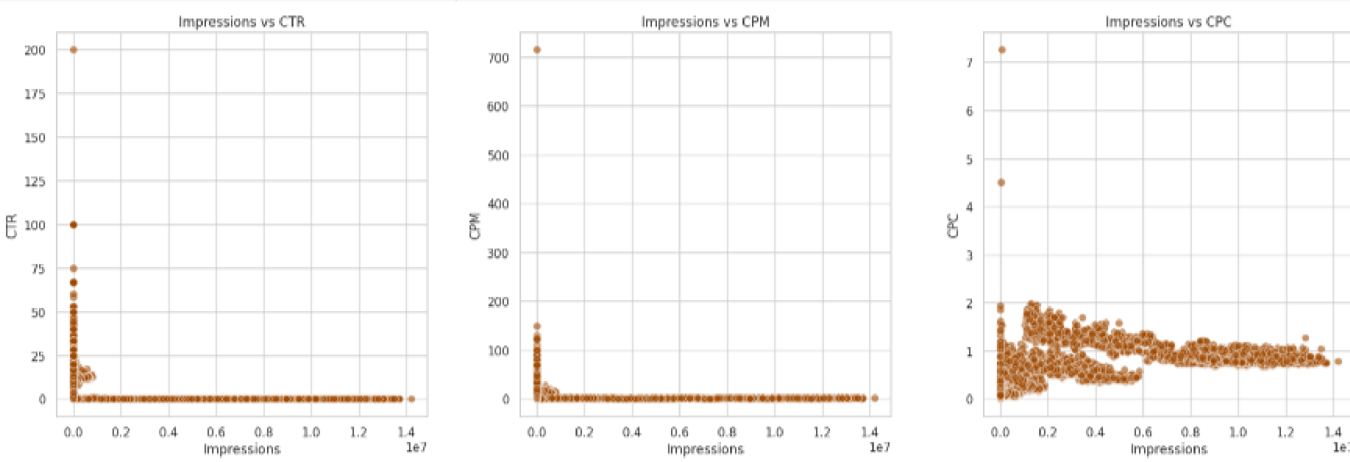




6. Impression:

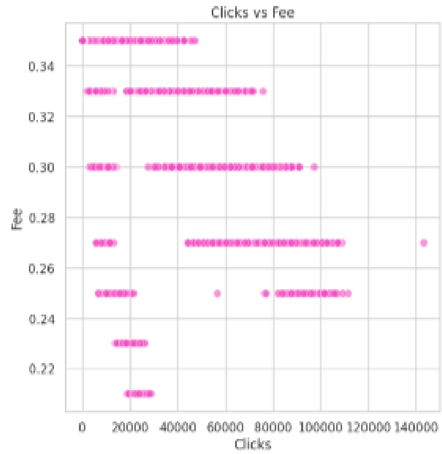

Impressions vs Clicks
Clicks
Impressions
1e7

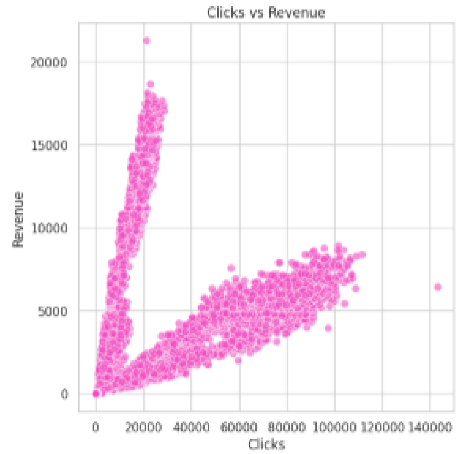

Impressions vs Spend
Spend
Impressions
Impressions vs Fee
Fee
Impressions
Impressions vs Revenue
Revenue
Impressions

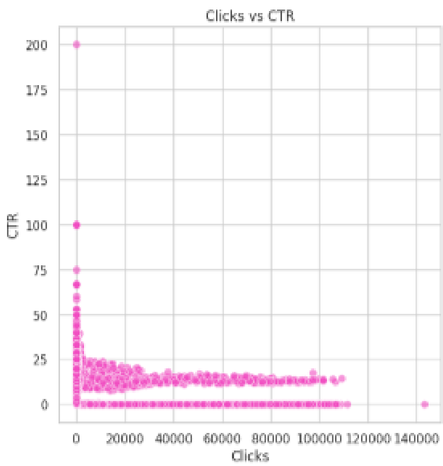

Impressions vs CTR
CTR
Impressions
Impressions vs CPM
CPM
Impressions
Impressions vs CPC
CPC
Impressions

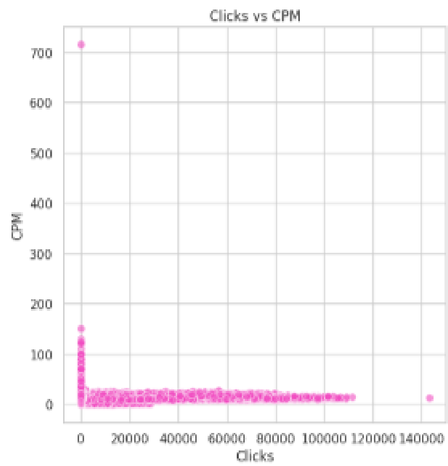
7. Clicks:

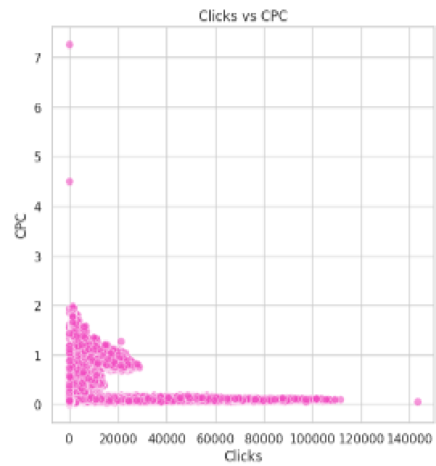

Clicks vs Spend
Spend
Clicks
25000
20000
15000
10000
5000
0
0 20000 40000 60000 80000 100000 120000 140000


Clicks vs Fee
Fee
Clicks
0.34
0.32
0.30
0.28
0.26
0.24
0.22
0 20000 40000 60000 80000 100000 120000 140000

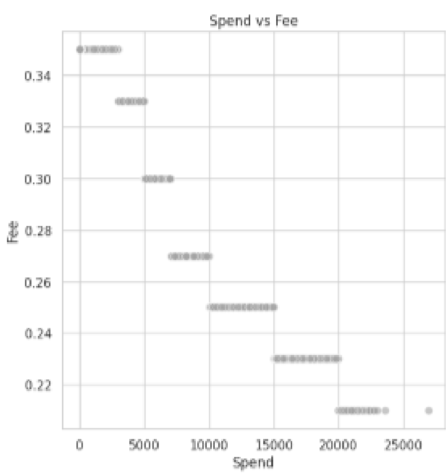

Clicks vs Revenue
Revenue
Clicks
20000
15000
10000
5000
0
0 20000 40000 60000 80000 100000 120000 140000

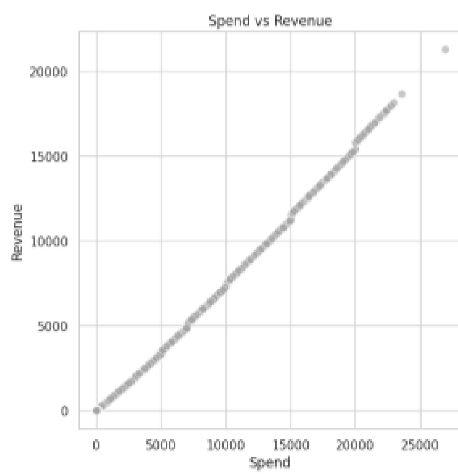

Clicks vs CTR
CTR
Clicks
200
175
150
125
100
75
50
25
0
0 20000 40000 60000 80000 100000 120000 140000

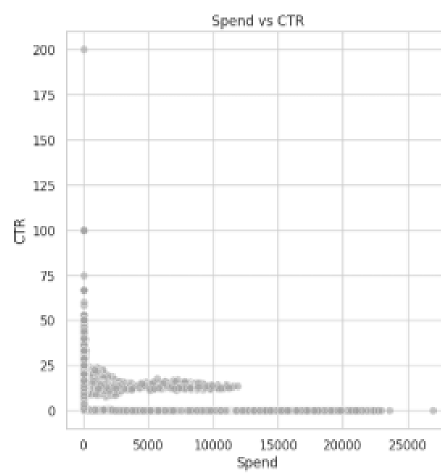

Clicks vs CPM
CPM
Clicks
700
600
500
400
300
200
100
0
0 20000 40000 60000 80000 100000 120000 140000

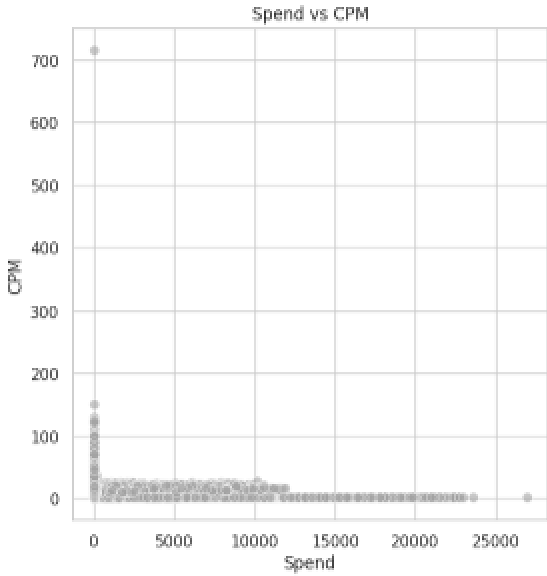

Clicks vs CPC
CPC
Clicks
7
6
5
4
3
2
1
0
0 20000 40000 60000 80000 100000 120000 140000

8. Spend:

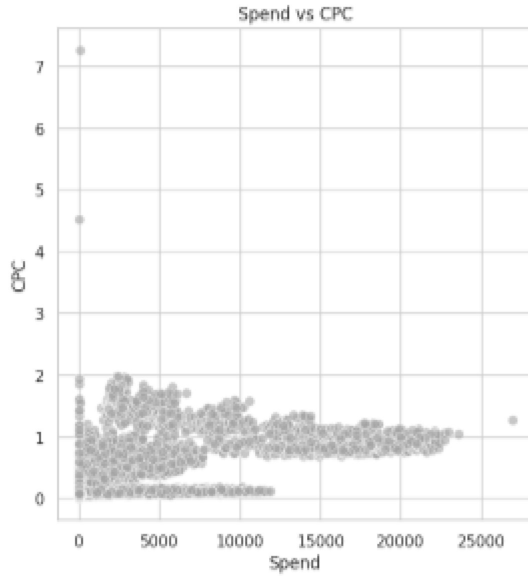

Spend vs Fee
Fee
Spend
0.34
0.32
0.30
0.28
0.26
0.24
0.22
0 5000 10000 15000 20000 25000


Spend vs Revenue
Revenue
Spend
20000
15000
10000
5000
0
0 5000 10000 15000 20000 25000


Spend vs CTR
CTR
Spend
200
175
150
125
100
75
50
25
0
0 5000 10000 15000 20000 25000



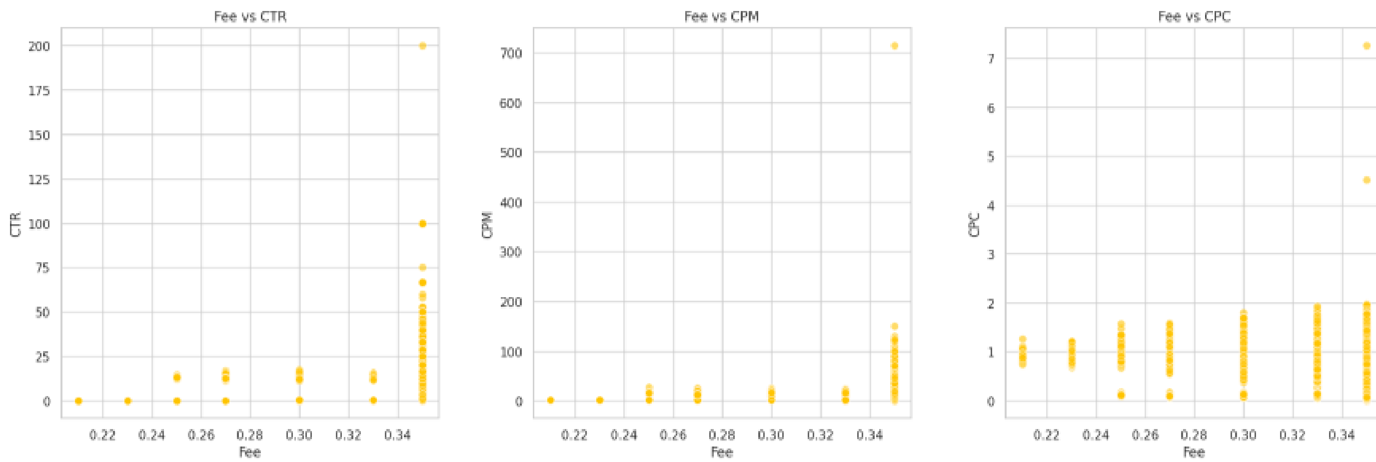
Spend vs CPM
CPM
Spend
0
100
200
300
400
500
600
700
0
5000
10000
15000
20000
25000



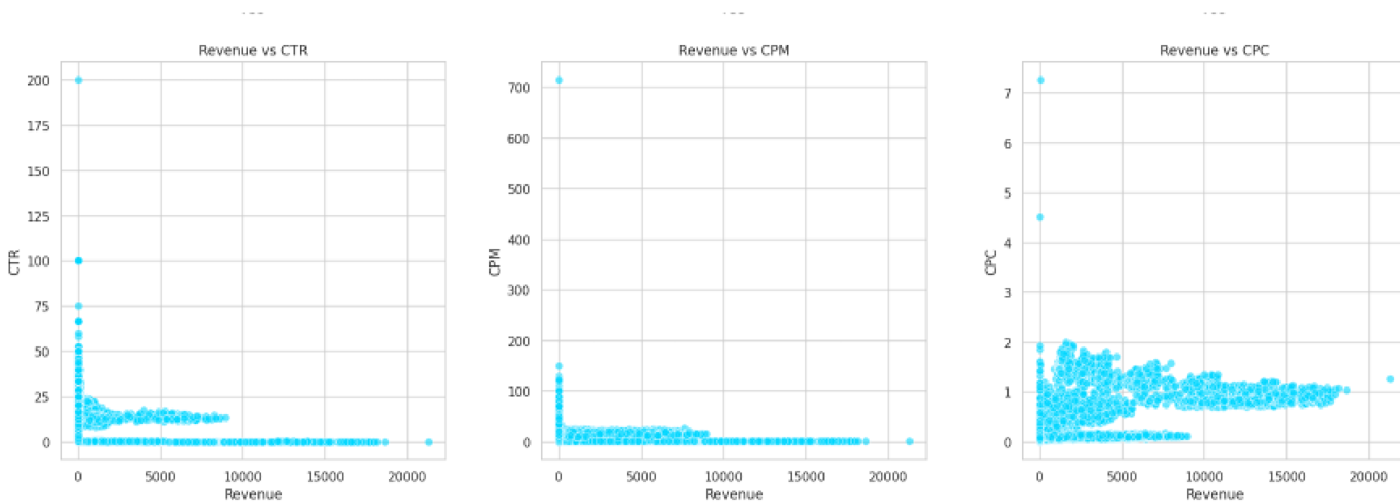
Spend vs CPC
CPC
Spend
0
1
2
3
4
5
6
7
0
5000
10000
15000
20000
25000

9. Fee:

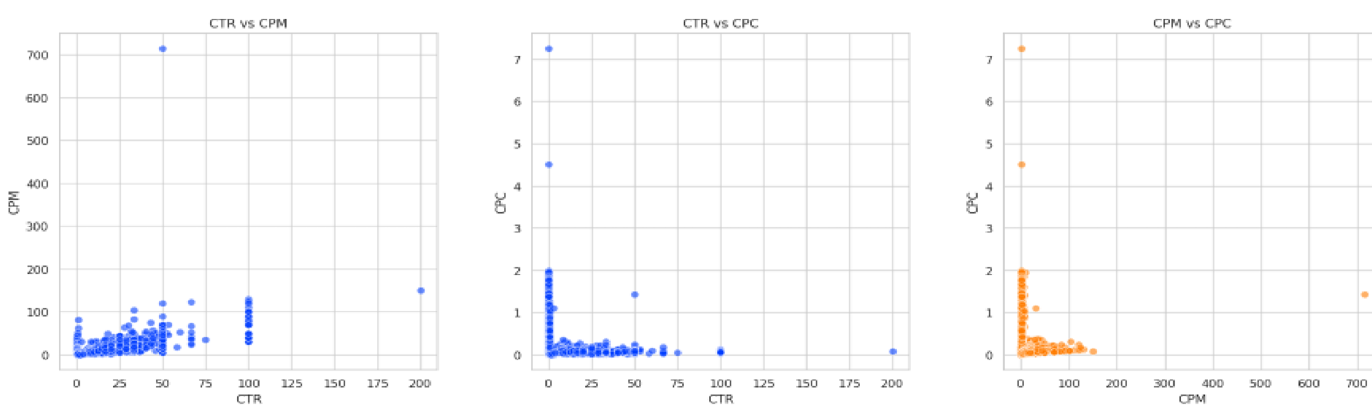
Fee vs Revenue
Revenue
Fee
0
5000
10000
15000
20000
0.22
0.24
0.26
0.28
0.30
0.32
0.34



10. Revenue:



11. CT



R:

Figure 4 Pairplot showing Relationship between Numerical Variables

Scatter Plots: These are prominent in the dataset where two numerical variables are plotted against each other. Each point represents an observation in the dataset with its position determined by the values of the two variables.

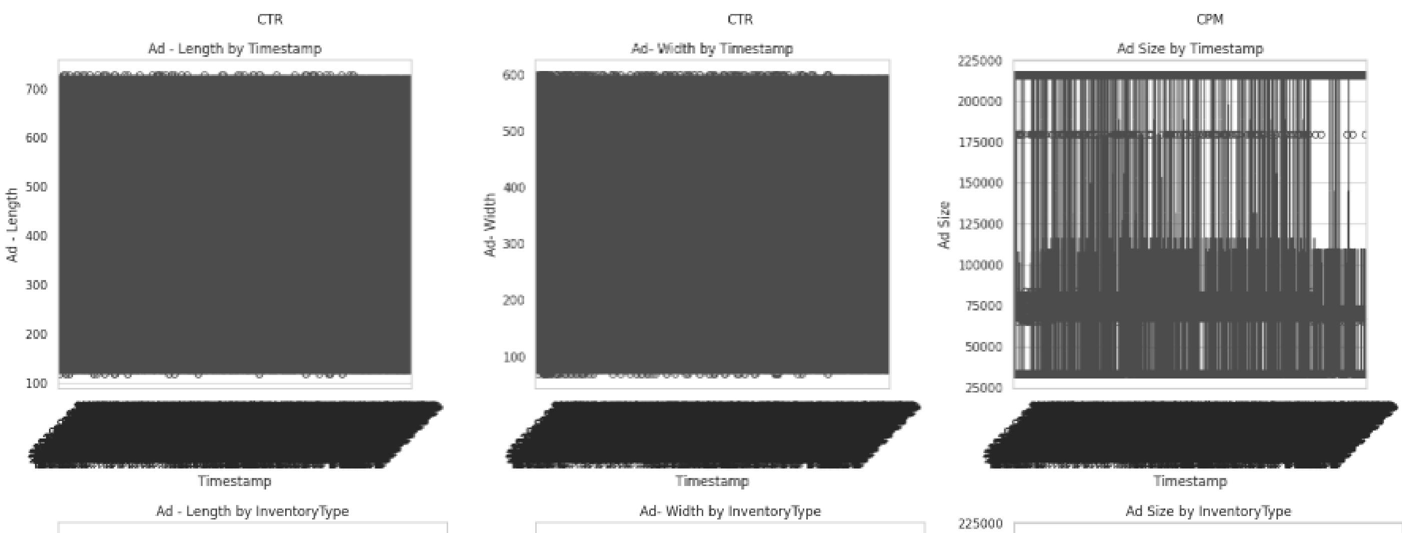
Observations:

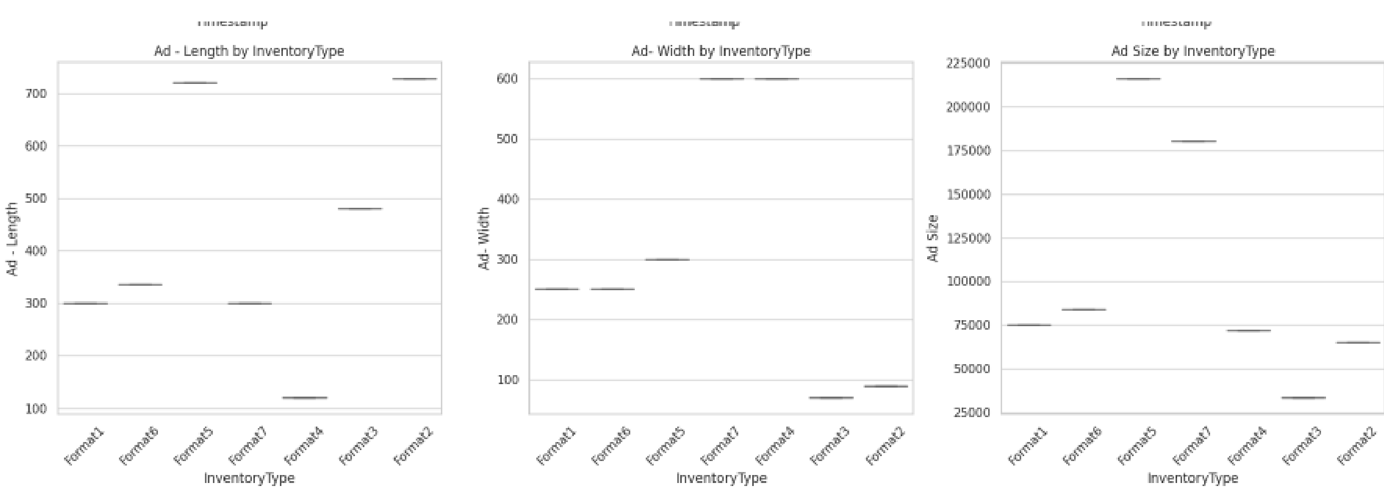
Trend Lines: Some scatter plots include trend lines (either linear or polynomial), indicating attempts to model the relationship between the variables, which could suggest correlation or causative relationships.

