MACHINE LEARNING

BUSINESS REPORT

BY

G S JAIGURURAM

TABLE OF CONTENTS

TABLE OF FIGURES

Problem 1:

The ads24x7 is a Digital Marketing company which has now got seed funding of $10 Million. They are expanding their wings in Marketing Analytics. They collected data from their Marketing Intelligence team and now wants you (their newly appointed data analyst) to segment type of ads based on the features provided. Use Clustering procedure to segment ads into homogeneous groups.

PROBLEM 1.1 DEFINE THE PROBLEM AND PERFORM EXPLORATORY DATA ANALYSIS:

Problem 1.1.1 Problem definition:

The task is to segment types of ads based on the features provided in the data collected by the Marketing Intelligence team of ads24x7, a Digital Marketing company that has received $10 million in seed funding. The goal is to use clustering procedures to group ads into homogeneous clusters based on their characteristics.

Problem 1.1.2 Check shape, Data types, statistical summary:

First, we import all the necessary libraries seaborn, numpy, pandas, sklearn, matplotlib etc. to perform our analysis

Next, we import the data set “Ads_Data”

Data Dictionary:

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 Advertisement.
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 particular 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.

Shape of the data:

Shape of the dataset is 23066 rows and 19 Columns.

Data Type:

```
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
```

There are total of 13 Numerical and 6 Categorical data type are available.

Statistical Summary:

	count	mean	std	min	25%	50%	75%	max
Ad - Length	23066.0	3.851631e+02	2.336514e+02	120.0000	120.000000	300.00000	7.200000e+02	728.00
Ad- Width	23066.0	3.378960e+02	2.030929e+02	70.0000	250.000000	300.00000	6.000000e+02	600.00
Ad Size	23066.0	9.667447e+04	6.153833e+04	33600.0000	72000.000000	72000.00000	8.400000e+04	216000.00
Available_Impressions	23066.0	2.432044e+06	4.742888e+06	1.0000	33672.250000	483771.00000	2.527712e+06	27592861.00
Matched_Queries	23066.0	1.295099e+06	2.512970e+06	1.0000	18282.500000	258087.50000	1.180700e+06	14702025.00
Impressions	23066.0	1.241520e+06	2.429400e+06	1.0000	7990.500000	225290.00000	1.112428e+06	14194774.00
Clicks	23066.0	1.067852e+04	1.735341e+04	1.0000	710.000000	4425.00000	1.279375e+04	143049.00
Spend	23066.0	2.706626e+03	4.067927e+03	0.0000	85.180000	1425.12500	3.121400e+03	26931.87
Fee	23066.0	3.351231e-01	3.196322e-02	0.2100	0.330000	0.35000	3.500000e-01	0.35
Revenue	23066.0	1.924252e+03	3.105238e+03	0.0000	55.365375	926.33500	2.091338e+03	21276.18
CTR	18330.0	7.366054e-02	7.515992e-02	0.0001	0.002600	0.08255	1.300000e-01	1.00
CPM	18330.0	7.672045e+00	6.481391e+00	0.0000	1.710000	7.66000	1.251000e+01	81.56
CPC	18330.0	3.510606e-01	3.433338e-01	0.0000	0.090000	0.16000	5.700000e-01	7.26

Insights:

1. The dataset contains 23066 entries (rows) and 19 features (columns).

2. We have 13 features that are numerical and 6 categorical type data.

3. We have 7 Int data type and 6 float data type

3. We have 6 object data type.

Problem 1.1.3 Univariate analysis:

Numerical Data type:

We Going to Plot Box plot for all Numerical Data.