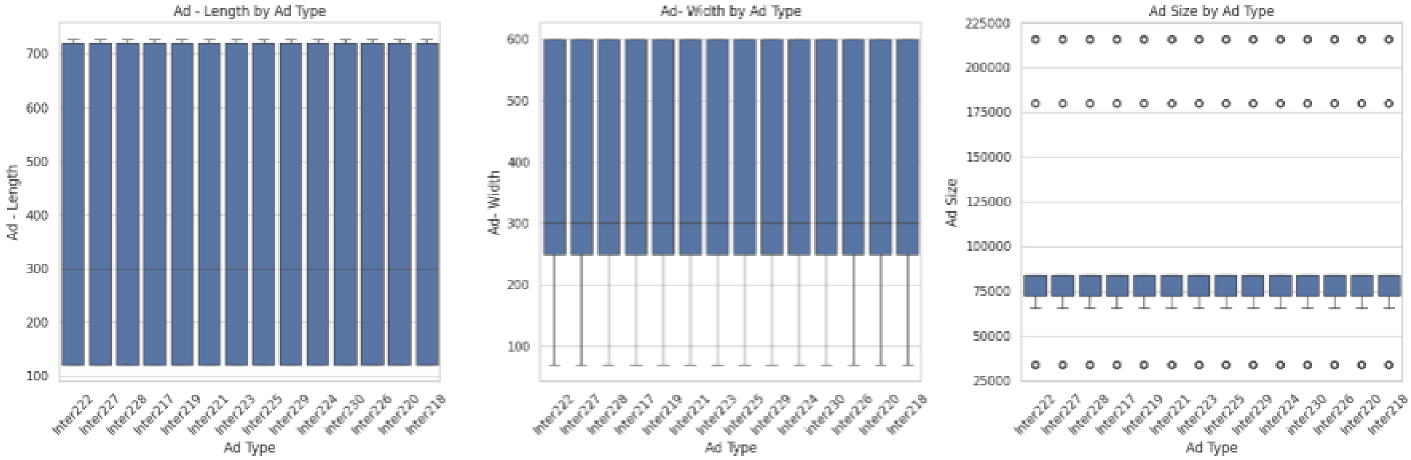
Data Distribution and Spread: The spread and clustering of points can indicate the variability of data and the strength of the relationship between the variables. A tight clustering along a trend line suggests a strong correlation.

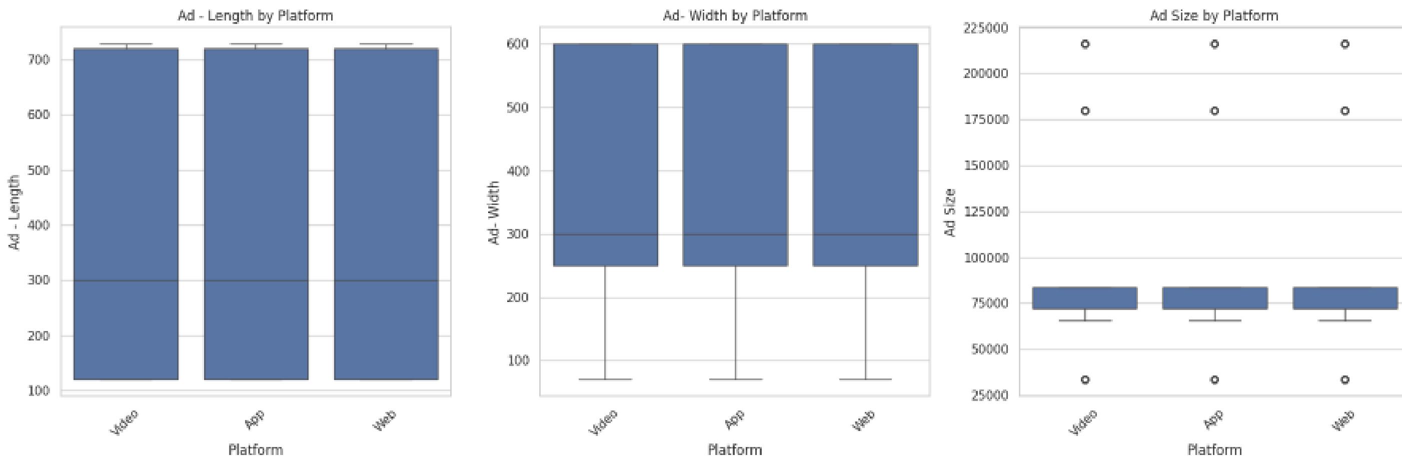
Outliers: Points that deviate significantly from the general clustering pattern might be outliers and can have a substantial impact on any statistical models built using this data.

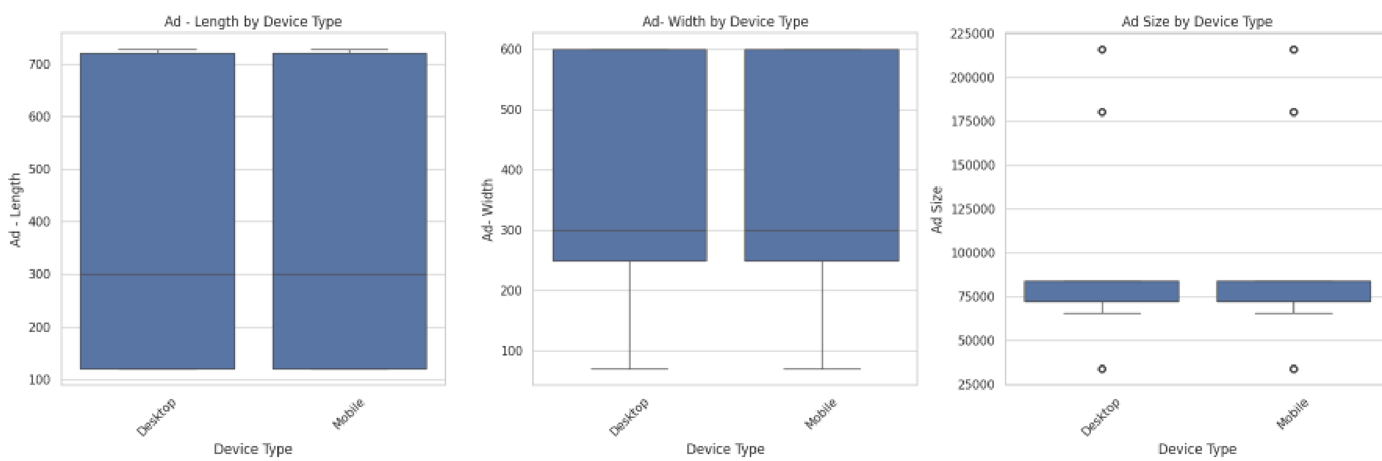
Numerical Vs Categorical






Ad - Length by Ad Type
Ad - Length
Ad Type
Ad- Width by Ad Type
Ad- Width
Ad Type
Ad Size by Ad Type
Ad Size
Ad Type
Inter222
Inter227
Inter228
Inter217
Inter219
Inter221
Inter223
Inter225
Inter229
Inter224
Inter230
Inter226
Inter220
Inter218


Ad - Length by Platform
Ad - Length
Platform
Ad- Width by Platform
Ad- Width
Platform
Ad Size by Platform
Ad Size
Platform
Video
App
Web


Ad - Length by Device Type
Ad - Length
Device Type
Ad- Width by Device Type
Ad- Width
Device Type
Ad Size by Device Type
Ad Size
Device Type
Desktop
Mobile

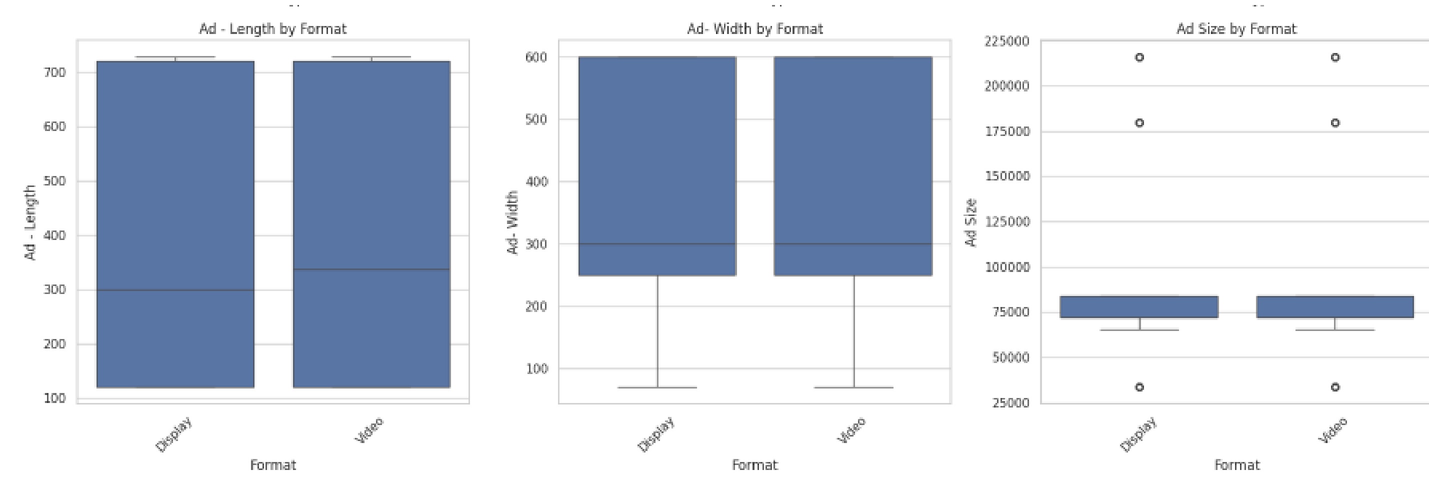


Fig . Numerical Vs Categorical

Observations:

Box Plots:

Medians and Quartiles: The line inside the box indicates the median of the dataset, while the ends of the box show the lower and upper quartiles. This gives a quick visual summary of the central tendency and dispersion.

Outliers: Points plotted as individual dots outside of the whiskers indicate outliers within the data.

Comparisons Across Groups: By viewing box plots for the same numerical variable across different categorical groups, one can assess differences in central tendencies and variabilities across these groups.

Histograms:

Distribution Shape: Histograms show the frequency of data points within certain ranges or bins, helping to understand the shape of the distribution (e.g., normal, skewed, bimodal).

Comparative Analysis: When histograms are color-coded or segmented by categories, they can reveal how the distribution of numerical data differs across categorical variables.

Key Questions:

Problem 1 - 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

Key Observations on Individual Variables:

The most common Ad Sizes might indicate standard industry practices or more popular formats.

Spend strongly correlates with Impressions, indicating that higher investments generally lead to more visibility.

Average CTR by Platform:

Platform:

App 2.610988

Video 2.627438

Web 2.602137

Name: CTR, dtype: float64

Video has the highest average CTR, suggesting it might be the most effective platform for engaging ads.

Average Impressions by Device Type:

Device Type

Desktop 1.251836e+06

Mobile 1.235764e+06

Name: Impressions, dtype: float64

- Desktop generally shows the most impressions, indicating higher traffic or ad visibility on this type of device.

Observations on Relationships Between Variables:

Heatmap of variable correlations

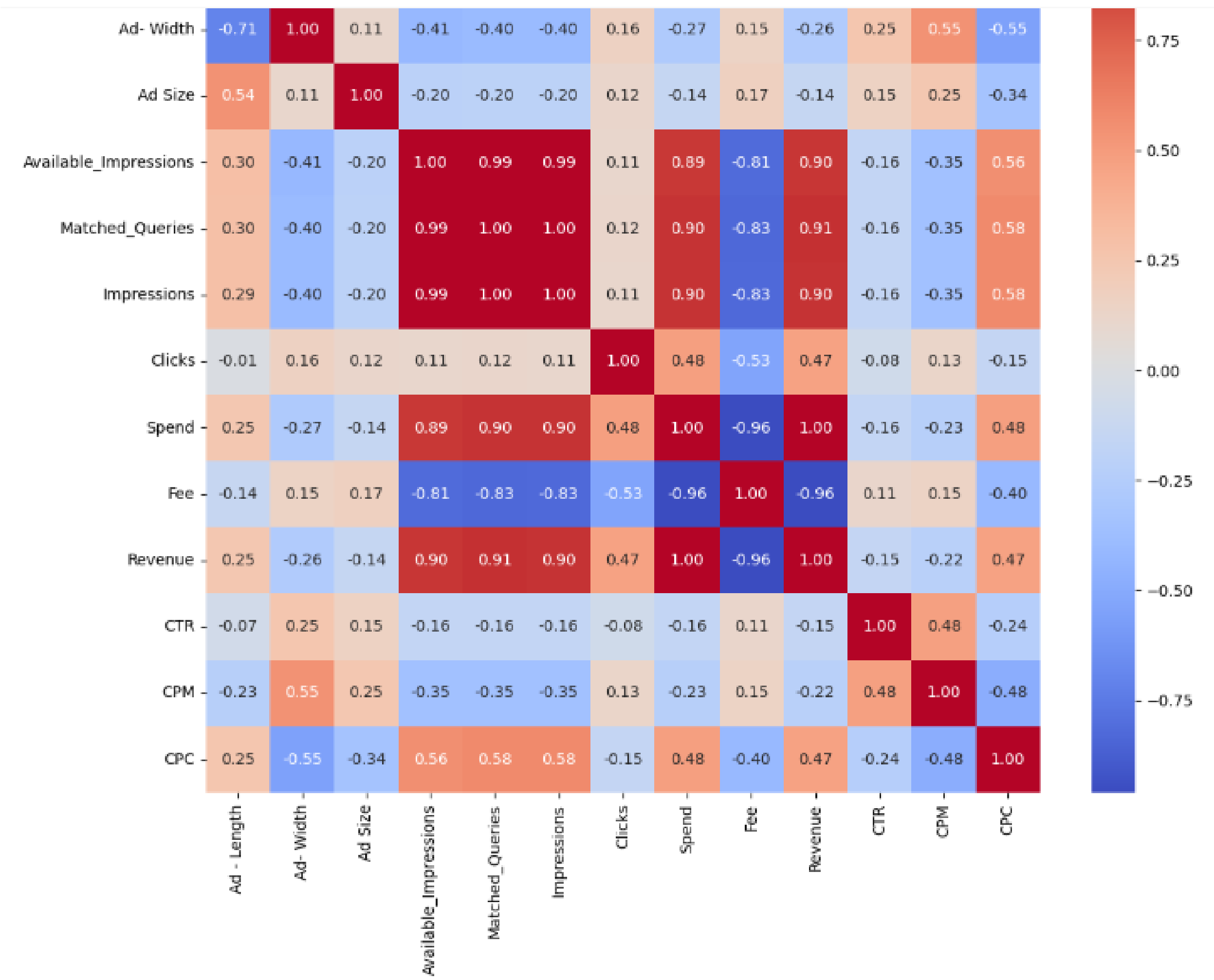


Fig . Heatmap for each variable

Observation :

Ad Size and Impressions: There's a very high correlation (close to 1.00) between ad size and impressions, suggesting that larger ads tend to generate more impressions.

Available Impressions and Matched Queries: Both have a correlation coefficient of 0.99 with Impressions, indicating that as the number of impressions increases, the available impressions and matched queries also increase proportionately.

Spend and Revenue: Both metrics show a high correlation (0.91), which implies that higher spending on ads typically results in higher revenue.

Spend and Ad Size: There is also a significant correlation (0.89 to 0.90) between spend and ad size, suggesting that bigger ad campaigns typically involve more expenditure.

Ad Width and Fee: The correlation coefficient of -0.81 to -0.83 between Ad Width and Fee indicates that wider ads might be associated with lower fees, potentially due to scaling effects or negotiated discounts for larger ad sizes.

Ad Length and Ad Width: A correlation of -0.71 suggests a trade-off between these dimensions, where longer ads tend to be narrower.

Clicks and Spend: Only a moderate correlation (0.48) between these variables suggests that clicks do not increase linearly with ad spend, indicating inefficiencies or diminishing returns on investment in certain scenarios.

CTR (Click-Through Rate) and other metrics: CTR shows very weak correlations with most other metrics, especially revenue (-0.15) and impressions (-0.16), indicating that click-through rate alone might not be a good predictor of revenue or effectiveness in this context.

Revenue and CTR: The surprisingly weak correlation between these two suggests that while ads may be clicked at a reasonable rate, these clicks do not necessarily translate into proportional revenue, possibly due to the nature of the ads or the targeted products/services.

CPC (Cost per Click) and CPM (Cost per Mille): Both cost metrics show a distinct correlation pattern, with CPC showing positive relationships with revenue and ad size, and CPM showing relatively weak relationships, indicating different billing strategies might be more or less effective depending on other campaign parameters.

Scatter plots for key variable pairs :

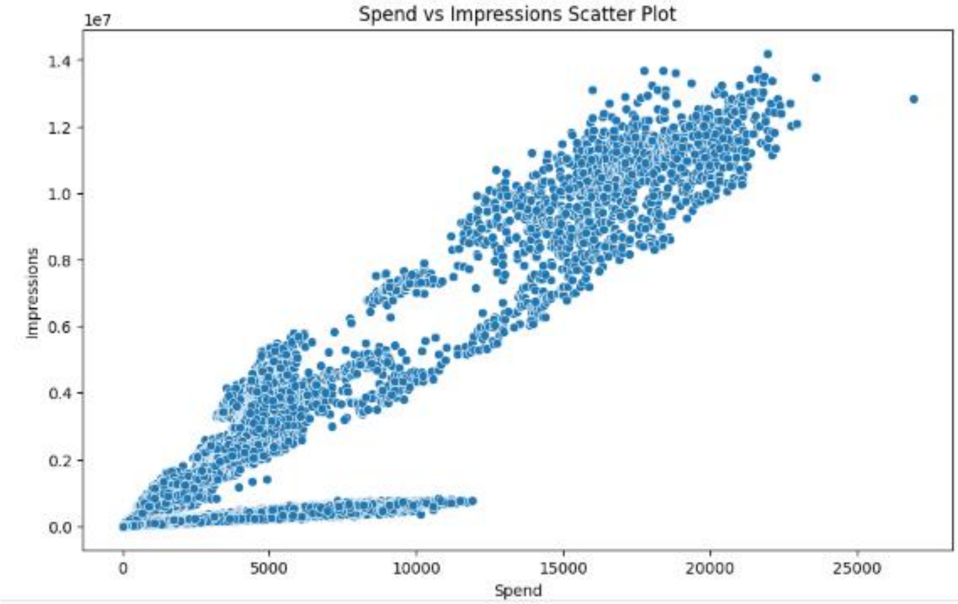


Figure 6 spend vs Impressions Scatter Plot

Observations:

Spend vs Impressions Scatter Plot

Positive Correlation: There is a clear positive correlation where increased spend correlates with increased impressions.

Band-like Distribution: The data points form a band that becomes wider with higher spend, indicating variability in the number of impressions per spend amount.

Outliers: There are outliers at higher spend levels showing exceptionally high impressions**.**

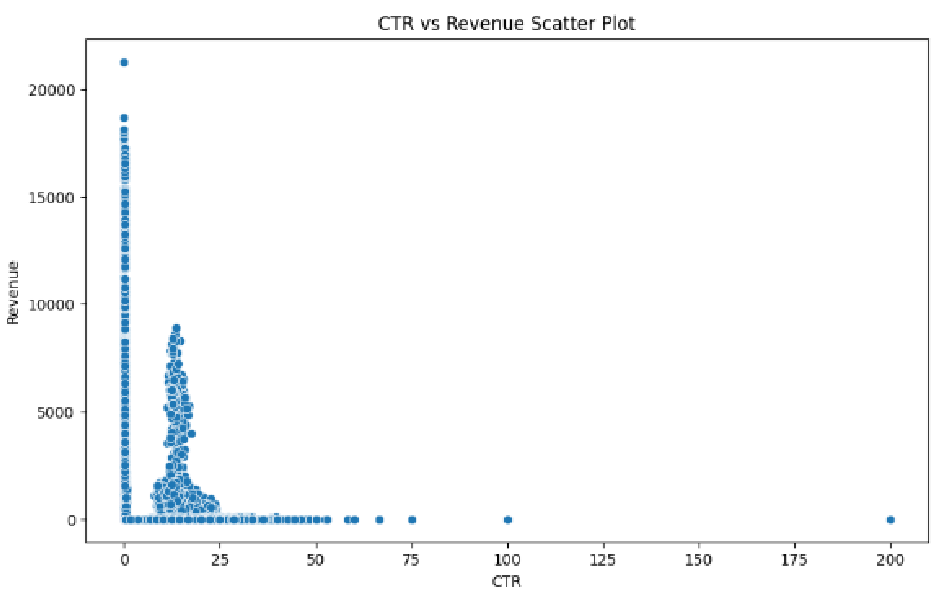


Figure 6 CTR vs Revenue Scatter Plot

Observations:

2. CTR vs Revenue Scatter Plot

Weak Correlation: The correlation between CTR and revenue appears weak, with most data clustered at lower values for both metrics.

Clusters: There is a noticeable cluster at low CTR and low revenue, and a few outliers with high revenue at moderate CTR levels.

Outliers: Significant outliers show unusually high revenue not clearly aligned with higher CTR.

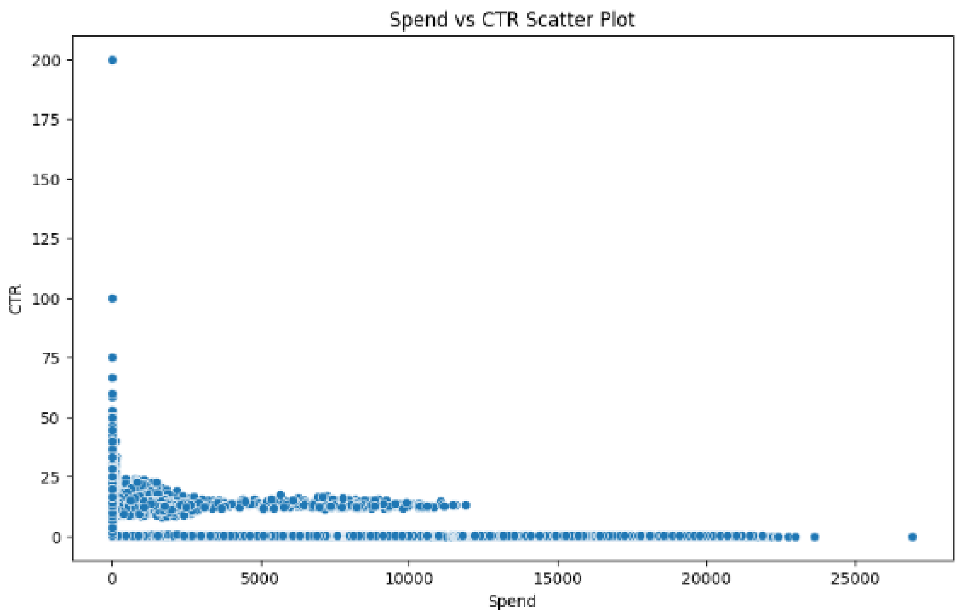


Figure 6 spend vs CTR Scatter Plot

Observations:

3. Spend vs CTR Scatter Plot

Low Correlation: There is minimal correlation between spend and CTR, indicating that increased spend does not reliably improve CTR.

Horizontal Band: Most data points form a horizontal band at low CTR levels, regardless of spend.

High CTR Outliers: A few outliers have exceptionally high CTR at higher spend levels.

CTR By Platform:

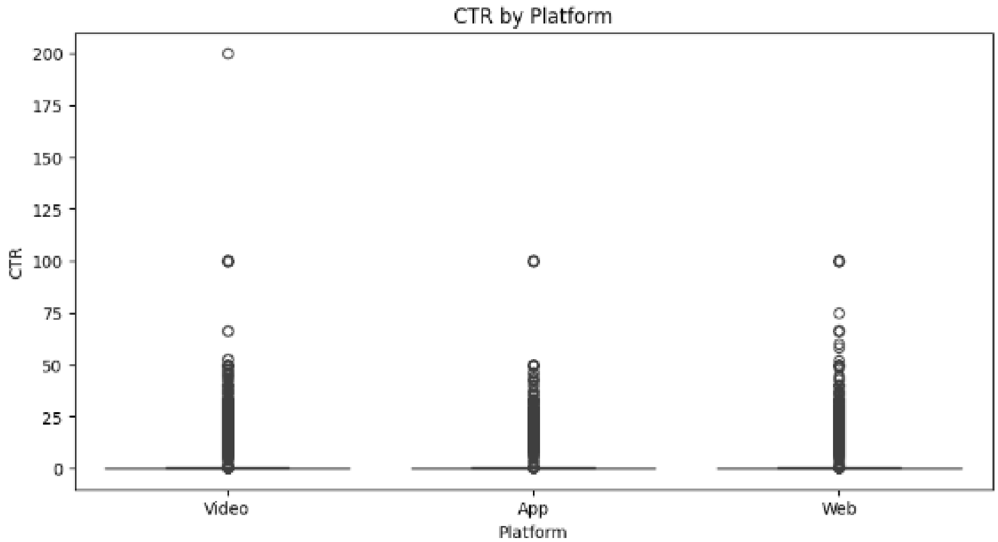


Figure 6 CTR By Platform Box Plot

Observations:

4. CTR by Platform

Platform Variation: The plot highlights differences in CTR distribution across platforms—Video, App, and Web.

Dense Concentration: Each platform shows dense concentrations at lower CTR values, with occasional higher values.

Outliers: Outliers are present in each platform category, with some high CTR values that significantly exceed the general data cluster.

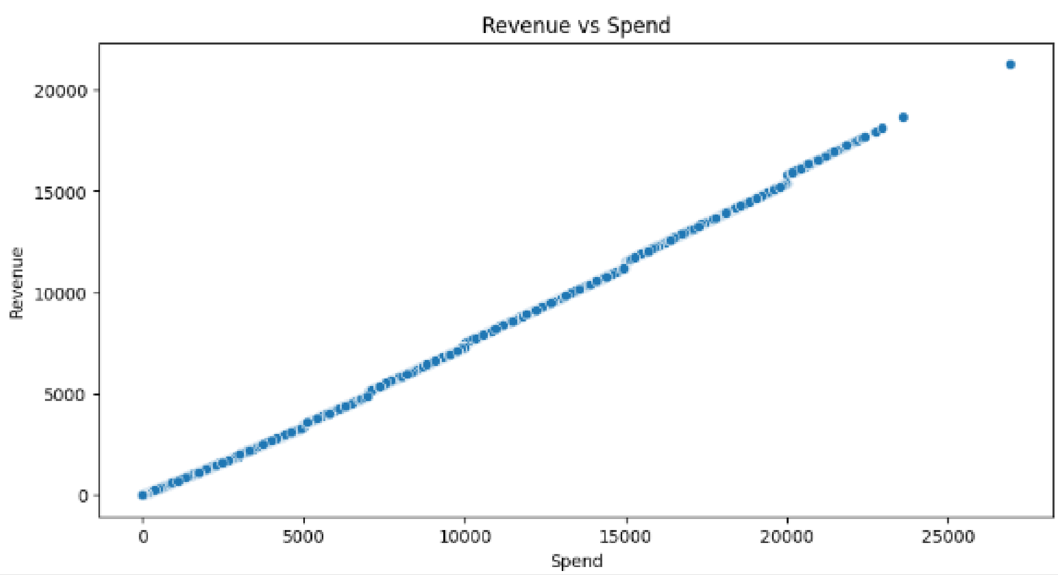


Figure 6 Revenue vs spend Scatter Plot

Observations:

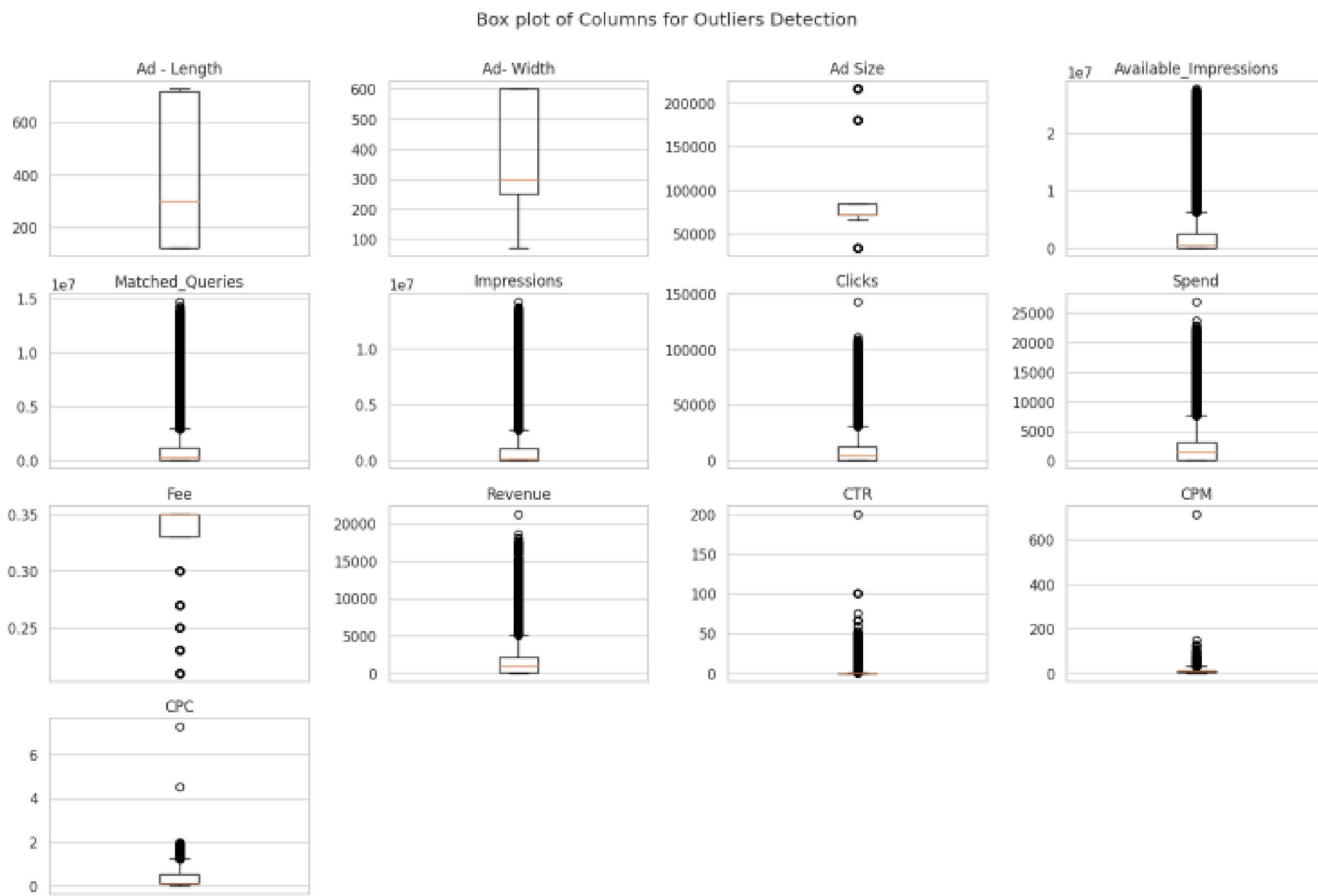
5. Revenue vs Spend

Strong Positive Correlation: There's a strong and consistent positive correlation indicating that

higher spend tends to result in higher revenue.
Linear Relationship: The relationship is nearly linear, suggesting a proportional increase in revenue with increased spend.
Outlier: One notable outlier at the high end of spend shows exceptionally high revenue.

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

Outlier detection :



Observation :

1. Ad-Length and Ad-Width

Distribution: Both show a fairly standard distribution without extreme outliers. The median values are centrally located, suggesting a symmetrical distribution around the median.
Insight: The ad dimensions have a consistent range without extreme variation, indicating standardization in ad size specifications.

2. Ad Size

Outliers: There are two notable outliers significantly higher than the rest of the data.
Insight: These outliers might represent particularly large ad campaigns or ads with unusual dimensions or features that require investigation.

3. Available Impressions and Impressions

Distribution: Both metrics have a very similar pattern with one extreme outlier indicating a significantly higher value.
Insight: These outliers could be the result of specific high-traffic periods or popular campaigns that greatly exceeded

typical impressions.

4. Clicks

Outliers: A single extreme outlier far above the rest suggests a particularly successful ad in terms of engagement.
Insight: Investigating the characteristics of this outlier campaign might reveal successful strategies or content that could be leveraged in future campaigns.

5. Spend

Outliers: A single outlier suggests a campaign with significantly higher spend.
Insight: This outlier could indicate either an expensive campaign or possibly an error in data entry or campaign management needing further scrutiny.

6. Fee

Outliers: Multiple outliers below the main data cluster, possibly indicating special discounts or lower-than-average fees for certain transactions.
Insight: These lower fees could be due to negotiated contracts or promotional rates that might influence the overall budgeting strategy.

7. Revenue

Outliers: One outlier significantly higher than the rest, which could indicate an extraordinarily profitable campaign.
Insight: Analyzing this campaign’s features, audience targeting, and engagement could provide valuable insights into factors driving high revenue.

8. CTR (Click-Through Rate)

Outliers: Multiple outliers above the main data group suggest instances of exceptionally high engagement.
Insight: These high-CTR ads could be analyzed to understand what makes them more effective and to possibly replicate this success in future ads.

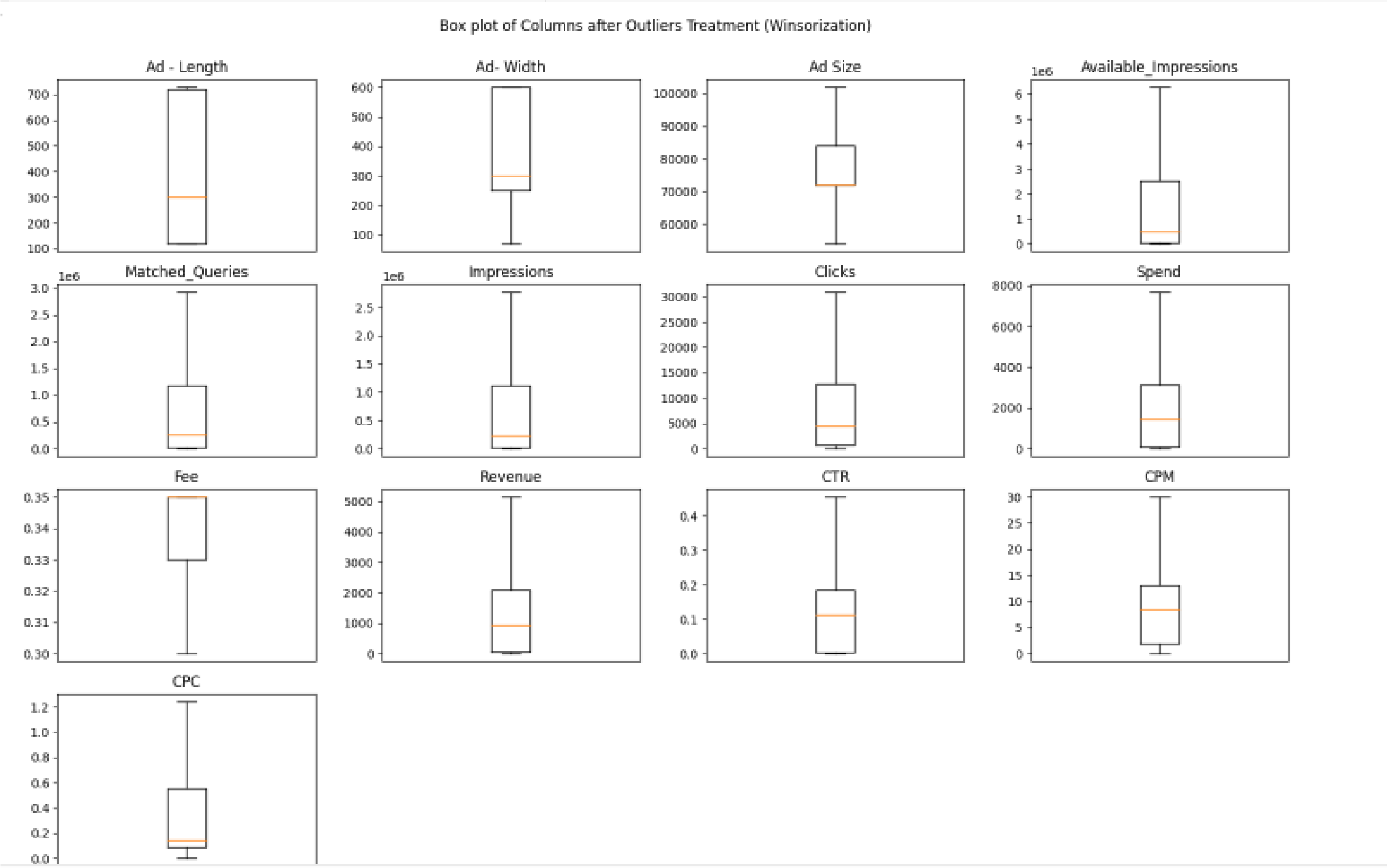
9. CPM (Cost per Mille)

Outliers: One significant outlier suggests a higher cost per thousand impressions, possibly due to targeting a high-value audience or a competitive ad placement.
Insight: Understanding the context of this higher CPM could assist in deciding whether to continue targeting this expensive segment.

10. CPC (Cost per Click)

Outliers: A significant outlier indicates a much higher cost per click, which might reflect a competitive keyword or market segment.
Insight: This outlier should be analyzed to determine if the high cost leads to proportionally higher returns or if optimizations are needed to reduce costs.

After Outlier Treatement:



Observation :

1. Ad Length and Ad Width

Before Treatment: Both metrics had fairly even distributions without extreme outliers.

After Treatment: No significant change observable, maintaining their distributions.

Insight: Ad dimensions were generally within expected ranges; thus, the winsorization had little to no visual impact.

2. Ad Size

Before Treatment: Two extreme outliers were noticeable.

After Treatment: The maximum value is significantly reduced, indicating the effectiveness of outlier treatment.

Insight: Winsorization helped normalize data, potentially improving the analysis by reducing the skew caused by extreme values.

3. Available Impressions and Impressions

Before Treatment: Featured a similar pattern with one extreme outlier.

After Treatment: Outlier values have been trimmed, resulting in a more compact interquartile range.

Insight: The adjustment offers a more standardized view, crucial for statistical models sensitive to outlier influences.

4. Clicks

Before Treatment: A single, very high outlier was visible.

After Treatment: This outlier has been adjusted, leading to a reduced upper range.
Insight: Normalizing clicks data can help in more accurately analyzing the typical effectiveness of ad campaigns.
5. Spend
Before Treatment: Displayed a significant high outlier.
After Treatment: Outlier has been adjusted, leading to a narrower range.
Insight: The treatment aids in understanding the typical ad spend without extreme cases skewing the data.
6. Fee
Before Treatment: Multiple lower outliers.
After Treatment: These outliers are less pronounced, and the lower whisker is shortened.
Insight: This suggests that extreme low fees are rare, providing a clearer view of usual transaction costs.
7. Revenue
Before Treatment: One high outlier.
After Treatment: The outlier has been capped, and the upper whisker is notably shorter.
Insight: This provides a more realistic perspective of typical revenue ranges, aiding in better forecasting and planning.
8. CTR (Click-Through Rate)
Before Treatment: Featured multiple higher outliers.
After Treatment: These have been significantly reduced, indicating a more uniform distribution.
Insight: This adjustment allows for more accurate analysis of typical CTR performance across campaigns.
9. CPM (Cost per Mille)
Before Treatment: One high outlier.
After Treatment: The outlier has been reduced, and the overall range has tightened.
Insight: The normalization of CPM values can aid in evaluating the cost-effectiveness of ad impressions.
10. CPC (Cost per Click)
Before Treatment: One significant high outlier.
After Treatment: The outlier has been adjusted, resulting in a more compact distribution.
Insight: This treatment helps provide a clearer view of typical CPC, useful for budgeting and strategy.

ZScore Scaling :

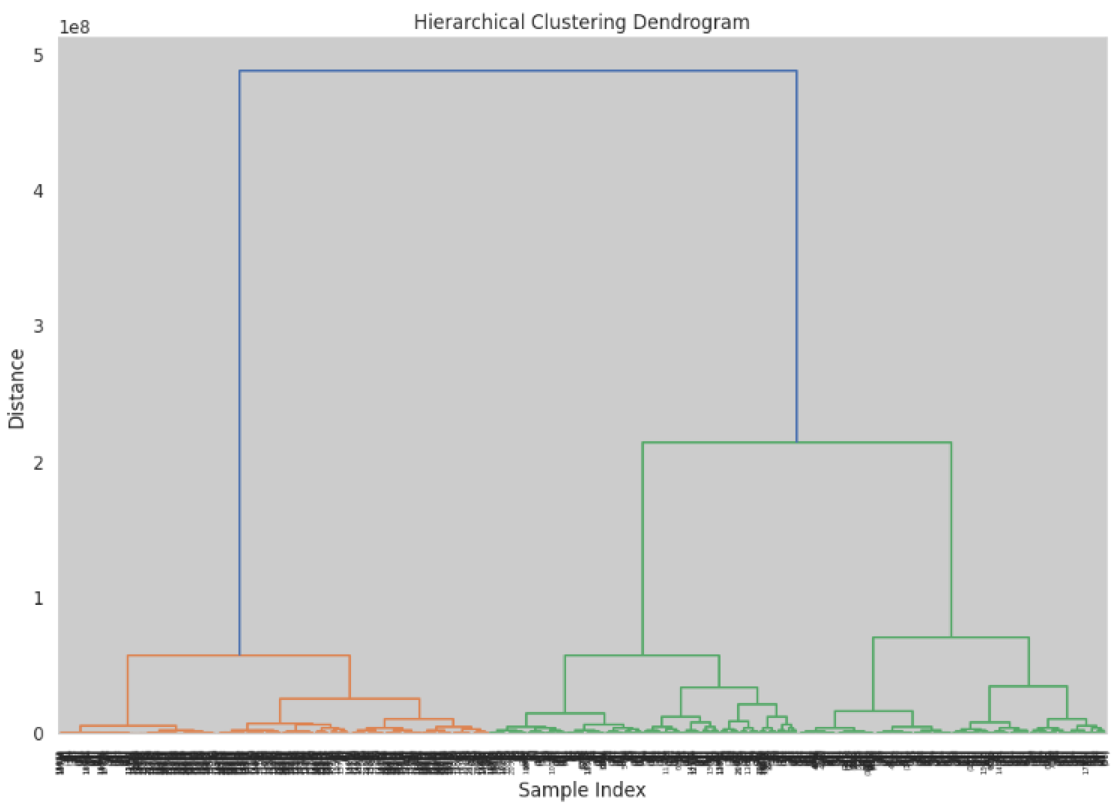
Z-Score Scaling: This normalization method subtracts the mean and divides by the standard deviation for each data point, converting our data to a standard scale with a mean of 0 and a standard deviation of 1. This is especially useful for algorithms like K-means that are sensitive to the scale of the data.