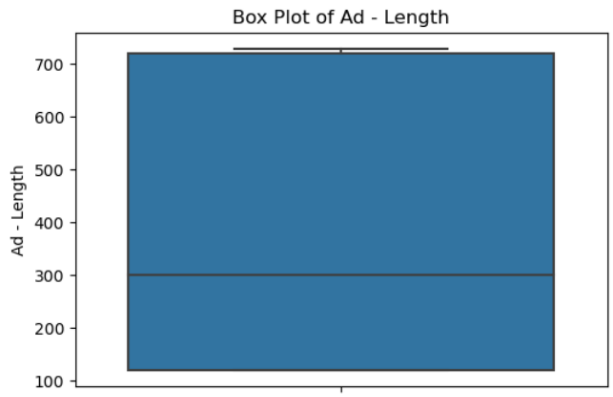


FIGURE 1

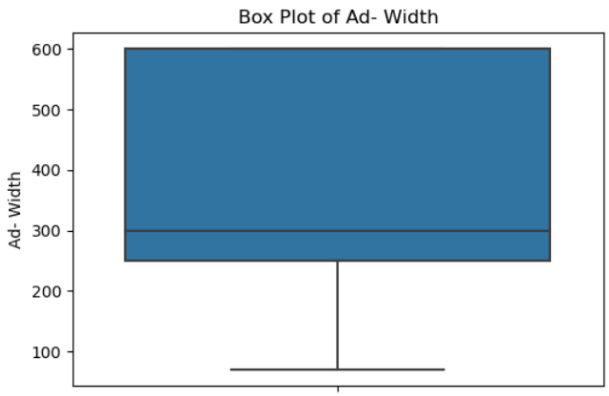


FIGURE 2

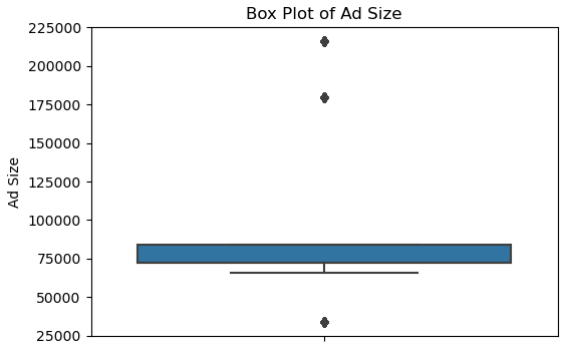


FIGURE 3

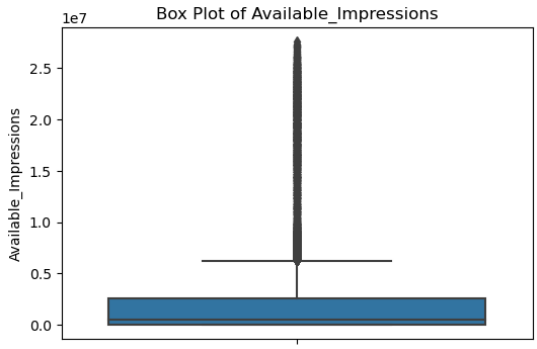


FIGURE 4

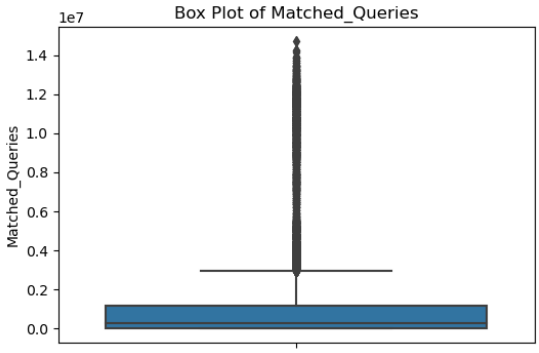


FIGURE 5

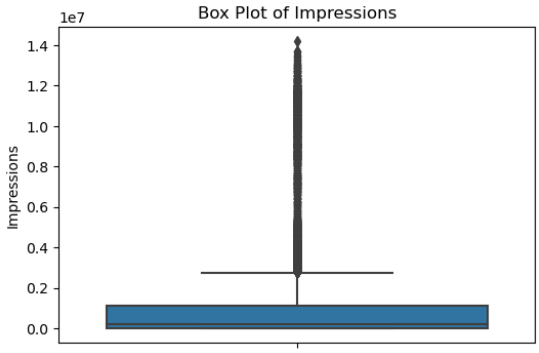


FIGURE 6

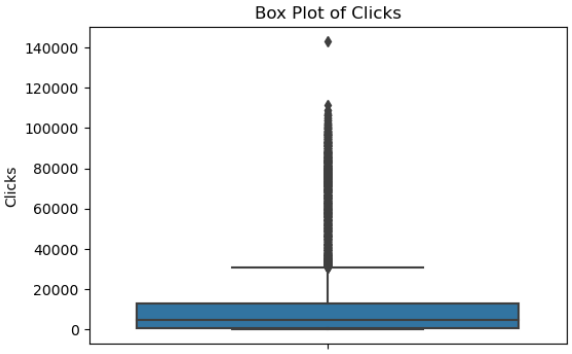


FIGURE 7

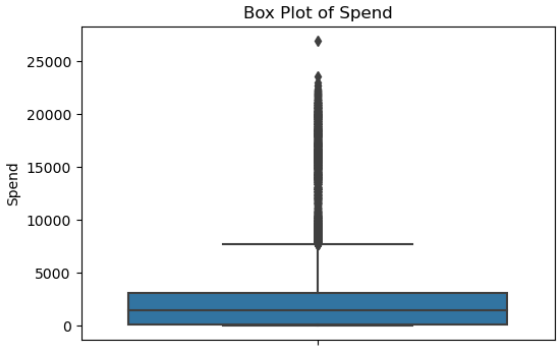


FIGURE 8

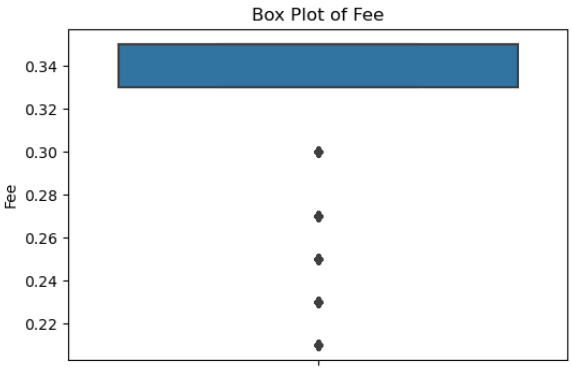


FIGURE 9

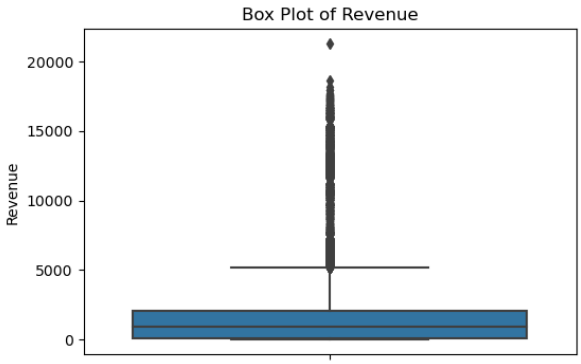


FIGURE 10

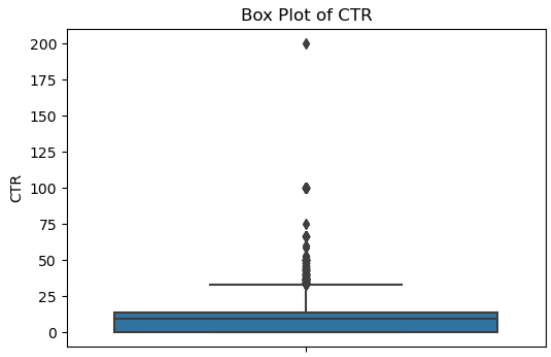


FIGURE 2

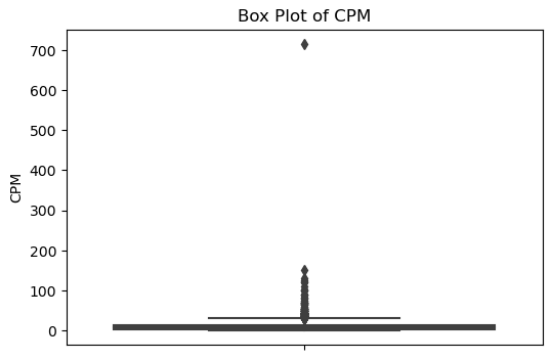


FIGURE 12

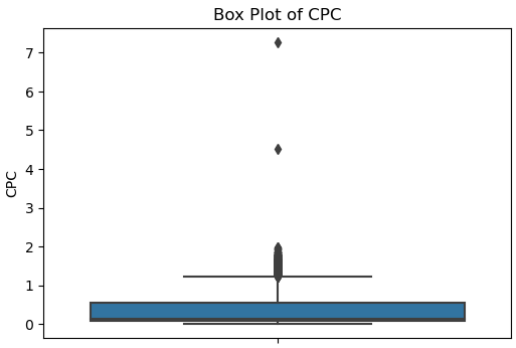


FIGURE 13

Insight:

Impressions, Matched_Queries, and Available_Impressions: These measures have a long tail and a highly skewed distribution, suggesting the existence of outliers with values noticeably higher than the median. Near the bottom end of the scale is where most of the data points are concentrated.

Clicks: Clicks also exhibit a right skewed distribution with outlier's present.

Spend, Revenue: Both Spend and Revenue show a right skewed distribution, with several outliers indicating instances of higher expenditure and revenue generation.

CTR: The CTR distribution is skewed towards the lower values with some outliers indicating exceptionally high CTR.

CPM: The CPM values are concentrated at the lower end with a few high value outliers.