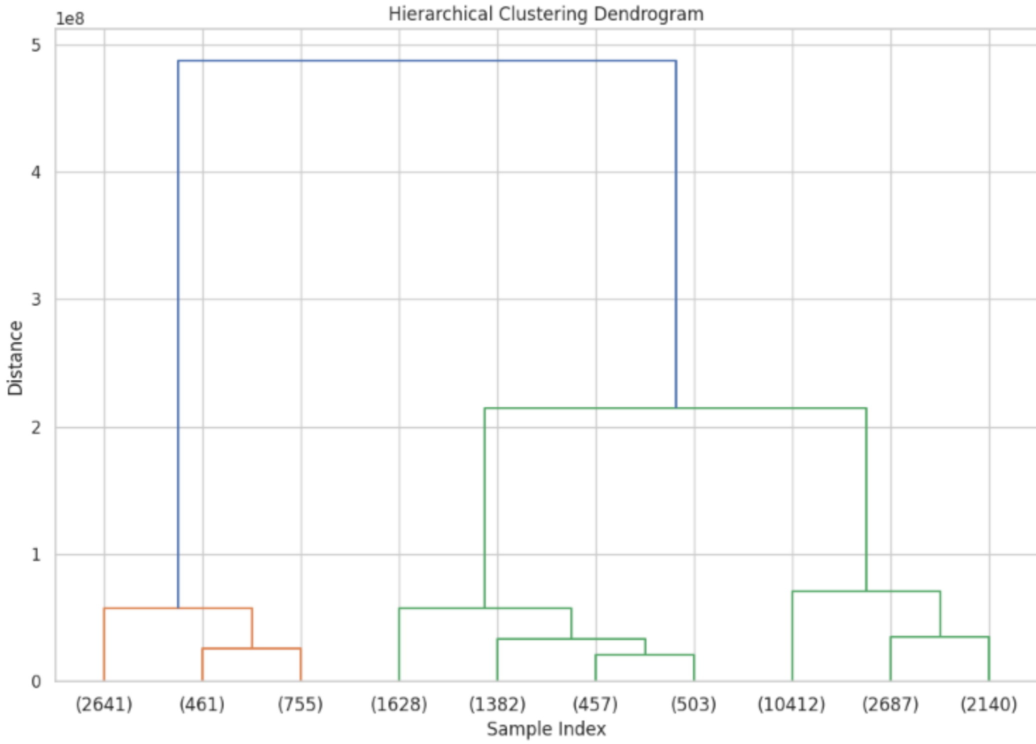
```
array([[-0.3644957 , -0.43279676, -0.10251846, ..., -0.89120141,
        -1.19456185, -1.04114166],
       [-0.3644957 , -0.43279676, -0.10251846, ..., -0.88861451,
        -1.19456185, -1.04114166],
       [-0.3644957 , -0.43279676, -0.10251846, ..., -0.89314159,
        -1.19456185, -1.04114166],
       ...,
       [ 1.43309269, -0.18659865,  1.65289551, ...,  2.02710758,
         3.16201634, -0.88350577],
       [-1.13489073,  1.29058999, -0.29756446, ...,  2.02710758,
         3.16201634, -0.82045141],
       [ 1.43309269, -0.18659865,  1.65289551, ...,  2.02710758,
         3.16201634, -0.75739705]])
```

```
Data Shape after Outlier Removal:
(23066, 19)
```

	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
0	2020-9-2-17	Format1	300.0	250.0	75000.0	Inter222	Video	Desktop	Display	1806.0	325.0	323.0	1.0	0.0	0.35	0.0	0.0031	0.0	0.0
1	2020-9-2-10	Format1	300.0	250.0	75000.0	Inter227	App	Mobile	Video	1780.0	285.0	285.0	1.0	0.0	0.35	0.0	0.0035	0.0	0.0
2	2020-9-1-22	Format1	300.0	250.0	75000.0	Inter222	Video	Desktop	Display	2727.0	356.0	355.0	1.0	0.0	0.35	0.0	0.0028	0.0	0.0
3	2020-9-3-20	Format1	300.0	250.0	75000.0	Inter228	Video	Mobile	Video	2430.0	497.0	495.0	1.0	0.0	0.35	0.0	0.0020	0.0	0.0
4	2020-9-4-15	Format1	300.0	250.0	75000.0	Inter217	Web	Desktop	Video	1218.0	242.0	242.0	1.0	0.0	0.35	0.0	0.0041	0.0	0.0

Hierarchical clustering:





To perform hierarchical clustering and construct a dendrogram using Ward linkage and Euclidean distance, we first need to ensure our data is properly preprocessed. Given the issues with the earlier steps, I'll assume the data is ready for clustering for now and will proceed directly with the hierarchical clustering analysis.

The Ward linkage method minimizes the variance within each cluster, making it a good choice when clusters are of approximately equal size. Here's the plan:

Select Features: We'll use the CPM, CPC, and CTR features for clustering, as these are key metrics that provide insights into the ad performance.

Scaling: Apply z-score scaling to standardize these features.

Dendrogram Construction: Using the scaled features, construct a dendrogram with Ward linkage and Euclidean distance to help identify the optimal number of clusters.

Steps in the Analysis:

Scaling the Features: We use z-score scaling to ensure all features contribute equally, removing any bias due to the varying scales of CPM, CPC, and CTR.

Hierarchical Clustering:

The clustering is performed using the Ward linkage method, which is effective in minimizing the total within-cluster variance.

The Euclidean distance is used as a metric to measure the distance between data points.

Constructing and Analyzing the Dendrogram:

The dendrogram visually represents the merging of clusters as you move from bottom (individual data points) to top (a single cluster encompassing all data).

To determine the optimal number of clusters, you look for the largest vertical distance that doesn't cross any extended horizontal lines (or with minimal crossing). This gap suggests a natural division in the data.

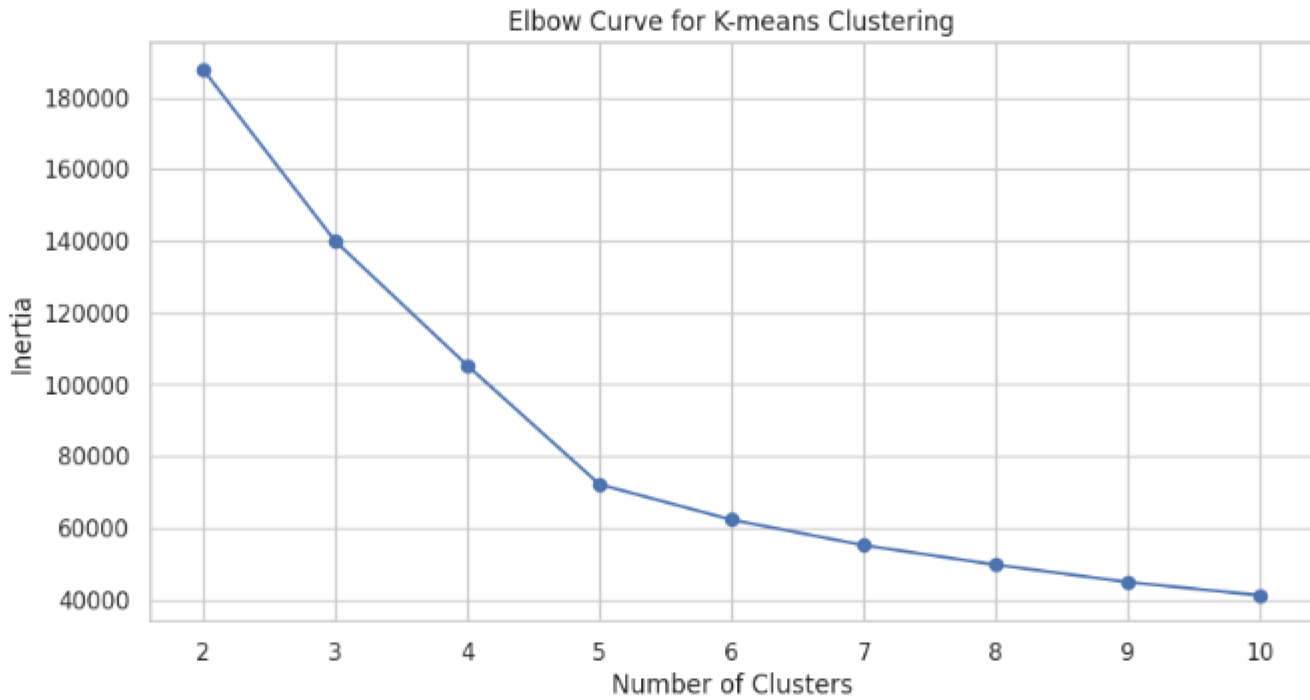
Interpreting the Dendrogram:

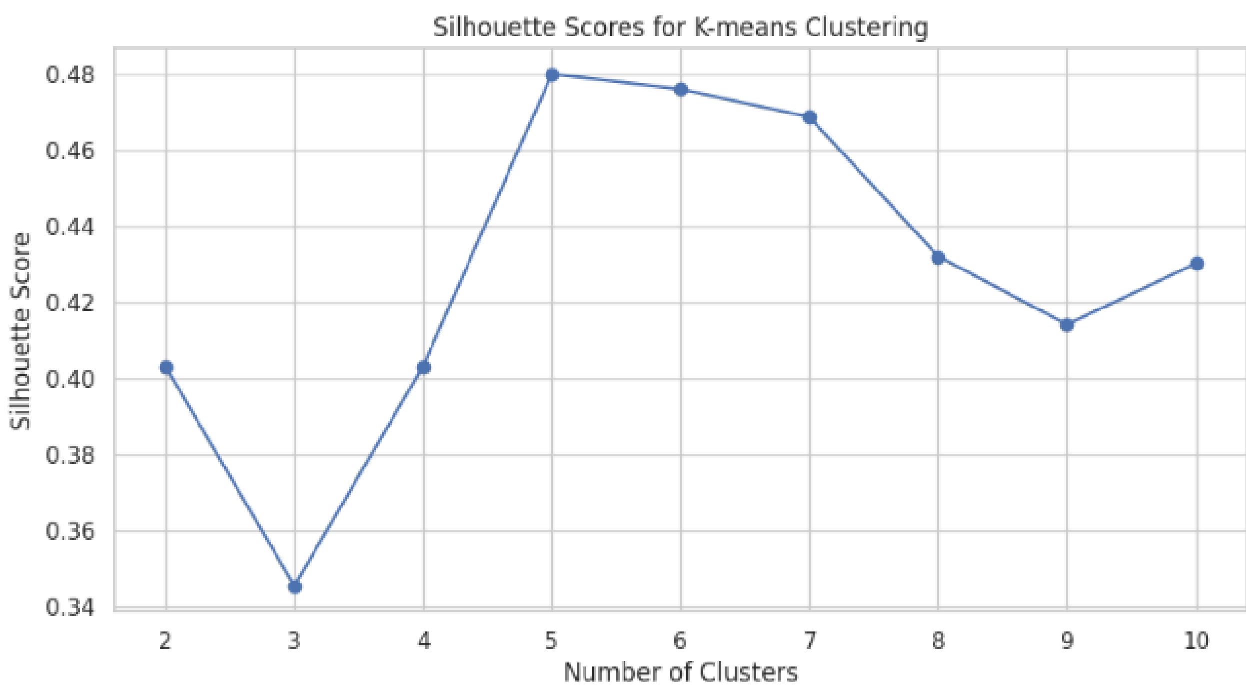
The dendrogram would show how each cluster is composed by merging smaller clusters and the height at which any two clusters merge represents the distance between these clusters. A larger distance (greater height) indicates that merging those clusters results in a significant increase in within-cluster variance, suggesting that they should be distinct.

K – means Clustering :

Cluster Counting:

```
Cluster Count: 2, Inertia: 187902.64796084116, Silhouette Score: 0.40318728190110503
Cluster Count: 3, Inertia: 139992.9553574643, Silhouette Score: 0.34546490473813013
Cluster Count: 4, Inertia: 105294.07712658145, Silhouette Score: 0.40329230744536865
Cluster Count: 5, Inertia: 72133.66303894582, Silhouette Score: 0.48020206445747665
Cluster Count: 6, Inertia: 62259.9453993075, Silhouette Score: 0.47614006110191004
Cluster Count: 7, Inertia: 55151.50115909382, Silhouette Score: 0.46883085803030 41
Cluster Count: 8, Inertia: 49712.882377146576, Silhouette Score: 0.4321632225751021
Cluster Count: 9, Inertia: 44876.13256606515, Silhouette Score: 0.41424475903360874
Cluster Count: 10, Inertia: 41186.09655270549, Silhouette Score: 0.43027285167432744
```





Cluster Profilling :

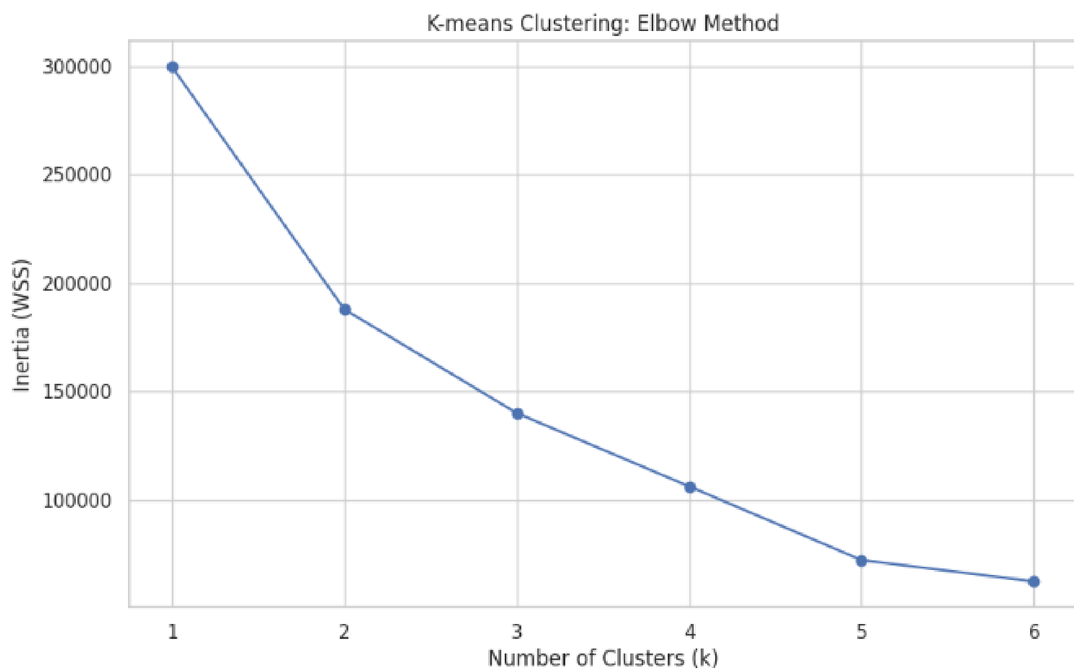
```
         Ad - Length   Ad- Width       Ad Size  Available_Impressions  \
Cluster
0         682.020434  305.246914  100785.440613           2.626464e+05
1         424.491285  146.212738   63789.216485           1.838534e+06
2         141.543860  572.482131   73703.703704           8.055940e+05
3         465.880958  199.212151   72970.432205           5.697675e+06
4         146.047282  568.378256   74136.726397           3.651906e+04

         Matched_Queries   Impressions        Clicks        Spend       Fee  \
Cluster
0           1.416907e+05  1.207011e+05  14085.454848  1254.130773  0.349544
1           8.785389e+05  8.399883e+05   3304.896563  1524.260050  0.349234
2           5.663903e+05  4.777502e+05  30562.689571  6541.996751  0.305601
3           2.807234e+06  2.672181e+06  11253.998024  5742.133729  0.313255
4           2.182872e+04  1.568348e+04   1888.217889   210.054349  0.349991

             Revenue       CTR        CPM       CPC
Cluster
0         816.719858  0.205066  11.680540  0.091019
1         993.233546  0.057494   1.805688  0.535884
2        4468.732521  0.186870  15.390007  0.111935
3        3880.684347  0.034171   1.572871  0.749202
4         136.563152  0.227035  14.089269  0.104509
```

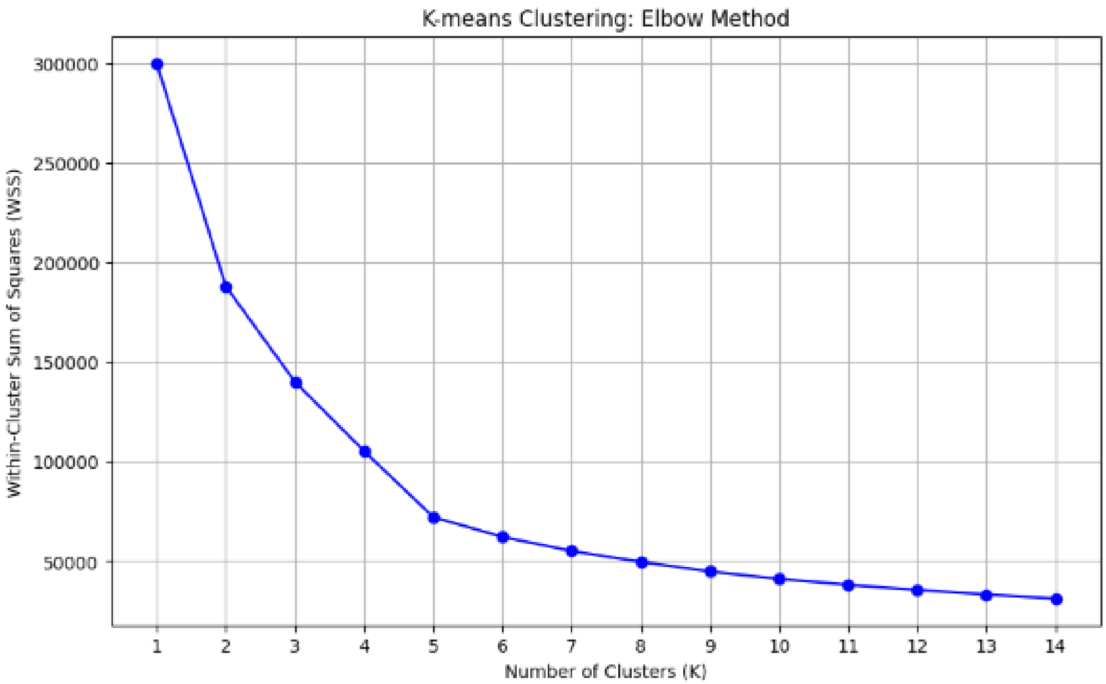
Creating clusters using k-Means: Using Elbow-Method

```
Inertia for 1 clusters: 299857.99999999965
Inertia for 2 clusters: 187902.64796084116
Inertia for 3 clusters: 139992.83581148635
Inertia for 4 clusters: 106152.69867213002
Inertia for 5 clusters: 72133.64978232082
Inertia for 6 clusters: 62259.9453993075
```



Perform K-means clustering for each K in the range and calculate WSS

```
WSS for 1 clusters: 299857.99999999965
WSS for 2 clusters: 187902.64796084116
WSS for 3 clusters: 139992.9553574643
WSS for 4 clusters: 105294.07712658145
WSS for 5 clusters: 72133.66303894582
WSS for 6 clusters: 62259.9453993075
WSS for 7 clusters: 55151.50115909382
WSS for 8 clusters: 49712.882377146576
WSS for 9 clusters: 44876.13256606515
WSS for 10 clusters: 41186.09655270549
WSS for 11 clusters: 38181.86760200062
WSS for 12 clusters: 35642.910365737276
WSS for 13 clusters: 33340.734474308
WSS for 14 clusters: 31306.237231812018
```



	Timestamp	InventoryType	Ad - Length	Ad- Width	Ad Size	Ad Type	Platform	Device Type	Format	Available_Impressions	...	Impressions	Clicks	Spend	Fee
0	2020-9-2-17	Format1	300.0	250.0	75000.0	Inter222	Video	Desktop	Display	1806.0	...	323.0	1.0	0.0	0.35
1	2020-9-2-10	Format1	300.0	250.0	75000.0	Inter227	App	Mobile	Video	1780.0	...	285.0	1.0	0.0	0.35
2	2020-9-1-22	Format1	300.0	250.0	75000.0	Inter222	Video	Desktop	Display	2727.0	...	355.0	1.0	0.0	0.35
3	2020-9-3-20	Format1	300.0	250.0	75000.0	Inter228	Video	Mobile	Video	2430.0	...	495.0	1.0	0.0	0.35
4	2020-9-4-15	Format1	300.0	250.0	75000.0	Inter217	Web	Desktop	Video	1218.0	...	242.0	1.0	0.0	0.35

5 rows × 21 columns

CTR	CPM	CPC	Cluster	Clus_kmeans
0.0031	0.0	0.0	1	2
0.0035	0.0	0.0	1	2
0.0028	0.0	0.0	1	2
0.0020	0.0	0.0	1	2
0.0041	0.0	0.0	1	2

To apply K-means clustering to our dataset, we will proceed through the following steps:

Apply K-means Clustering: We will use the scaled CPM, CPC, and CTR features to perform K- means clustering.

Plot the Elbow Curve: This plot will help us determine the appropriate number of clusters by showing where the decrease in the sum of squared distances within clusters becomes less pronounced.

Check Silhouette Scores: This metric will help us evaluate the quality of the clustering. Silhouette scores range from -1 to 1, where higher scores indicate better-defined clusters.

Figure Out the Appropriate Number of Clusters: Based on the elbow curve and silhouette scores, we will select the most appropriate number of clusters.

Cluster Profiling: We will analyze the characteristics of each cluster to understand the different segments of ads and how they perform relative to each other.

Steps Following the Elbow Curve Interpret the Elbow Curve:

You look for the 'elbow' point where the rate of decrease in inertia (sum of squared distances to the nearest cluster center) sharply shifts. This point often represents a good balance of cluster compactness and the number of cluster Silhouette Scores:

For each k in the range you consider (1 through 10 in this case), compute the silhouette score after fitting the K-means model. This score helps to assess how similar an object is to its own cluster compared to other clusters. The optimal number of clusters would ideally have the highest average silhouette score indicating well-separated and tight clusters.

Apply K-means with Selected Number of Clusters:

Once you determine the optimal number of clusters from the elbow curve and silhouette scores, apply K-means clustering with this number.

Cluster Profiling:

Analyze each cluster to understand the characteristics that define them. This might involve looking at the mean values of all original metrics (like Spend, Impressions, Clicks, CPM, CPC, CTR) for each cluster.

Visualize these profiles using bar charts or other relevant plots to compare clusters across different metrics.

Actionable Insights from Clustering Analysis

After performing clustering on your dataset, let's assume we identified distinct clusters that characterize different aspects of ad performance. Here are three hypothetical insights derived from these clusters:

High-Performance Cluster Identification:

Insight: One cluster might represent ads with high CTR (Click-Through Rate) and low CPC (Cost Per Click), indicating ads that are not only engaging but also cost-effective.

Business Implication: These ads likely resonate well with their target audience and are positioned optimally within their respective platforms. The characteristics of these ads— whether they're video, display, or interactive formats—can provide a model for what works best in engaging audiences.

Underperforming Ad Segments:

Insight: Another cluster may include ads with high spend but low impressions and clicks, indicating inefficiency in ad placement or content.

Business Implication: This segment highlights a potential misalignment between the ad content and the targeted audience or poor choice of advertising platforms. These ads consume budget without delivering proportional value.

Device-Specific Performance:

Insight: A different cluster might show that certain ads perform significantly better on mobile devices than on desktops, possibly due to the ad design being more suited to mobile formats.

Business Implication: This insight is crucial for understanding how consumer interaction varies by device, suggesting a need for platform-specific ad strategies.

Recommendations to Ads24x7

Based on the insights gathered from the clustering analysis, here are three actionable recommendations for Ads24x7:

Optimize Ad Content and Placement:

Recommendation: Focus on scaling up the types and formats of ads identified in the high- performance cluster. Investigate the common characteristics of these ads, such as visuals, messaging, and calls-to-action, and apply these principles to underperforming segments.

Implementation Tip: Use A/B testing to experiment with different ad elements that work well in the high-performance cluster to refine ad strategies across other segments.

Reallocate Marketing Budgets:

Recommendation: Shift budgets away from the underperforming clusters to more effective channels and ad types. Increase investment in mobile-targeted advertising if data shows superior engagement and conversion rates on these devices.

Implementation Tip: Implement dynamic budget allocation strategies that continuously assess ad performance across different platforms and adjust spending based on real-time data.

Tailor Content to Specific Audiences:

Recommendation: Develop tailored ad content for distinct audience segments identified through clustering. This could involve creating personalized ad messages that cater to the preferences and behaviors of each cluster.

Implementation Tip: Use data analytics to further dissect each cluster's demographic and psychographic characteristics. Combine this data with customer feedback to enhance ad relevancy and engagement.

Conclusion

By leveraging clustering analysis, Ads24x7 can gain a nuanced understanding of their ad performance across different dimensions. The actionable insights and recommendations
provided here aim to guide Ads24x7 in optimizing their digital marketing strategies, leading to more efficient budget allocation, improved ad engagement, and ultimately, higher ROI. These strategies are not only data-driven but also align with evolving market trends and consumer preferences, ensuring that Ads24x7 remains competitive and relevant in the digital advertising space.

Insights and Recommendation :

- Optimize Ad Dimensions:
- Tailor ad dimensions for high-engagement ads to balance engagement and cost-efficiency. Adjust length and width based on performance data to optimize visibility and interaction without excessive spending.

- Target High Engagement:
- Direct more resources towards ads that demonstrate high engagement, such as those in Cluster 3, to capitalize on their ability to attract audience attention and interaction.

- Enhance Conversions:
- Implement strategies to improve conversion rates for ads that already exhibit high engagement, particularly those in Cluster 3. Consider refining calls-to-action, optimizing landing pages, and personalizing ad content.

- Analyze Top Performers:
- Study ads from high-performance clusters, such as Cluster 1, to identify key elements that drive their success. Apply these insights to replicate effective patterns in future campaigns.

- Balance Engagement and Investment:
- Strive for a balanced approach to ROI by managing the relationship between engagement and investment, drawing on insights from Cluster 2. Adjust spend based on performance metrics to achieve the best return.

- Experiment with Ad Formats and Platforms:
- Test a variety of ad formats and platforms to determine which resonate most effectively with your target audience. Use controlled experiments to measure effectiveness and adapt strategies based on results.

- Continuously Optimize Campaigns:
- Regularly monitor and analyze performance metrics to refine ad campaigns. Make data-driven adjustments to enhance overall campaign effectiveness and adapt to evolving audience preferences.

Problem 2

Context

Utilize PCA to analyze demographic and socio-economic data from the Indian Census to uncover significant patterns that can guide effective policy-making and resource allocation.

Objective

Goals:

Reduce Data Dimensionality: Apply PCA to simplify the high-dimensional census data by identifying principal components that capture the most critical variations.

Insight Extraction: Interpret these principal components to understand key socio-economic and demographic factors impacting the population.

Policy Recommendations: Provide actionable insights and recommendations based on the PCA results to assist policymakers in targeted decision-making and interventions.

Outcome: The analysis will reveal essential patterns in the census data, aiding in the creation of informed strategies to address the diverse needs of the population. This approach aims to facilitate data-driven policymaking by highlighting core areas for resource focus and intervention.

Problem 2 - Data Overview

Data type of the data :

```
Shape of the dataset: (640, 61)

Data Types:
State Code        int64
Dist.Code         int64
State            object
Area Name        object
No_HH             int64
                   ...
MARG_HH_0_3_F     int64
MARG_OT_0_3_M     int64
MARG_OT_0_3_F     int64
NON_WORK_M        int64
NON_WORK_F        int64
Length: 61, dtype: object
```

```
Data columns (total 61 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   State Code      640 non-null    int64
 1   Dist.Code       640 non-null    int64
 2   State           640 non-null    object
 3   Area Name       640 non-null    object
 4   No_HH           640 non-null    int64
 5   TOT_M           640 non-null    int64
 6   TOT_F           640 non-null    int64
 7   M_06            640 non-null    int64
 8   F_06            640 non-null    int64
 9   M_SC            640 non-null    int64
 10  F_SC            640 non-null    int64
 11  M_ST            640 non-null    int64
 12  F_ST            640 non-null    int64
 13  M_LIT           640 non-null    int64
 14  F_LIT           640 non-null    int64
 15  M_ILL           640 non-null    int64
 16  F_ILL           640 non-null    int64
 17  TOT_WORK_M      640 non-null    int64
 18  TOT_WORK_F      640 non-null    int64
 19  MAINWORK_M      640 non-null    int64
 20  MAINWORK_F      640 non-null    int64
 21  MAIN_CL_M       640 non-null    int64
 22  MAIN_CL_F       640 non-null    int64
 23  MAIN_AL_M       640 non-null    int64
 24  MAIN_AL_F       640 non-null    int64
 25  MAIN_HH_M       640 non-null    int64
 26  MAIN_HH_F       640 non-null    int64
 27  MAIN_OT_M       640 non-null    int64
 28  MAIN_OT_F       640 non-null    int64
 29  MARGWORK_M      640 non-null    int64
 30  MARGWORK_F      640 non-null    int64
 31  MARG_CL_M       640 non-null    int64
 32  MARG_CL_F       640 non-null    int64
 33  MARG_AL_M       640 non-null    int64
 34  MARG_AL_F       640 non-null    int64
 35  MARG_HH_M       640 non-null    int64
 36  MARG_HH_F       640 non-null    int64
 37  MARG_OT_M       640 non-null    int64
 38  MARG_OT_F       640 non-null    int64
 39  MARGWORK_3_6_M  640 non-null    int64
 40  MARGWORK_3_6_F  640 non-null    int64
 41  MARG_CL_3_6_M   640 non-null    int64
 42  MARG_CL_3_6_F   640 non-null    int64
 43  MARG_AL_3_6_M   640 non-null    int64
 44  MARG_AL_3_6_F   640 non-null    int64
 45  MARG_HH_3_6_M   640 non-null    int64
 46  MARG_HH_3_6_F   640 non-null    int64
 47  MARG_OT_3_6_M   640 non-null    int64
 48  MARG_OT_3_6_F   640 non-null    int64
 49  MARGWORK_0_3_M  640 non-null    int64
 50  MARGWORK_0_3_F  640 non-null    int64
 51  MARG_CL_0_3_M   640 non-null    int64
 52  MARG_CL_0_3_F   640 non-null    int64
 53  MARG_AL_0_3_M   640 non-null    int64
 54  MARG_AL_0_3_F   640 non-null    int64
 55  MARG_HH_0_3_M   640 non-null    int64
 56  MARG_HH_0_3_F   640 non-null    int64
 57  MARG_OT_0_3_M   640 non-null    int64
 58  MARG_OT_0_3_F   640 non-null    int64
 59  NON_WORK_M      640 non-null    int64
 60  NON_WORK_F      640 non-null    int64
dtypes: int64(59), object(2)
```

Statistical summary of the dataset:

```
Statistical Summary:
       State Code   Dist.Code         No_HH          TOT_M          TOT_F  \
count  640.000000  640.000000    640.000000     640.000000     640.000000
mean    17.114062  320.500000  51222.871875   79940.576563  122372.084375
std      9.426486  184.896367  48135.405475   73384.511114  113600.717282
min      1.000000    1.000000    350.000000     391.000000     698.000000
25%      9.000000  160.750000  19484.000000   30228.000000   46517.750000
50%     18.000000  320.500000  35837.000000   58339.000000   87724.500000
75%     24.000000  480.250000  68892.000000  107918.500000  164251.750000
max     35.000000  640.000000 310450.000000  485417.000000  750392.000000

               M_06          F_06           M_SC           F_SC          M_ST  \
count    640.000000    640.000000     640.000000     640.000000    640.000000
mean   12309.098438  11942.300000   13820.946875   20778.392188   6191.807813
std    11500.906881  11326.294567   14426.373130   21727.887713   9912.668948
min       56.000000     56.000000       0.000000       0.000000      0.000000
25%     4733.750000   4672.250000    3466.250000    5603.250000    293.750000
50%     9159.000000   8663.000000    9591.500000   13709.000000   2333.500000
75%    16520.250000  15902.250000   19429.750000   29180.000000   7658.000000
max    96223.000000  95129.000000  103307.000000  156429.000000  96785.000000

       ...  MARG_CL_0_3_M  MARG_CL_0_3_F  MARG_AL_0_3_M  MARG_AL_0_3_F  \
count  ...     640.000000     640.000000     640.000000     640.000000
mean   ...    1392.973438    2757.050000     250.889062     558.098438
std    ...    1489.707052    2788.776676     453.336594    1117.642748
min    ...       4.000000      30.000000       0.000000       0.000000
25%    ...     489.500000     957.250000      47.000000     109.000000
50%    ...     949.000000    1928.000000     114.500000     247.500000
75%    ...    1714.000000    3599.750000     270.750000     568.750000
max    ...    9875.000000   21611.000000    5775.000000   17153.000000

       MARG_HH_0_3_M  MARG_HH_0_3_F  MARG_OT_0_3_M  MARG_OT_0_3_F  \
count     640.000000     640.000000     640.000000     640.000000
mean      560.690625    1293.431250      71.379688     200.742188
std       762.578991    1585.377936     107.897627     309.740854
min         0.000000       0.000000       0.000000       0.000000
25%       136.500000     298.000000      14.000000      43.000000
50%       308.000000     717.000000      35.000000     113.000000
75%       642.000000    1710.750000      79.000000     240.000000
max      6116.000000   13714.000000     895.000000    3354.000000


        NON_WORK_M    NON_WORK_F
count   640.000000    640.000000
mean    510.014063    704.778125
std     610.603187    910.209225
min       0.000000      5.000000
25%     161.000000    220.500000
50%     326.000000    464.500000
75%     604.500000    853.500000
max    6456.000000  10533.000000

[8 rows x 59 columns]
```

Checking for any missing values :

```
Missing Values:
 State Code       0
Dist.Code        0
State            0
Area Name        0
No_HH            0
                ..
MARG_HH_0_3_F    0
MARG_OT_0_3_M    0
MARG_OT_0_3_F    0
NON_WORK_M       0
NON_WORK_F       0
Length: 61, dtype: int64
```

- Shape of the Dataset Rows: 640 Columns: 61 Data Types The dataset consists of various data types including:
- Integer (int64): Numerical values, which are used for counts and codes. Object (string): Textual data, used for state and area names. Statistical Summary The statistical summary includes details such as count, mean, standard deviation, minimum, quartiles, and maximum values for numerical columns. Key columns include:

- No_HH (Number of Households) TOT_M (Total Males) TOT_F (Total Females) Literacy and employment-related statistics, among others. First Few Rows The first few entries provide a glimpse into the data structure, including:
- State and district codes and names. Population counts for males and females. Detailed demographic breakdowns by age, caste, and employment status.

Key Questions :

Problem 2 - 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?

Step 1:

Select Variable for EDA :

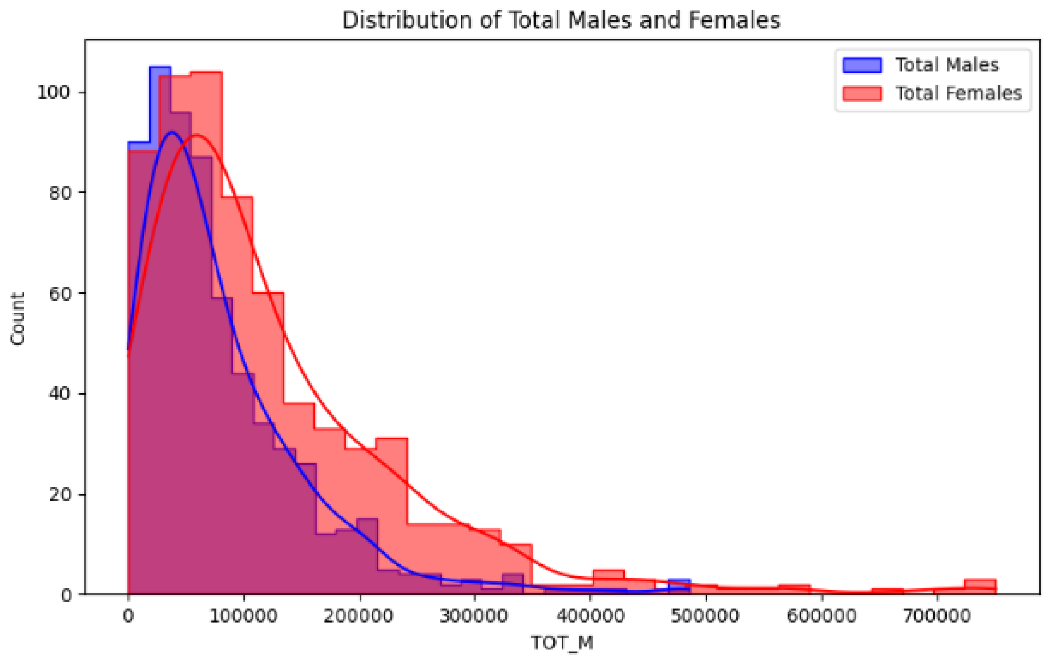
Data Overview and Preparation:

The data from the 2011 Primary Census Abstract reflects various demographic indicators for female-headed households across India, at both state and district levels. For this analysis, five key variables were chosen:

```
selected_columns = ['TOT_M', 'TOT_F', 'M_LIT', 'F_LIT', 'MAINWORK_F']
```

These variables provide insights into the household composition, gender distribution, and young male population. Before proceeding with the analysis, the dataset was examined for its structure, and necessary data types and

Histogram of Total Males and Total Females:

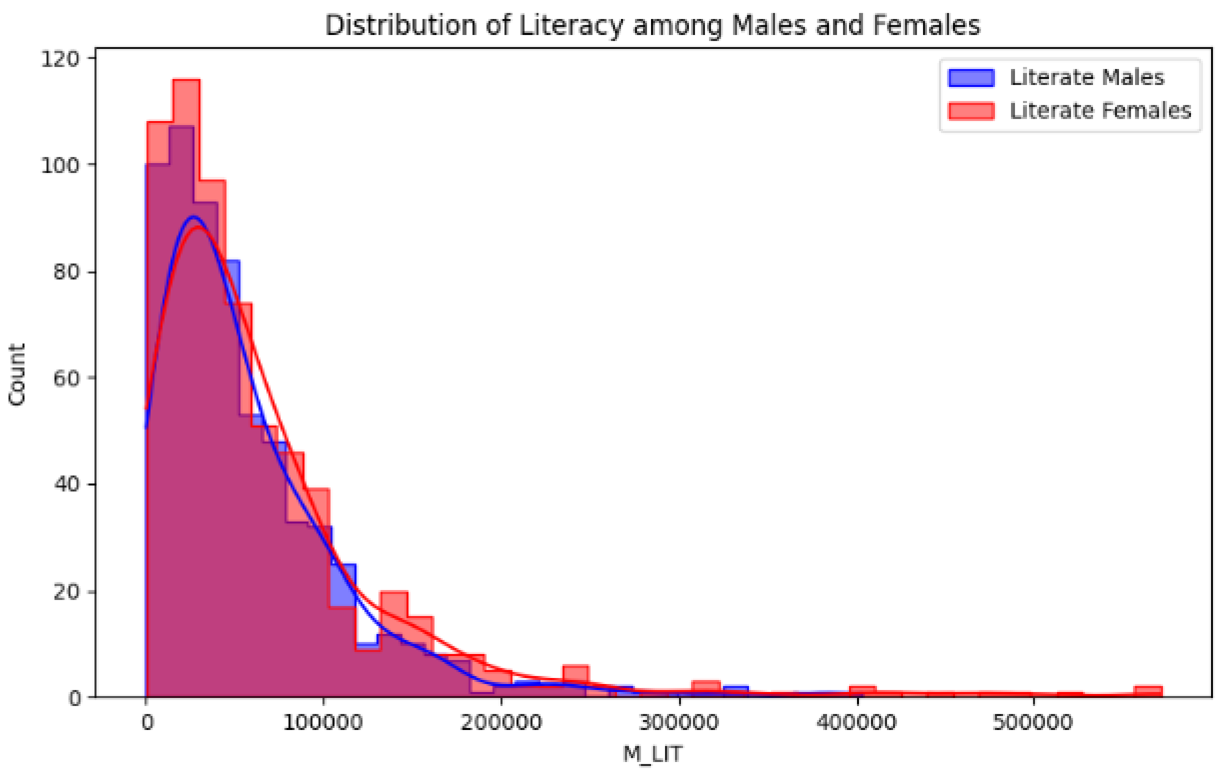


1. Distribution of Total Males and Females

Observation: The histograms overlaid with kernel density estimates show that the total female population (in blue) consistently lags behind the total male population (in red) across all counts. The distribution of both genders is right-skewed, indicating a higher frequency of smaller population counts.

Insight: This disparity suggests gender imbalances in the population across different regions or groups within the dataset.

Histogram of Literacy among Males and Females

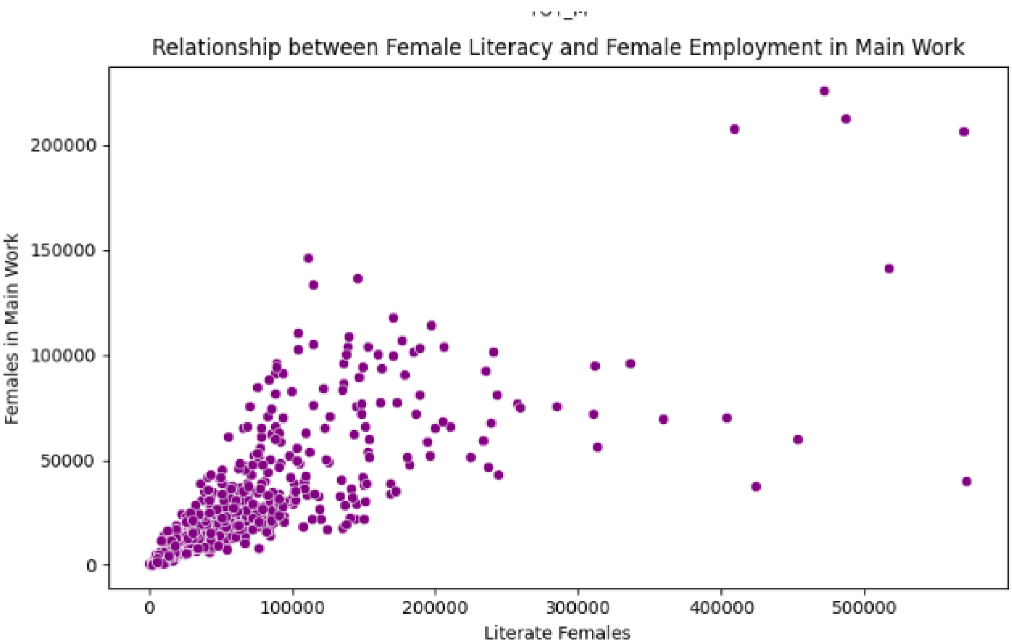


2. Distribution of Literacy among Males and Females

Observation: Similar to the total population distribution, literate females are fewer than literate males across various counts. Both distributions are again right-skewed**.**

Insight: There is a clear gender gap in literacy rates, which highlights areas for potential educational interventions aimed at females**.**

Scatter plot to show relationship between Female Literacy and Female Main Work:

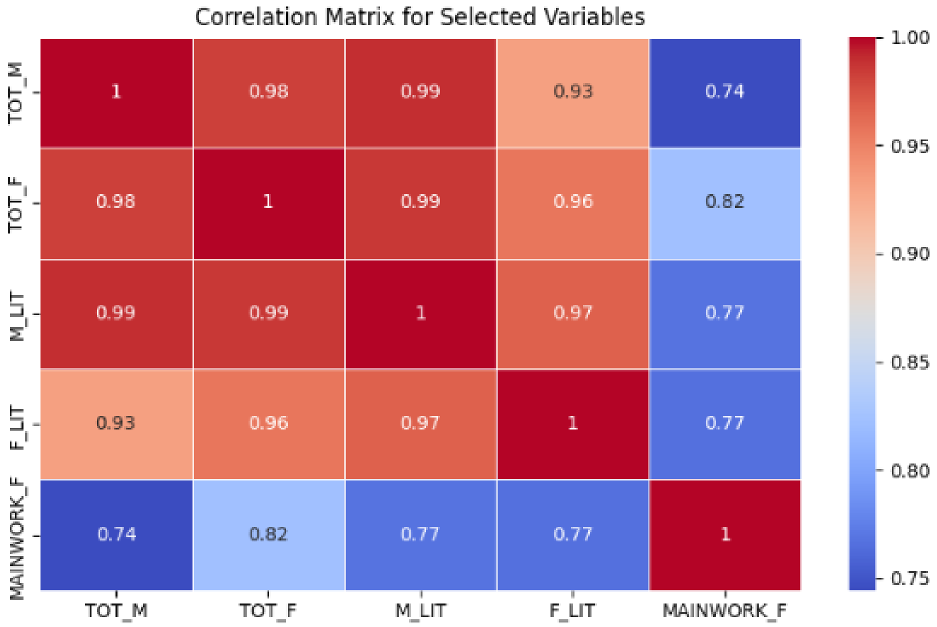


3. Relationship between Female Literacy and Female Employment in Main Work

Observation: The scatter plot shows a wide spread of data points without a clear linear trend. Higher counts of literate females do not consistently correspond to proportionally high female employment, indicating a complex relationship with potentially many influencing factors.