CPC: The CPC distribution has a concentration of lower values with some outliers on the higher end.

Categorical Data Type:

We are going to plot a bar plot for Categorical data.

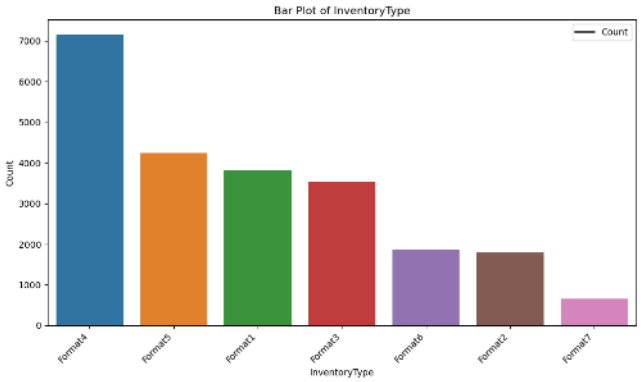


FIGURE 14

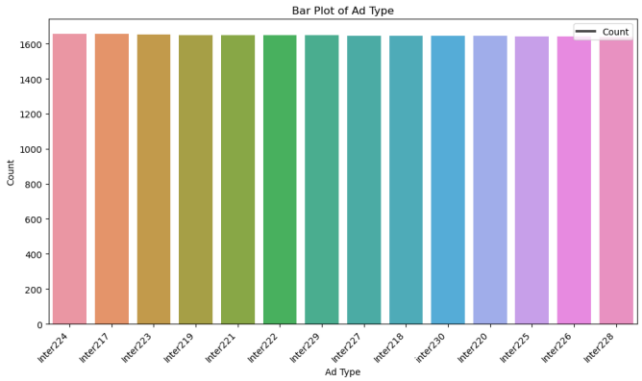


FIGURE 15

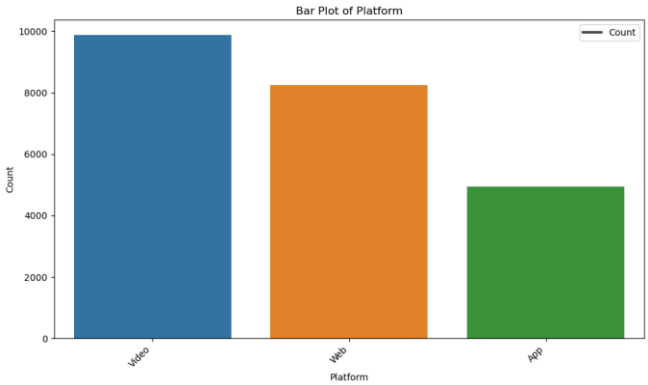


FIGURE 16

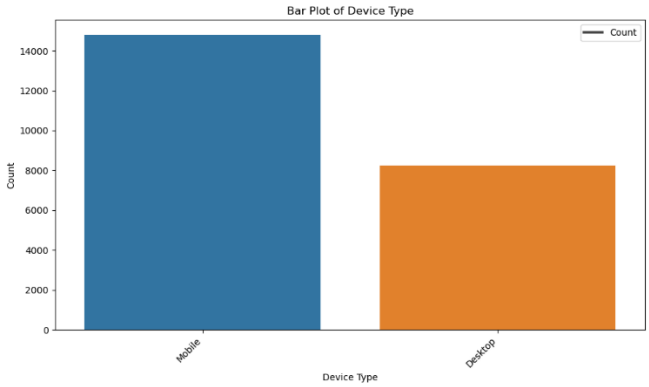


FIGURE 17

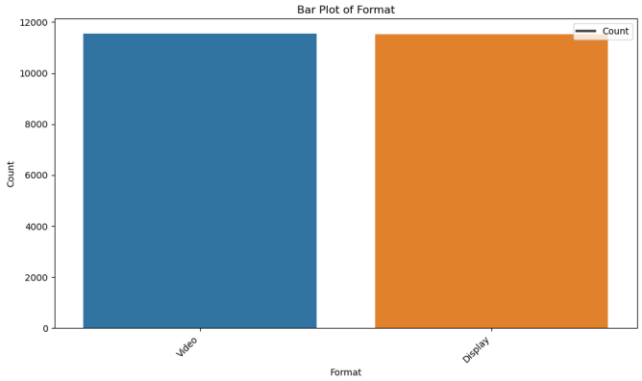


FIGURE 18

Insights:

InventoryType Distribution: The bar plot for 'InventoryType' shows that Format4 is the most common category, followed by Format5, Format1, and Format3. The least common is Format7.