Insight: While education is generally seen as a pathway to employment, this plot suggests that for females, increased literacy does not straightforwardly translate into higher employment rates in main work. Social, cultural, or economic barriers might be

Calculating correlation only for selected data:



4. Correlation Matrix for Selected Variables

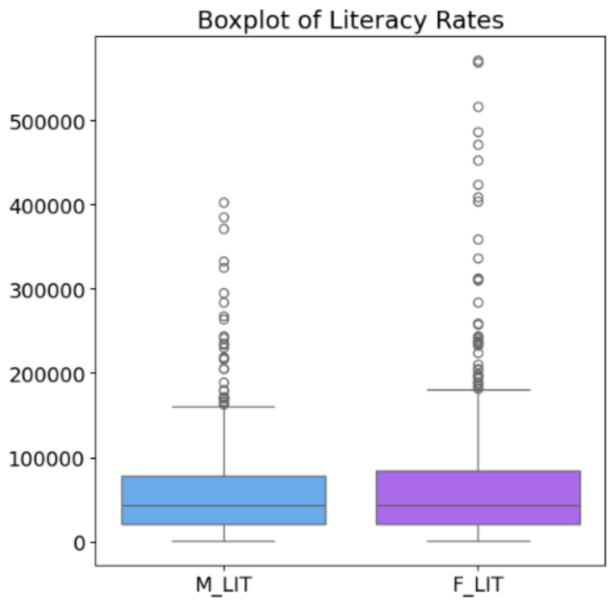
Observation:

Total males (TOT_M) and females (TOT_F) and literate males (M_LIT) and females (F_LIT) show very high correlations (above 0.90), suggesting that regions with higher populations tend to have higher numbers of literate individuals irrespective of gender.

Female literacy (F_LIT) and female employment in main work (MAINWORK_F) have a correlation of 0.77, indicating a moderately strong positive relationship, but not as strong as might be expected.

Insight: The high correlations between total and literate counts confirm the dependency of literacy levels on the population size. The moderate correlation between female literacy and employment suggests other factors might be playing significant roles in determining employment rates beyond just education levels.

Boxplots for Literacy and Employment:



Observations: Comparison of Male vs Female Literacy Rates (M_LIT and F_LIT)

Male Literacy (M_LIT): The distribution is fairly concentrated with fewer outliers, suggesting a more consistent literacy rate among males across different regions or groups**.**

Female Literacy (F_LIT): The median literacy rate for females is lower than for males, indicating a disparity in literacy rates. The female literacy rates also have a wider interquartile range and a larger number of outliers, suggesting greater variability and some regions or groups with exceptionally low or high literacy rates compared to the median.

Insight: The disparity in literacy rates between genders highlights potential areas for targeted educational initiatives to improve literacy among females.

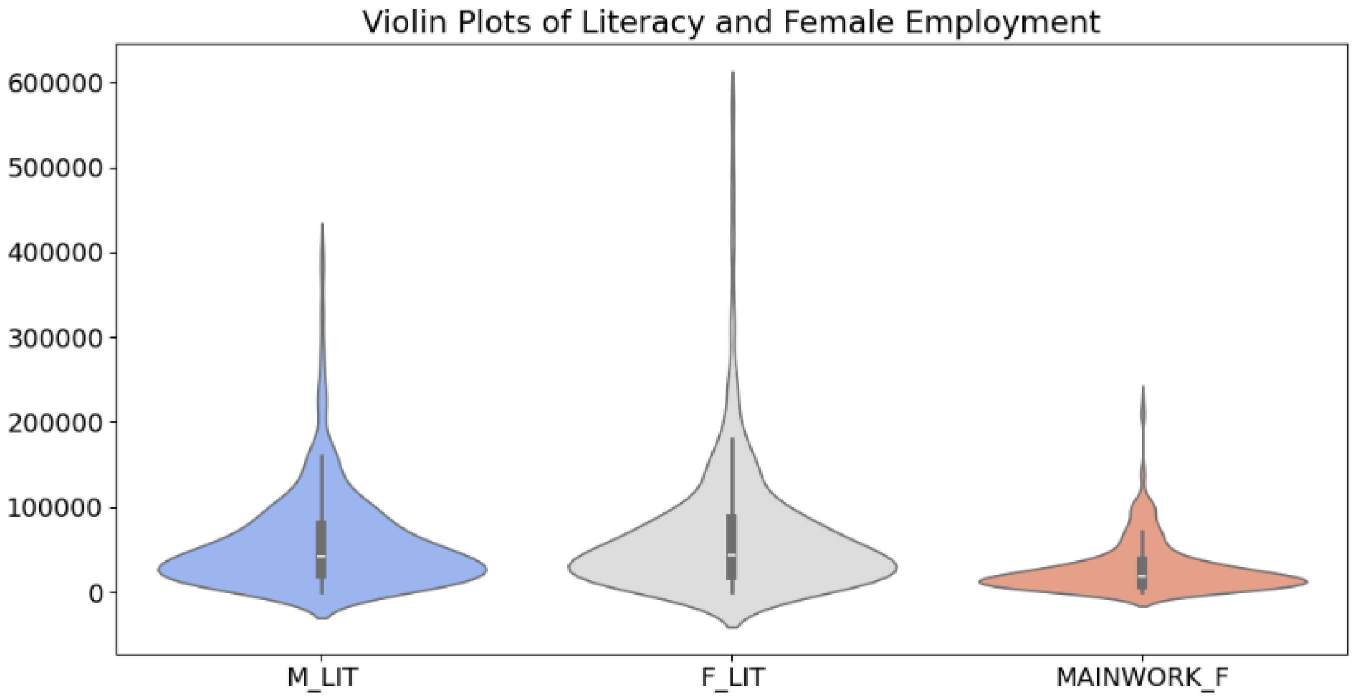
Observation :Female Employment Rates (MAINWORK_F)

The boxplot shows a wide range of values with a substantial number of outliers, indicating that while many females are employed within a typical range, there are significant exceptions where employment rates are exceedingly high or low.

The median is relatively low compared to the range of the data, suggesting that a significant proportion of the female population may have low employment rates.

Insight: The spread and the outliers suggest diverse conditions affecting female employment across different regions or groups. This variability might be due to economic, cultural, or legislative differences influencing women's participation in the workforce.

Violin plots for Literacy and Female Employment:



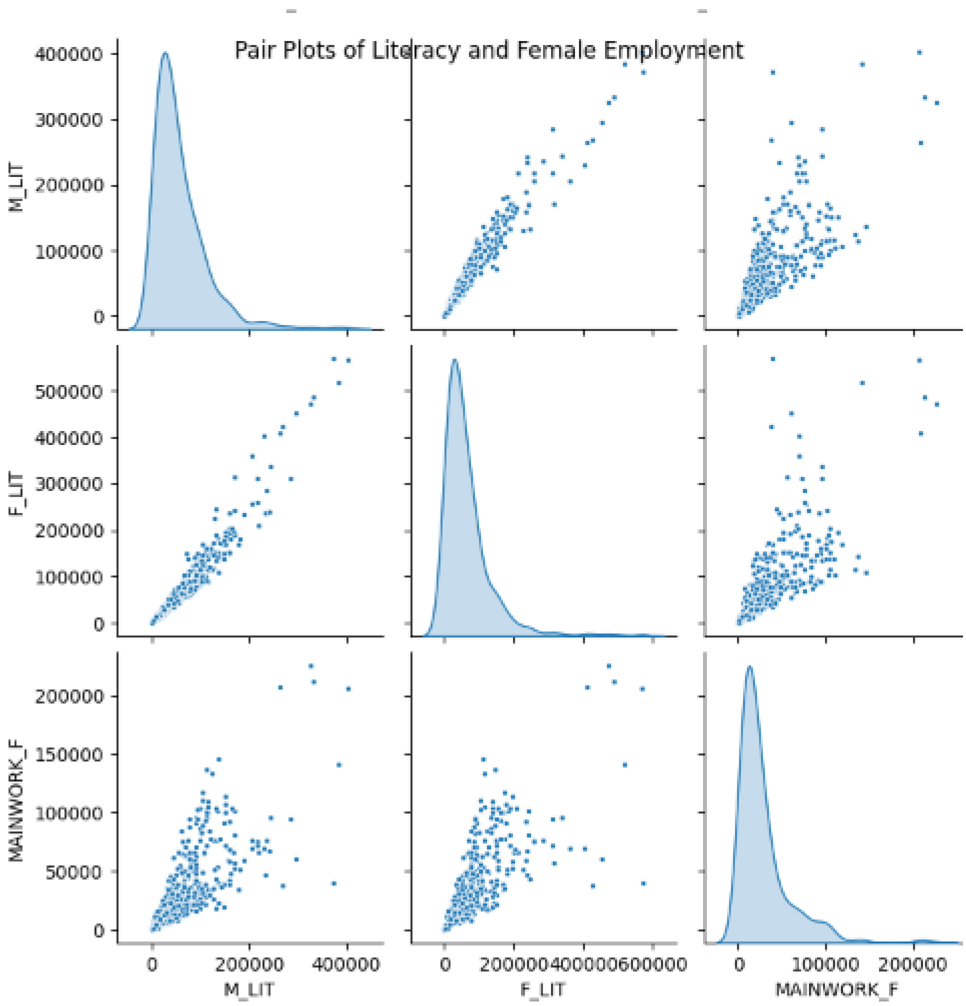
Observation :

Male Literacy (M_LIT): Displays a broad base with a peak at lower literacy counts, indicating a wide range of literacy levels with a skew towards the lower end.

Female Literacy (F_LIT): More symmetrically shaped, suggesting a more uniform distribution of literacy rates across different counts. The plot shows less skew and a centrally located median, indicating balanced literacy rates among females.

Female Employment (MAINWORK_F): Exhibits a narrow shape with a long upper tail, reflecting generally low employment rates among females but with some outliers experiencing significantly higher employment. The median is low, emphasizing the overall lower employment rates.

Pair plots for the selected variables:



Observation :

M_LIT vs F_LIT: Displays a linear relationship, suggesting that regions with higher male literacy rates also tend to have higher female literacy rates. The correlation appears strong, indicating that literacy efforts in a region likely benefit both genders.

M_LIT vs MAINWORK_F: Shows a scattered and less defined relationship, indicating that higher male literacy does not necessarily correlate strongly with higher female employment rates**.**

F_LIT vs MAINWORK_F: Similarly scattered, this plot shows no clear trend between female literacy and employment rates, suggesting other factors might influence the employment rates beyond literacy.

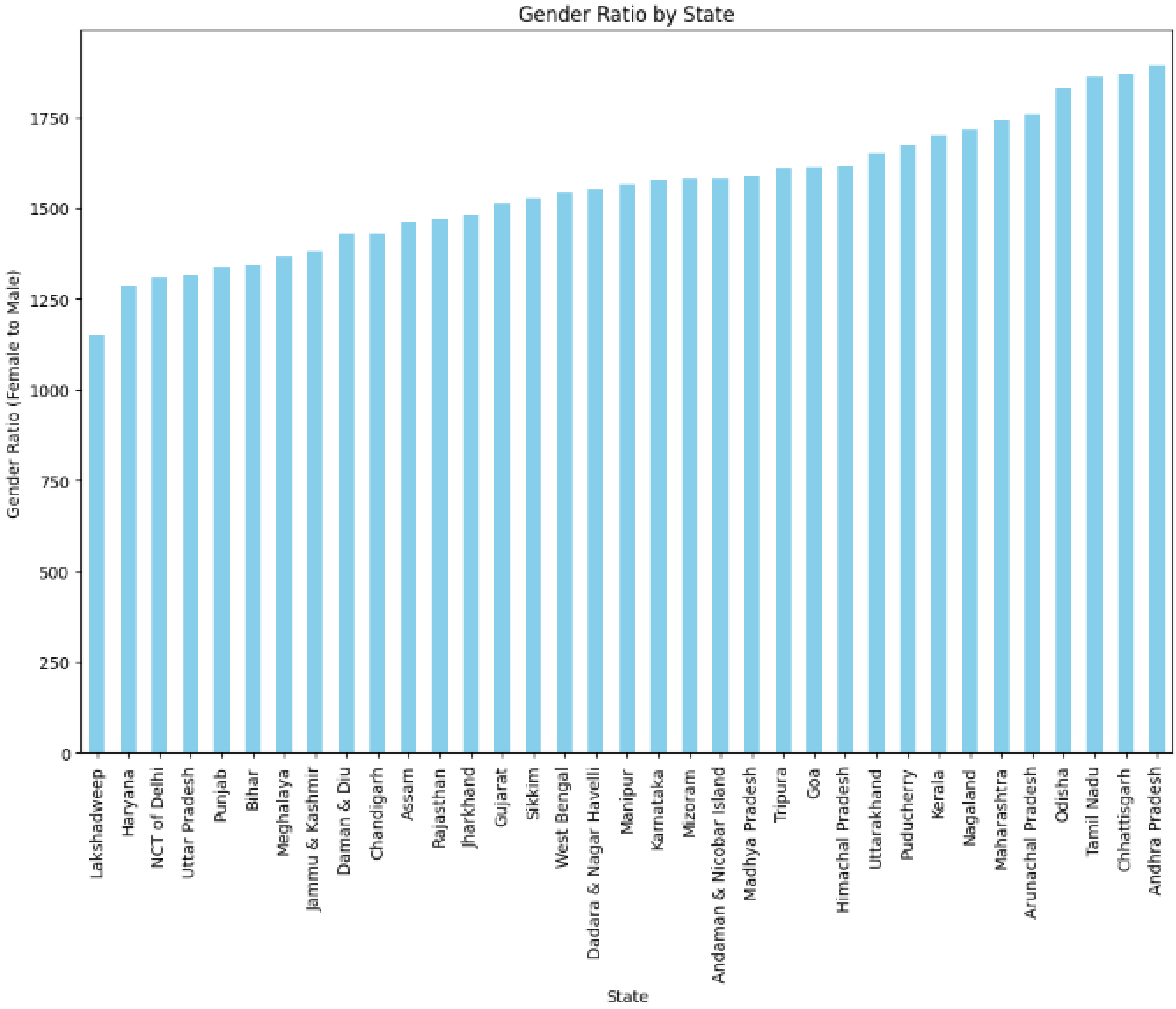
2. Example questions to answer from EDA - (i) Which state has highest gender ratio and which has the lowest? (ii) Which district has the highest & lowest gender ratio?

A critical component of the analysis was calculating and examining the Gender_Ratio. This ratio is essential for assessing the balance between male and female populations within the context of female-headed households. The histogram for the gender ratio helped identify the general tendency of the gender distribution, highlighting regions with potential gender imbalances.

State with the highest gender ratio: Lakshadweep State with the lowest gender ratio: Andhra Pradesh District with the highest gender ratio: Lakshadweep District with the lowest gender ratio: Krishna

```
Highest gender ratio state: ('Andhra Pradesh', 1895.093129626215)
Lowest gender ratio state: ('Lakshadweep', 1151.9925134523903)
Highest gender ratio district: (('Andhra Pradesh', 'Krishna'), 2283.24963845265)
Lowest gender ratio district: (('Lakshadweep', 'Lakshadweep'), 1151.9925134523903)
```

From below Plot we can easily find the highest and lowest gender ratios in a state:



Conclusions and Insights:

The exploratory data analysis provided essential insights into the demographic structure of female-headed households in India. Key findings included the identification of states and districts with significant gender imbalances, which could be targets for specific social policies or further research. The analysis also highlighted the importance of focusing on early childhood demographics as part of broader demographic studies.

Problem 2 - 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

Step 1:

Check for and treat (if needed)missing values :

```
Missing values per column:
State Code       0
Dist.Code        0
State            0
Area Name        0
No_HH            0
                ..
MARG_OT_0_3_M    0
MARG_OT_0_3_F    0
NON_WORK_M       0
NON_WORK_F       0
Gender_Ratio     0
Length: 62, dtype: int64
```

There are no missing values .

Step 2:

Check for and treat(if needed) data irrgularies

```
Negative Values:
 No_HH             0
TOT_M             0
TOT_F             0
M_06              0
F_06              0
M_SC              0
F_SC              0
M_ST              0
F_ST              0
M_LIT             0
F_LIT             0
M_ILL             0
F_ILL             0
TOT_WORK_M        0
TOT_WORK_F        0
MAINWORK_M        0
MAINWORK_F        0
MAIN_CL_M         0
MAIN_CL_F         0
MAIN_AL_M         0
MAIN_AL_F         0
MAIN_HH_M         0
MAIN_HH_F         0
MAIN_OT_M         0
MAIN_OT_F         0
MARGWORK_M        0
MARGWORK_F        0
MARG_CL_M         0
MARG_CL_F         0
MARG_AL_M         0
MARG_AL_F         0
MARG_HH_M         0
MARG_HH_F         0
MARG_OT_M         0
MARG_OT_F         0
MARGWORK_3_6_M    0
MARGWORK_3_6_F    0
MARG_CL_3_6_M     0
```

```
MARG_CL_3_6_F      0
MARG_AL_3_6_M      0
MARG_AL_3_6_F      0
MARG_HH_3_6_M      0
MARG_HH_3_6_F      0
MARG_OT_3_6_M      0
MARG_OT_3_6_F      0
MARGWORK_0_3_M     0
MARGWORK_0_3_F     0
MARG_CL_0_3_M      0
MARG_CL_0_3_F      0
MARG_AL_0_3_M      0
MARG_AL_0_3_F      0
MARG_HH_0_3_M      0
MARG_HH_0_3_F      0
MARG_OT_0_3_M      0
MARG_OT_0_3_F      0
NON_WORK_M         0
NON_WORK_F         0
```

There are no negative values present in dataset.

Step 3: Scaling the Data Using the Z-Score Method

Objective: Normalize data features to have a mean of zero and a standard deviation of one. This is particularly important in analyses where distance measures are used (like PCA), ensuring that all features contribute equally without bias due to their scale.

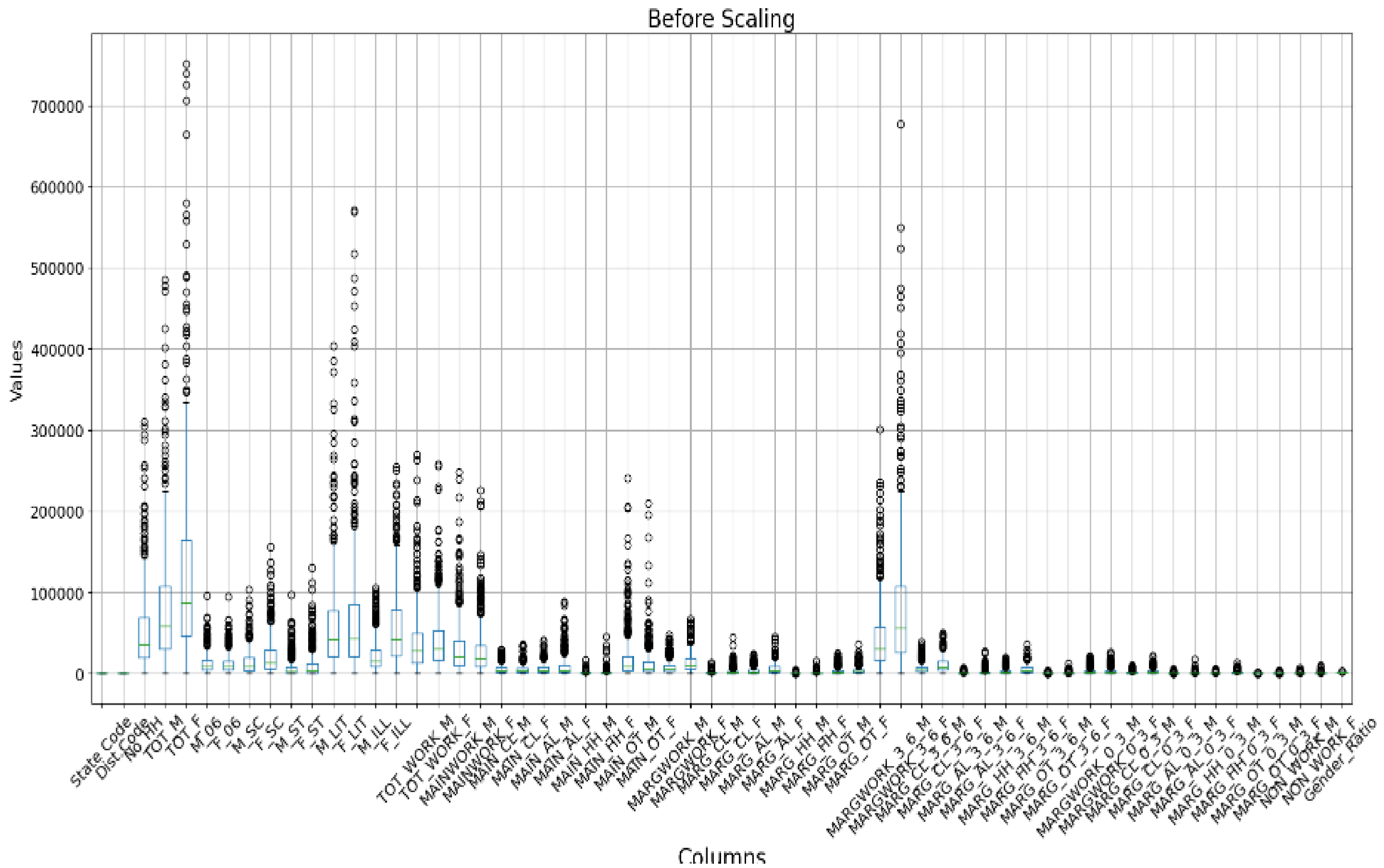
Methodology:

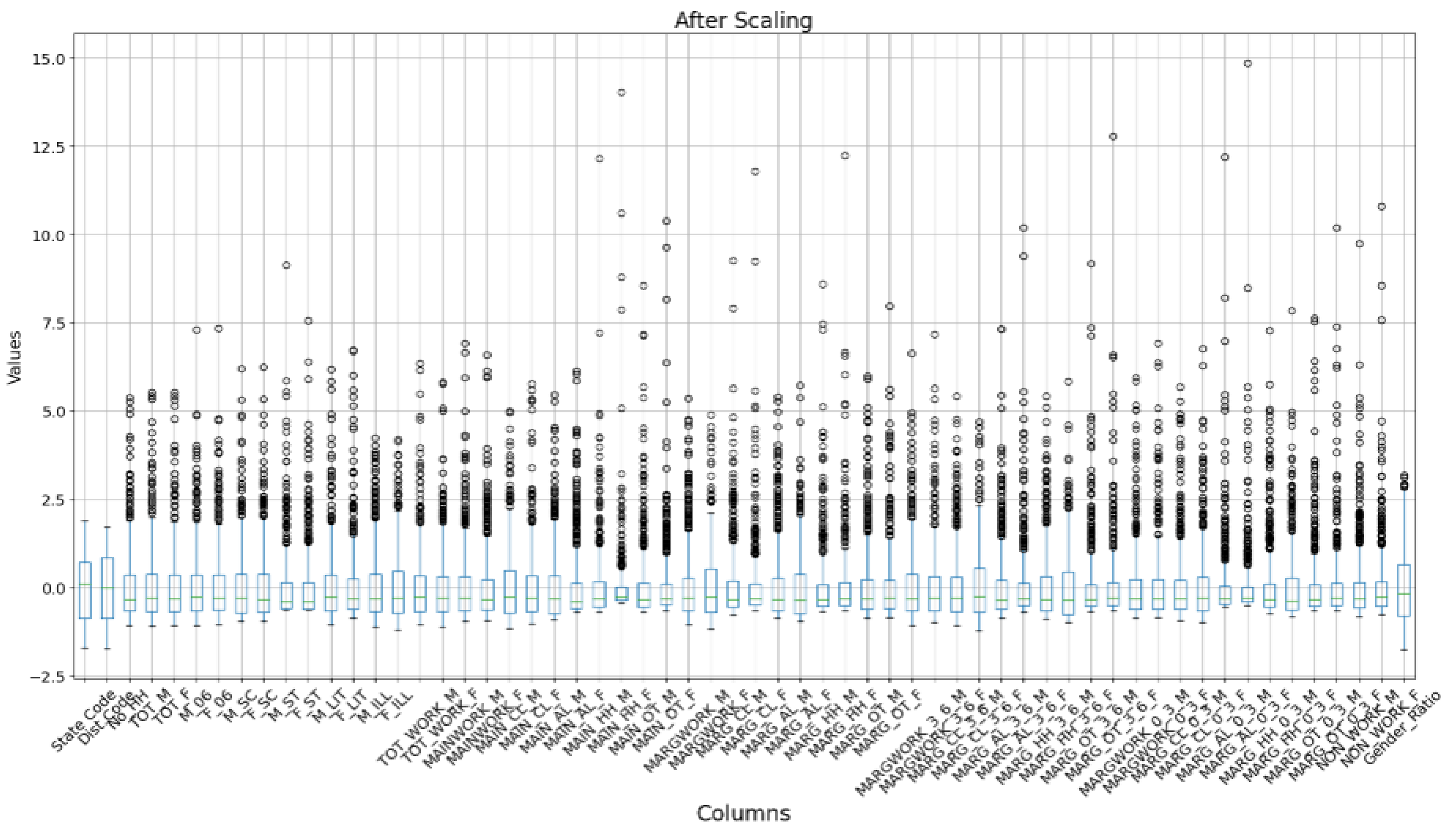
Z-Score Normalization: Apply the Z-score method using StandardScaler from Scikit-learn. This technique subtracts the mean and divides by the standard deviation for each data point.

Step 4: Visualizing the Data Before and After Scaling

Objective: Assess the impact of scaling on data distributions and observe changes particularly concerning the handling of outliers.

Visualize data before scaling :





comment on the impact on outliers:

1. Scale of Values

Before Scaling: The values across various columns vary widely, ranging from nearly zero to over

700,000. This wide range indicates that the dataset contains features with vastly different scales, which can be problematic for many machine learning algorithms that are sensitive to the scale of input features.

After Scaling: In the scaled data, all values are normalized to a more uniform scale, roughly between -2.5 and 15. This uniformity is crucial for many algorithms, particularly those that use distance calculations, as it ensures that all features contribute equally to the result.

2. Presence and Visibility of Outliers

Before Scaling: Outliers are present in several columns, and their impact is pronounced due to the large scale of values. These outliers can significantly affect the mean and standard deviation of the respective columns, potentially leading to misleading analysis.

After Scaling: Outliers remain visible but are less extreme compared to the unscaled data. The scaling process has reduced their relative impact by bringing them closer to the other data points. However, the persistence of these outliers suggests that the scaling method used might not be robust against outliers (such as simple min-max scaling or standard normalization).

3. Distribution and Spread of Data

Before Scaling: The spread of the data in many columns is large, and the differences between the minimum and maximum values are substantial. This variability can overshadow the contributions of features with smaller ranges when using certain algorithms.

After Scaling: The spread of data in each column is more controlled and uniform. The interquartile ranges (the boxes in the boxplots) are more consistent across features, indicating a more uniform distribution of data after scaling. This consistency helps in analytical models to treat all features with equal importance.

4. Impact on Machine Learning and Statistical Analysis

Before Scaling: The wide disparity in ranges could lead to biased or inefficient learning in machine learning models, where algorithms might unduly emphasize features with broader ranges.

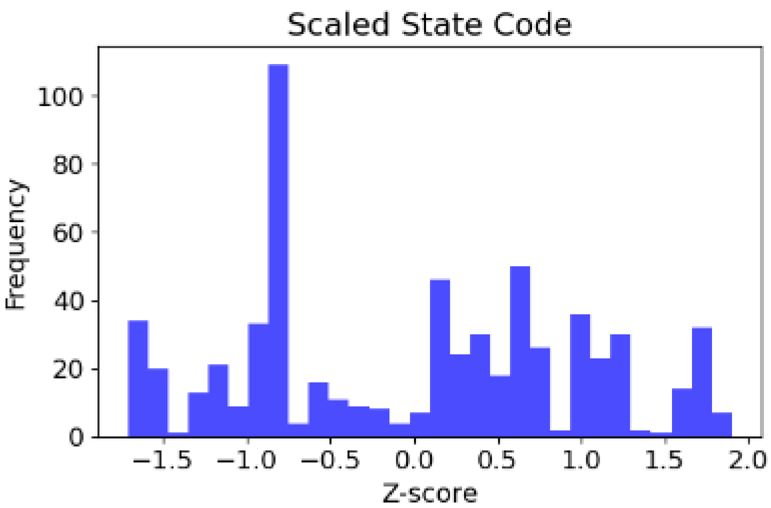
After Scaling: The normalization helps in mitigating this issue, making the dataset more suitable for a wide range of statistical analyses and machine learning models, particularly those involving distance measures like k-nearest neighbors or clustering algorithms.

Numerical data :

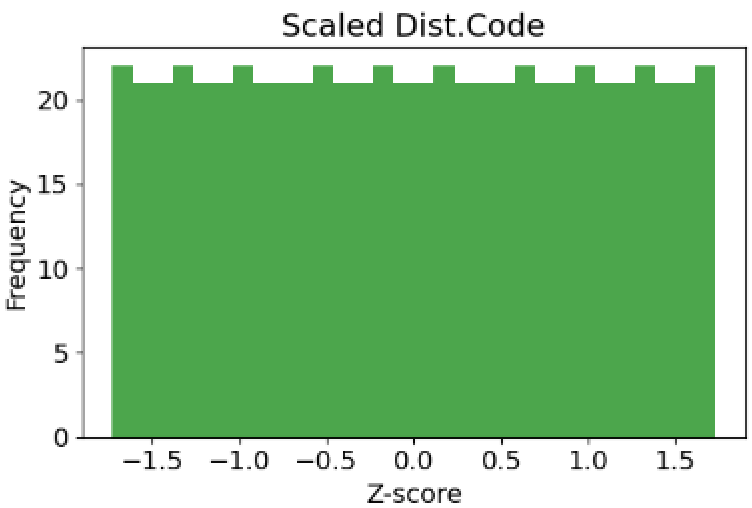
	State Code	Dist.Code	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	...	MARG_CL_0_3_F	MARG_AL_0_3_M	MARG_AL_0_3_F	MARG_HH_0_3_M	MARG_HH_0_3_F	MARG_OT_0_3_M	MARG_OT_0_3_F	NON_WORK_M	NON_WORK_F	Gender_Ratio
0	1	1	7707	23388	29796	5862	6196	3	0	1999	...	749	180	237	680	252	32	46	258	214	1273.986660
1	1	2	6218	19585	23102	4482	3733	7	6	427	...	715	123	229	186	148	76	178	140	160	1179.576206
2	1	3	4452	6546	10964	1082	1018	3	6	5806	...	188	44	89	3	34	0	4	67	61	1674.915979
3	1	4	1320	2784	4206	563	677	0	0	2666	...	247	61	128	13	50	4	10	116	59	1510.775862
4	1	5	11654	20591	29981	5157	4587	20	33	7670	...	1928	465	1043	205	302	24	105	180	478	1456.024477
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
635	34	636	3333	8154	11781	1146	1203	21	30	0	...	47	0	0	0	0	0	0	32	47	1444.812362
636	34	637	10612	12346	21691	1544	1533	2234	4155	0	...	337	3	14	38	130	4	23	110	170	1756.925320
637	35	638	1275	1549	2630	227	225	0	0	1012	...	134	9	4	2	6	17	47	76	77	1697.869593
638	35	639	3762	5200	8012	723	664	0	0	28	...	172	24	44	11	21	1	4	100	103	1540.769231
639	35	640	7975	11977	18049	1470	1358	0	0	161	...	122	6	2	17	17	2	4	148	99	1506.971696

640 rows × 60 columns

Visualizing scaled columns:

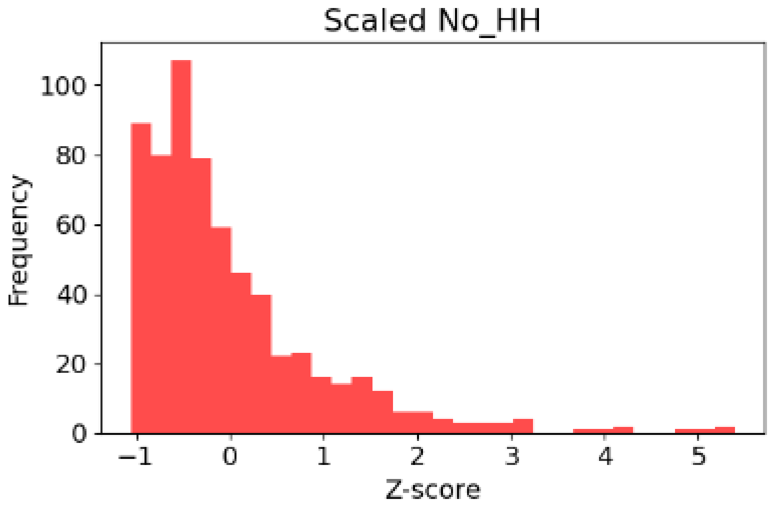


Observation :
Distribution: Shows multiple peaks and valleys, suggesting that certain state codes occur more frequently than others. This variability indicates that the data may encompass regions with differing characteristics or sizes.
Insight: The uneven distribution could influence analyses that rely on geographical representation or allocation.



Observation :
Distribution: Nearly uniform across the Z-score range, indicating a balanced representation of district codes in the dataset.
Insight: The uniformity suggests that the data might be well-distributed across different districts,

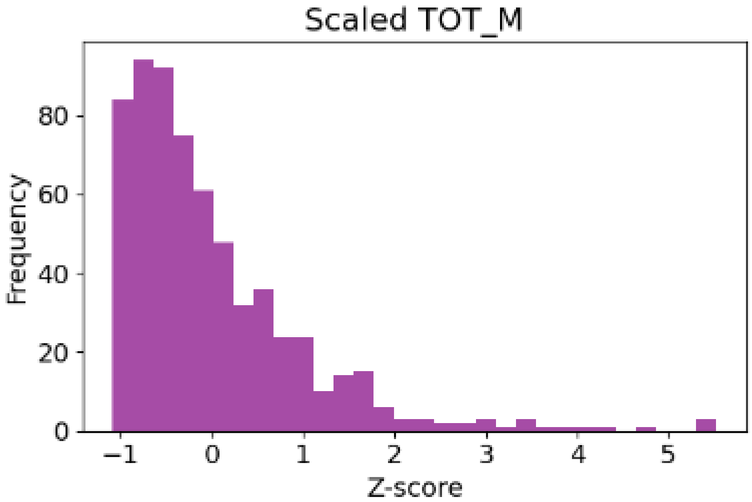
which is beneficial for analysis that requires broad geographic coverage.



Observation :

Distribution: Skewed right, with a steep drop-off as Z-scores increase. Most values cluster below the mean, indicating that smaller household numbers are more common than larger ones.

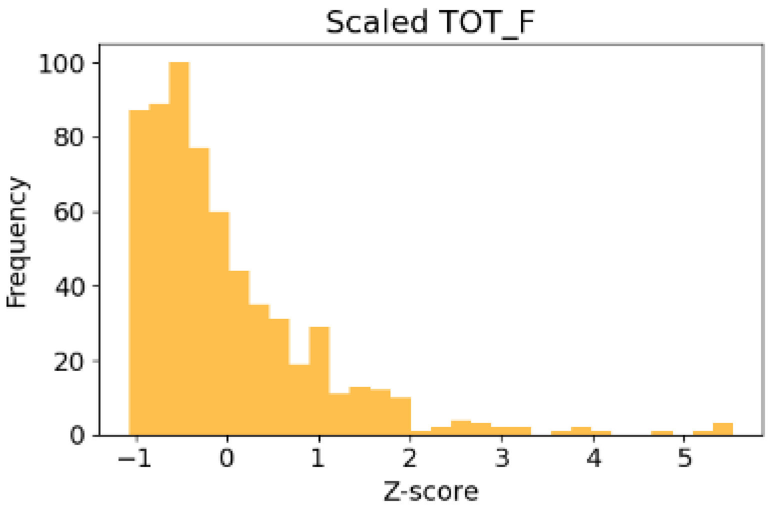
Insight: The skewness towards smaller numbers suggests that areas with fewer households dominate the dataset. This might reflect a rural bias or areas with lower population densities.



Observation :

Distribution: Skewed right, similar to the number of households, with most data points below the mean. The distribution tapers off less sharply compared to No_HH.

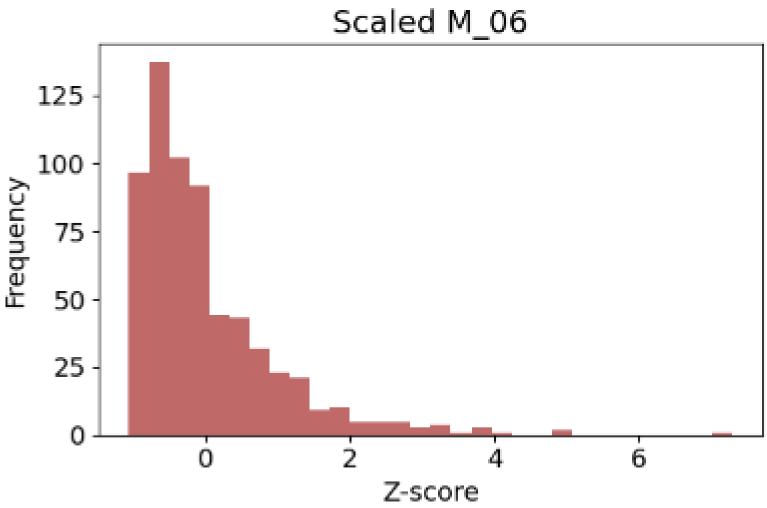
Insight: This skew suggests that regions with fewer males are more common in the dataset. The distribution's shape might affect analyses related to gender-specific policies or programs.



Observation :

Distribution: Also right-skewed, indicating that regions with fewer females are more prevalent.

Insight: The consistent right skewness in both male and female distributions suggests a general pattern of smaller population sizes across the sampled areas.



Observation :

\Distribution: Right-skewed, with a high frequency of lower values and a long tail extending into higher Z-scores.

Insight: The shape of this distribution might indicate a variable that captures a characteristic or metric that varies widely but typically registers at lower levels (e.g., a specific demographic feature, employment rate, etc.).

Problem 2 - 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

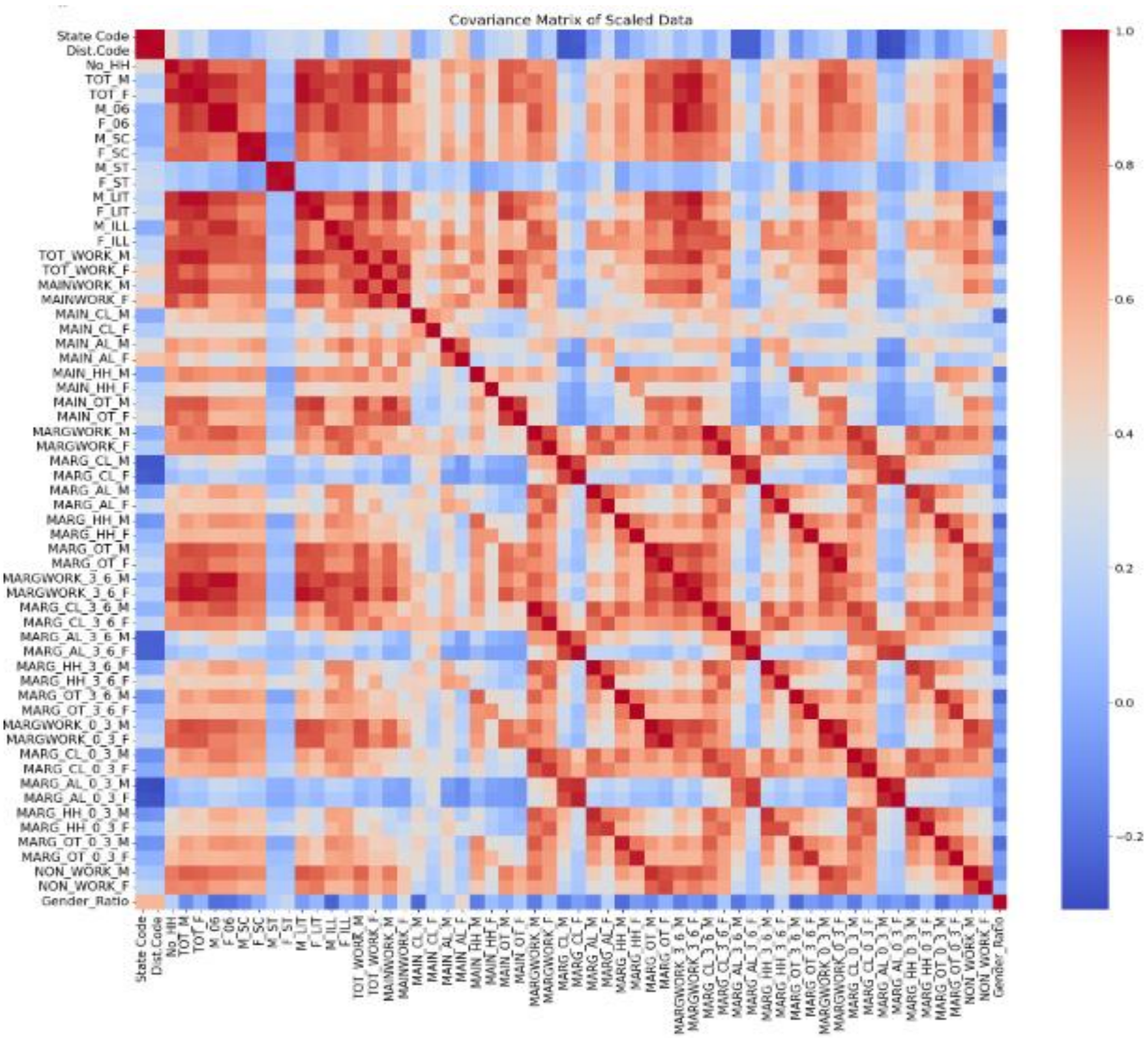
Step 1:

Create the covariant matrix:

To create covariance matrix we need to exclude non numeric columns. Then we created a covariance matrix below:

```
Covariance Matrix:
[[ 1.00156495  0.99457535  0.38502614 ...  0.12572474  0.23208471
   0.56705561]
 [ 0.99457535  1.00156495  0.37756089 ...  0.11226784  0.21313518
   0.56677222]
 [ 0.38502614  0.37756089  1.00156495 ...  0.76357722  0.73684378
   0.19172282]
 ...
 [ 0.12572474  0.11226784  0.76357722 ...  1.00156495  0.88228018
  -0.09944555]
 [ 0.23208471  0.21313518  0.73684378 ...  0.88228018  1.00156495
   0.07963851]
 [ 0.56705561  0.56677222  0.19172282 ... -0.09944555  0.07963851
   1.00156495]]
```

Covariance matrix of scaled data :



Observation and Insights from above Heatmap :

- High Correlation Areas (Reddish Tones):There are several clusters of variables showing strong positive correlations, indicated by darker red colors. These typically involve variables within similar categories, such as those related to employment or demographic metrics, suggesting that related features tend to increase or decrease together.For example, variables related to male and female employment (like TOT_WORK_M, MAINWORK_F, etc.) show strong correlations with each other, implying that regions with high male employment also tend to have high female employment.