Platform Distribution: The 'Platform' distribution indicates that Video is the most frequent platform, with Web being the second most common and App being the least common.

Device Type Distribution: For 'Device Type', Mobile devices are more prevalent than Desktop devices in the dataset.

Format Distribution: The distribution of 'Format' is nearly even, with Video being slightly more common than Display.

Ad Type: The distribution ‘Ad Type’ is even along all the types of Ad.

Problem 1.1.4 Bivariate analysis:

Numerical Vs Numerical Data: We going to plot a scatter plot.

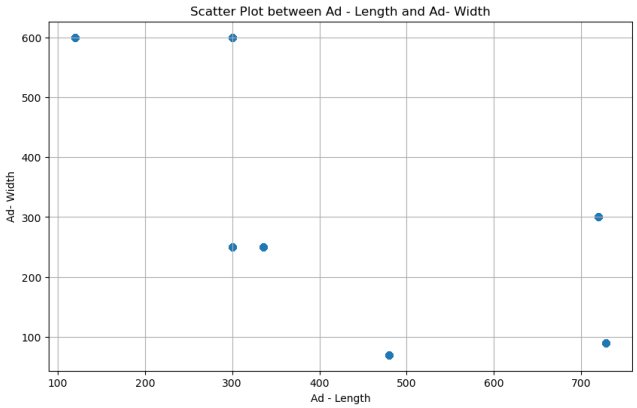


FIGURE 19

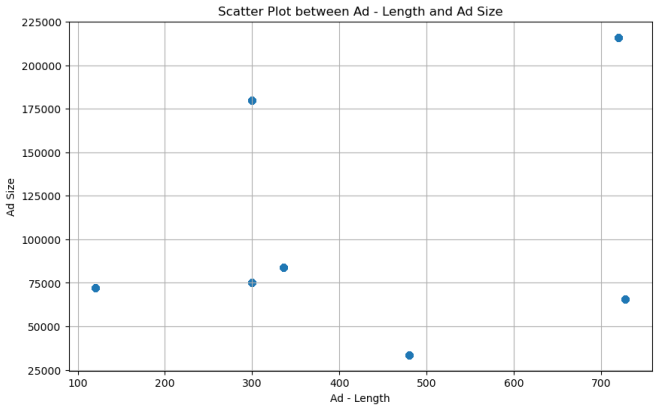


FIGURE 20

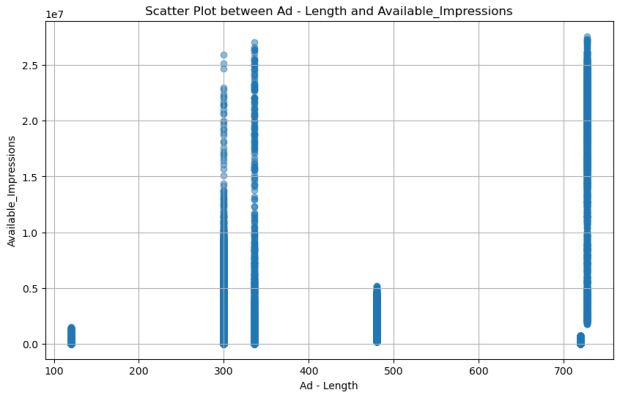


FIGURE 3

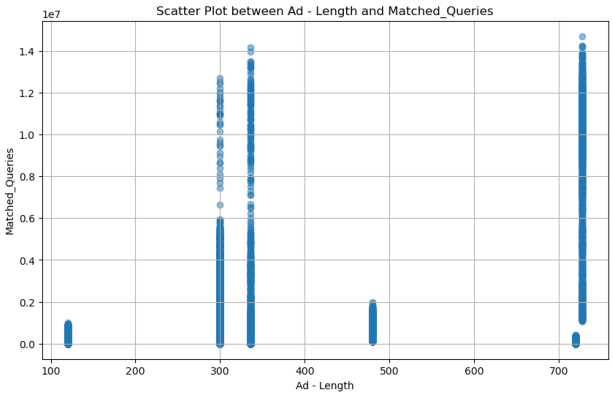


FIGURE 22

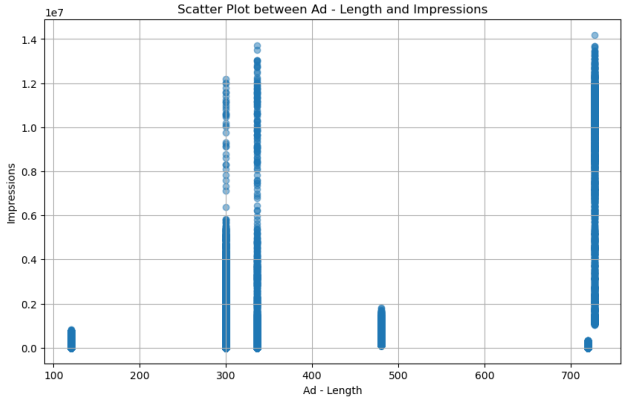


FIGURE 23