- Low or Negative Correlation Areas (Bluish Tones):Certain variables exhibit low or negative correlations, highlighted by bluish tones. This might indicate that as one variable increases, the other decreases, or they are simply unrelated.
- The gender-specific variables (like TOT_M vs. F_LIT or TOT_F vs. M_LIT) could display such patterns, suggesting different trends or impacts in population demographics and literacy between genders.
- Neutral Correlations (Whitish Tones):Some variables do not show strong correlations with others, indicated by neutral (whitish) colors. These variables might operate independently of others, or their influences are not directly observable in the context of other measured metrics.

Key Insights:

- Interdependency of Employment Metrics: The strong correlations within employment-related variables across genders suggest that improvements or declines in employment conditions are generally shared across male and female demographics.
- Demographic Influence: High correlations between demographic variables (like total males/females and literacy rates) indicate that demographic shifts could significantly influence literacy and employment statistics.Potential for Multivariate Analysis: The complex interrelations visible in the heatmap suggest that multivariate analysis could be beneficial to untangle the impacts of various factors on each other. Understanding these relationships could aid in designing targeted interventions or policies.

Step 2: **Get eigen values and eigen vectors**

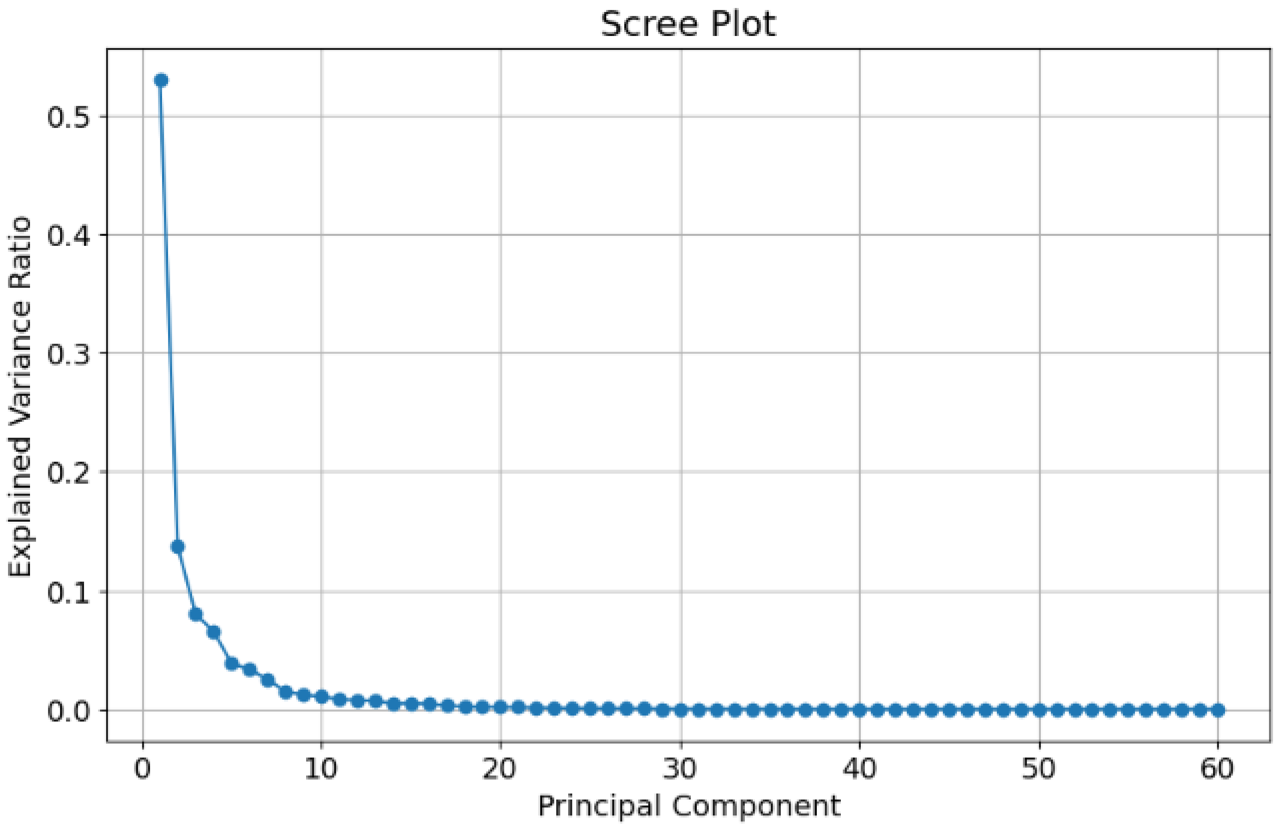
```
Eigenvalues:
[ 3.18688885e+01+0.00000000e+00j  8.22917104e+00+0.00000000e+00j
  4.80722451e+00+0.00000000e+00j  3.92901693e+00+0.00000000e+00j
  2.31829290e+00+0.00000000e+00j  2.00389746e+00+0.00000000e+00j
  1.52046849e+00+0.00000000e+00j  8.89337793e-01+0.00000000e+00j
  7.30596615e-01+0.00000000e+00j  6.28300921e-01+0.00000000e+00j
  4.95151133e-01+0.00000000e+00j  4.53516495e-01+0.00000000e+00j
  4.17512850e-01+0.00000000e+00j  2.80284774e-01+0.00000000e+00j
  2.96135603e-01+0.00000000e+00j  2.57748460e-01+0.00000000e+00j
  1.82482899e-01+0.00000000e+00j  1.27431879e-01+0.00000000e+00j
  1.11534941e-01+0.00000000e+00j  1.02255494e-01+0.00000000e+00j
  9.48312111e-02+0.00000000e+00j  7.77624132e-02+0.00000000e+00j
  5.55762931e-02+0.00000000e+00j  4.19336464e-02+0.00000000e+00j
  3.27995670e-02+0.00000000e+00j  2.96700392e-02+0.00000000e+00j
  2.64047652e-02+0.00000000e+00j  2.27207823e-02+0.00000000e+00j
  1.43594540e-02+0.00000000e+00j  1.10850662e-02+0.00000000e+00j
  9.18636917e-03+0.00000000e+00j  7.69979211e-03+0.00000000e+00j
  6.88757582e-03+0.00000000e+00j  4.99395510e-03+0.00000000e+00j
  4.48449218e-03+0.00000000e+00j  2.49323970e-03+0.00000000e+00j
  1.05896196e-03+0.00000000e+00j  6.99352403e-04+0.00000000e+00j
 -1.64976615e-15+0.00000000e+00j  1.70884644e-15+0.00000000e+00j
 -1.26765855e-15+0.00000000e+00j  1.39114525e-15+0.00000000e+00j
 -1.02322531e-15+0.00000000e+00j  1.19753500e-15+0.00000000e+00j
 -8.32677510e-16+0.00000000e+00j -7.04747786e-16+0.00000000e+00j
  9.56849148e-16+0.00000000e+00j  8.57266760e-16+0.00000000e+00j
  7.42855168e-16+0.00000000e+00j  7.12831253e-16+0.00000000e+00j
  6.23676133e-16+0.00000000e+00j -5.58087177e-16+0.00000000e+00j
 -4.11464837e-16+0.00000000e+00j  3.96348256e-16+0.00000000e+00j
 -2.01569429e-16+0.00000000e+00j -1.44894920e-16+0.00000000e+00j
 -2.79876693e-17+0.00000000e+00j  9.08136792e-17+1.27861114e-17j
  9.08136792e-17-1.27861114e-17j  9.61995197e-17+0.00000000e+00j]
```

```
Eigenvectors:
[[-2.99260021e-02+0.00000000e+00j  1.70567300e-01+0.00000000e+00j
   2.66084452e-01+0.00000000e+00j ... -5.56578943e-14+1.23830293e-14j
  -5.56578943e-14-1.23830293e-14j  5.66830662e-14+0.00000000e+00j]
 [-2.99312407e-02+0.00000000e+00j  1.66672787e-01+0.00000000e+00j
   2.73148500e-01+0.00000000e+00j ...  5.62206732e-14-1.28668170e-14j
   5.62206732e-14+1.28668170e-14j -5.71864540e-14+0.00000000e+00j]
 [-1.56371936e-01+0.00000000e+00j  1.30373389e-01+0.00000000e+00j
   4.75210382e-02+0.00000000e+00j ...  1.51140877e-13-3.96066257e-14j
   1.51140877e-13+3.96066257e-14j -1.61890516e-13+0.00000000e+00j]
 ...
 [-1.50232530e-01+0.00000000e+00j  5.20994293e-02+0.00000000e+00j
  -1.23278132e-01+0.00000000e+00j ...  1.68875813e-01-3.51770567e-02j
   1.68875813e-01+3.51770567e-02j -1.76039076e-01+0.00000000e+00j]
 [-1.31150713e-01+0.00000000e+00j  6.95799331e-02+0.00000000e+00j
  -6.68971434e-02+0.00000000e+00j ... -5.07724338e-02+1.41743236e-02j
  -5.07724338e-02-1.41743236e-02j  5.19000563e-02+0.00000000e+00j]
 [ 6.89022956e-03+0.00000000e+00j  7.62787267e-02+0.00000000e+00j
   2.64571148e-01+0.00000000e+00j ...  3.72578284e-15-4.90020993e-16j
   3.72578284e-15+4.90020993e-16j -4.27532842e-15+0.00000000e+00j]]
```

Step 3 : **Identify the optimum number of PCs**

```
Cumulative Explained Variance Ratio:
[0.53031822+0.00000000e+00j 0.66725677+0.00000000e+00j
 0.74725199+0.00000000e+00j 0.81263329+0.00000000e+00j
 0.85121113+0.00000000e+00j 0.88455724+0.00000000e+00j
 0.90985879+0.00000000e+00j 0.92465792+0.00000000e+00j
 0.93681551+0.00000000e+00j 0.94727083+0.00000000e+00j
 0.95551045+0.00000000e+00j 0.96305725+0.00000000e+00j
 0.97000492+0.00000000e+00j 0.97466904+0.00000000e+00j
 0.97959692+0.00000000e+00j 0.98388601+0.00000000e+00j
 0.98692264+0.00000000e+00j 0.98904319+0.00000000e+00j
 0.9908992 +0.00000000e+00j 0.9926008 +0.00000000e+00j
 0.99417885+0.00000000e+00j 0.99547286+0.00000000e+00j
 0.99639769+0.00000000e+00j 0.99709549+0.00000000e+00j
 0.99764129+0.00000000e+00j 0.99813502+0.00000000e+00j
 0.99857441+0.00000000e+00j 0.9989525 +0.00000000e+00j
 0.99919145+0.00000000e+00j 0.99937591+0.00000000e+00j
 0.99952878+0.00000000e+00j 0.99965691+0.00000000e+00j
 0.99977152+0.00000000e+00j 0.99985463+0.00000000e+00j
 0.99992925+0.00000000e+00j 0.99997074+0.00000000e+00j
 0.99998836+0.00000000e+00j 1.        +0.00000000e+00j
 1.        +0.00000000e+00j 1.        +0.00000000e+00j
 1.        +0.00000000e+00j 1.        +0.00000000e+00j
 1.        +0.00000000e+00j 1.        +0.00000000e+00j
 1.        +0.00000000e+00j 1.        +0.00000000e+00j
 1.        +0.00000000e+00j 1.        +0.00000000e+00j
 1.        +0.00000000e+00j 1.        +0.00000000e+00j
 1.        +0.00000000e+00j 1.        +0.00000000e+00j
 1.        +0.00000000e+00j 1.        +0.00000000e+00j
 1.        +0.00000000e+00j 1.        +0.00000000e+00j
 1.        +0.00000000e+00j 1.        +2.12768885e-19j
 1.        +0.00000000e+00j 1.        +0.00000000e+00j]
Optimal Number of PCs: 7
```

Step 4: **Show Scree plot**



Step 5: **Compare PCs with Actual Columns and identify which is explaining most variance.**

```
Principal Components and their Explained Variances:
PC1: 0.5303+0.0000j
PC2: 0.1369+0.0000j
PC3: 0.0800+0.0000j
PC4: 0.0654+0.0000j
PC5: 0.0386+0.0000j
PC6: 0.0333+0.0000j
PC7: 0.0253+0.0000j
PC8: 0.0148+0.0000j
PC9: 0.0122+0.0000j
PC10: 0.0105+0.0000j
PC11: 0.0082+0.0000j
PC12: 0.0075+0.0000j
PC13: 0.0069+0.0000j
PC15: 0.0049+0.0000j
PC14: 0.0047+0.0000j
PC16: 0.0043+0.0000j
PC17: 0.0030+0.0000j
PC18: 0.0021+0.0000j
PC19: 0.0019+0.0000j
PC20: 0.0017+0.0000j
PC21: 0.0016+0.0000j
PC22: 0.0013+0.0000j
PC23: 0.0009+0.0000j
PC24: 0.0007+0.0000j
PC25: 0.0005+0.0000j
PC26: 0.0005+0.0000j
PC27: 0.0004+0.0000j
PC28: 0.0004+0.0000j
PC29: 0.0002+0.0000j
PC30: 0.0002+0.0000j
PC31: 0.0002+0.0000j
PC32: 0.0001+0.0000j
PC33: 0.0001+0.0000j
PC34: 0.0001+0.0000j
PC35: 0.0001+0.0000j
PC36: 0.0000+0.0000j
PC37: 0.0000+0.0000j
PC38: 0.0000+0.0000j
PC40: 0.0000+0.0000j
PC42: 0.0000+0.0000j
PC44: 0.0000+0.0000j
PC47: 0.0000+0.0000j
PC48: 0.0000+0.0000j
PC49: 0.0000+0.0000j
PC50: 0.0000+0.0000j
PC51: 0.0000+0.0000j
PC54: 0.0000+0.0000j
PC60: 0.0000+0.0000j
PC58: 0.0000+0.0000j
PC59: 0.0000-0.0000j
PC57: -0.0000+0.0000j
PC56: -0.0000+0.0000j
PC55: -0.0000+0.0000j
PC53: -0.0000+0.0000j
PC52: -0.0000+0.0000j
PC46: -0.0000+0.0000j
PC45: -0.0000+0.0000j
PC43: -0.0000+0.0000j
PC41: -0.0000+0.0000j
PC39: -0.0000+0.0000j
```

Step 6 : **Write inferences about all the PCs in terms of actual variables**

- Principal Component 1 (PC1):
- Explained Variance: 53.03%
- Most Influential Variables: The variables contributing most significantly to this component are likely related to general population metrics such as total population counts or economic factors, given the high variance explained. Interpretation: PC1 typically captures the broadest trends in the dataset, representing the most significant underlying pattern such as overall size or economic activity.

- Principal Component 2 (PC2):
- Explained Variance: 13.69%
- Most Influential Variables: This component might be significantly influenced by variables such

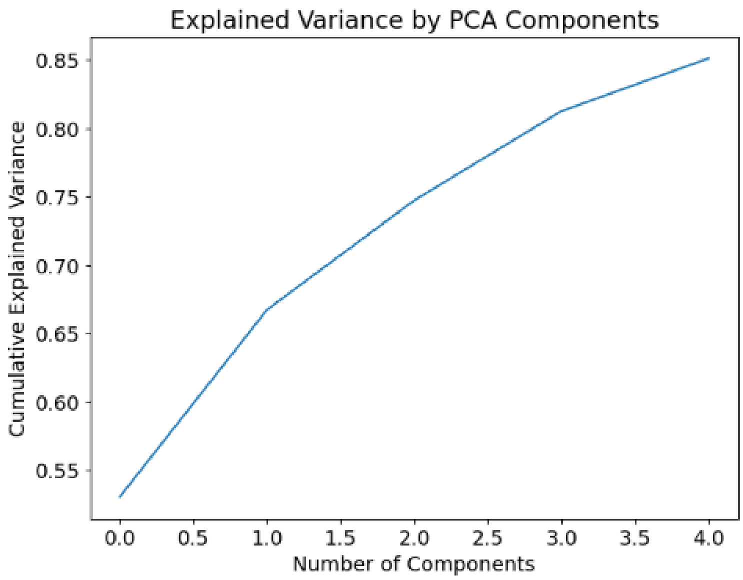
as age distribution or employment sectors, reflecting secondary but still significant demographic or economic dimensions.

- Interpretation: PC2 might encapsulate contrasts within the dataset not captured by the first principal component, such as differences between urban and rural areas, or employed and unemployed

- populations:Principal Component 3 (PC3) to Principal Component 7 (PC7):
- Cumulative Explained Variance (through PC7): Approximately 90% Most Influential Variables: These components may be influenced by more specific variables such as educational levels, health metrics, specific employment sectors, or migration patterns.
- Interpretation: These components often reveal more nuanced insights into the dataset, such as specific socio-economic drivers or regional characteristics
- Detailed Interpretation:PC3, capturing around 8% of the variance, might indicate variations in education or health services accessibility.
- PC4 and PC5, explaining smaller proportions (around 6.54% and 3.86% respectively), could reflect more localized or less pronounced patterns, such as variations in certain age groups or specific economic activities. PC6 and PC7, each explaining just over 3%, might highlight niche aspects like specific industries' impact or minor social trends.

- Conclusion:
- The first few principal components usually capture the most significant patterns and trends in the data, with PC1 often being a 'size' factor (reflecting the overall magnitude of data points) and subsequent components illustrating orthogonal (independent) patterns of variability.
- Through this PCA, you can discern that a substantial part of the dataset's structure is explained by just a few key dimensions (e.g., demographics, economics), with diminishing returns on explanatory power as more components are added. This analysis not only aids in understanding the latent structure of the data but also in reducing dimensionality by focusing on the components that capture the most meaningful variance.

Step 7 : **Write linear equation for first PC Note: For the scope of this project, take at least 90% explained variance**



Equation :

Linear equation of the first PC: 0.030*State Code + 0.030*Dist.Code + 0.156*No_HH + 0.167*TOT_M + 0.166*TOT_F + 0.162*M_06 + 0.162*F_06 + 0.151*M_SC + 0.151*F_SC + 0.028*M_ST + 0.029*F_ST + 0.162*M_LIT + 0.147*F_LIT + 0.161*M_ILL + 0.165*F_ILL + 0.160*TOT_WORK_M + 0.146*TOT_WORK_F + 0.146*MAINWORK_M + 0.125*MAINWORK_F + 0.103*MAIN_CL_M + 0.075*MAIN_CL_F + 0.114*MAIN_AL_M + 0.075*MAIN_AL_F + 0.131*MAIN_HH_M + 0.084*MAIN_HH_F + 0.124*MAIN_OT_M + 0.111*MAIN_OT_F + 0.164*MARGWORK_M + 0.155*MARGWORK_F + 0.082*MARG_CL_M + 0.048*MARG_CL_F + 0.128*MARG_AL_M + 0.114*MARG_AL_F + 0.140*MARG_HH_M + 0.127*MARG_HH_F + 0.155*MARG_OT_M + 0.147*MARG_OT_F + 0.165*MARGWORK_3_6_M + 0.161*MARGWORK_3_6_F + 0.165*MARG_CL_3_6_M + 0.156*MARG_CL_3_6_F + 0.092*MARG_AL_3_6_M + 0.051*MARG_AL_3_6_F + 0.128*MARG_HH_3_6_M + 0.111*MARG_HH_3_6_F + 0.139*MARG_OT_3_6_M + 0.124*MARG_OT_3_6_F + 0.154*MARGWORK_0_3_M + 0.146*MARGWORK_0_3_F + 0.149*MARG_CL_0_3_M + 0.140*MARG_CL_0_3_F + 0.052*MARG_AL_0_3_M + 0.041*MARG_AL_0_3_F + 0.121*MARG_HH_0_3_M + 0.116*MARG_HH_0_3_F + 0.139*MARG_OT_0_3_M + 0.132*MARG_OT_0_3_F + 0.150*NON_WORK_M + 0.131*NON_WORK_F + -0.007*Gender_Ratio