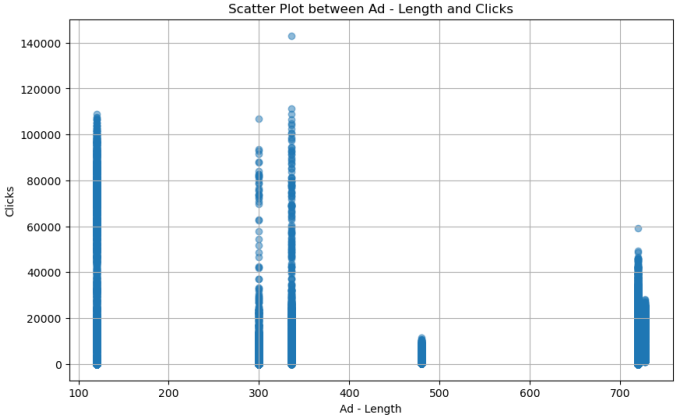


FIGURE 24

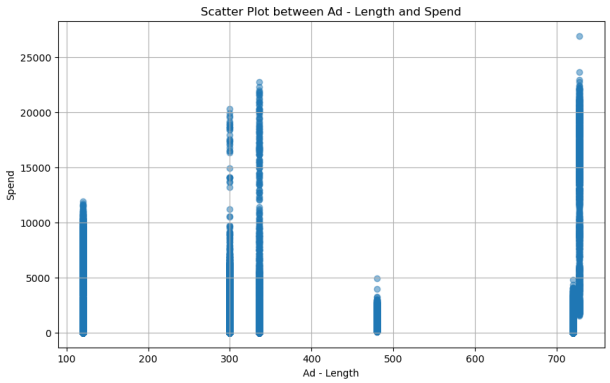


FIGURE 25

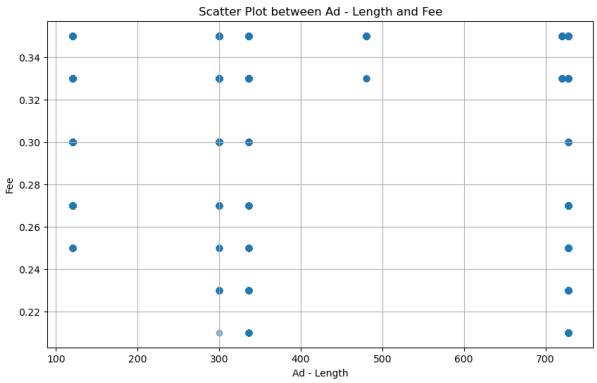


FIGURE 26

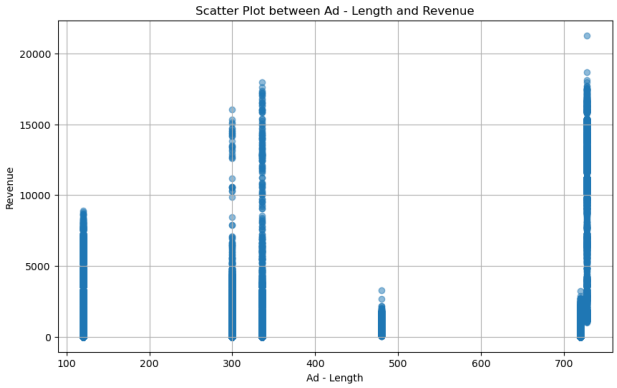


FIGURE 27

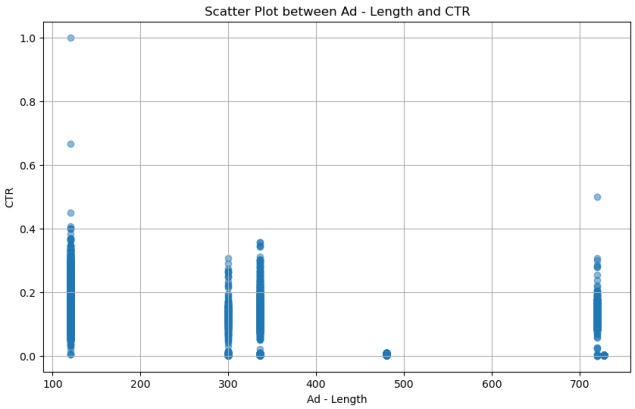


FIGURE 28

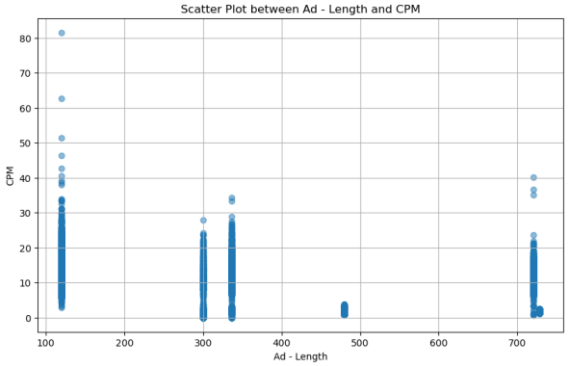


FIGURE 29

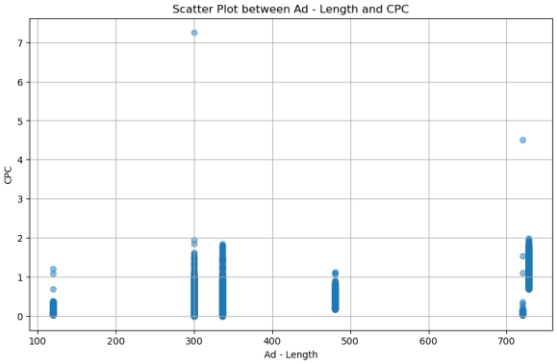


FIGURE 30

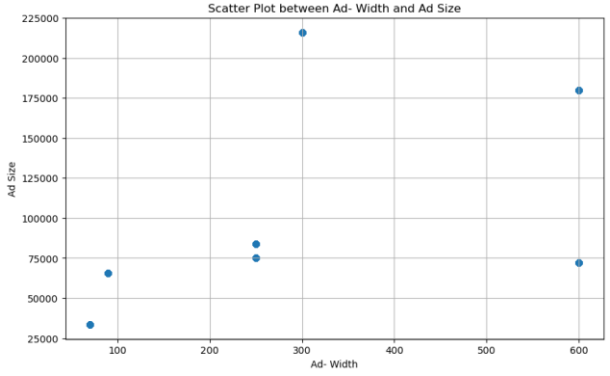


FIGURE 31

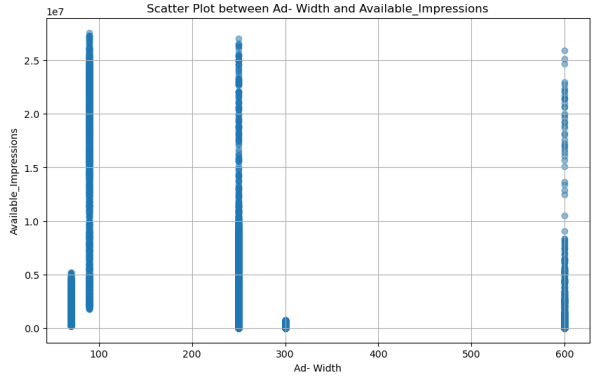


FIGURE 32

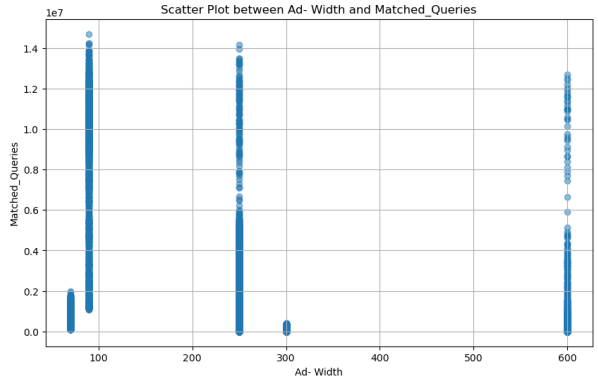


FIGURE 33

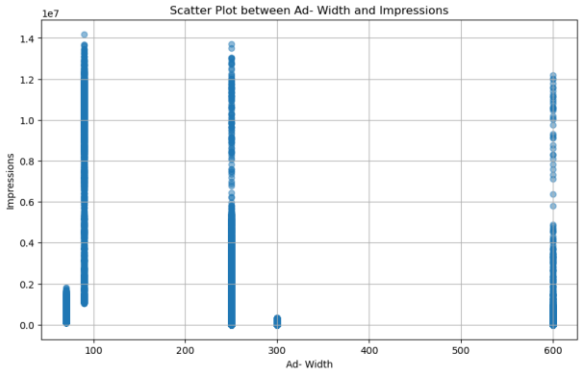


FIGURE 34

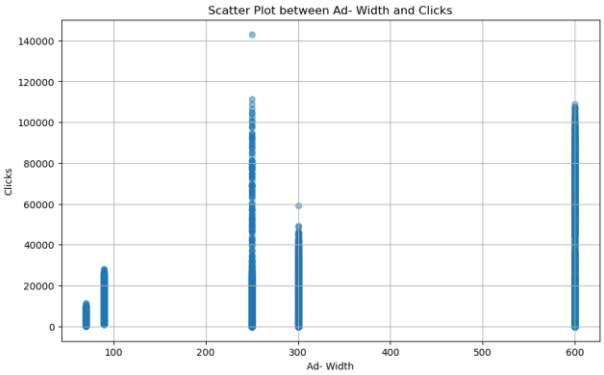


FIGURE 35

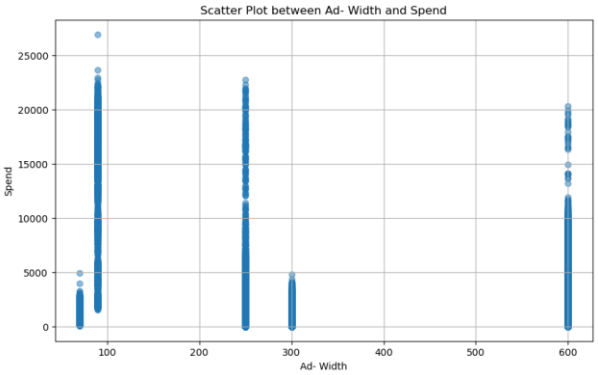


FIGURE 36

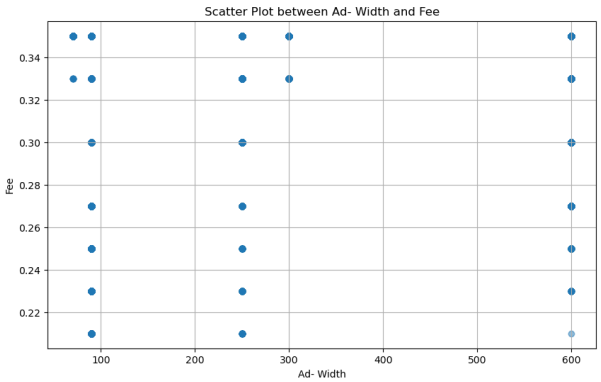


FIGURE 37

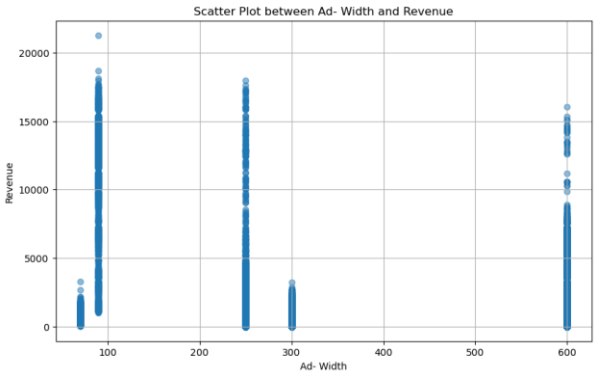


FIGURE 38

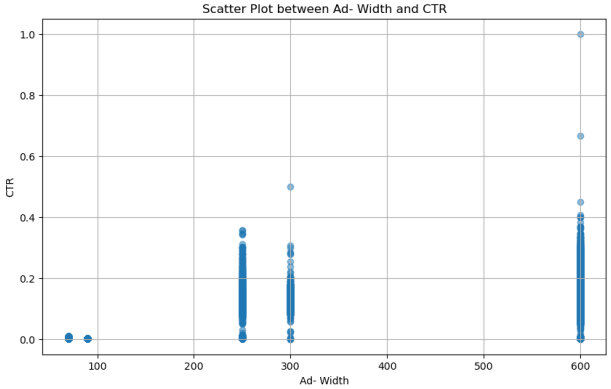


FIGURE 39

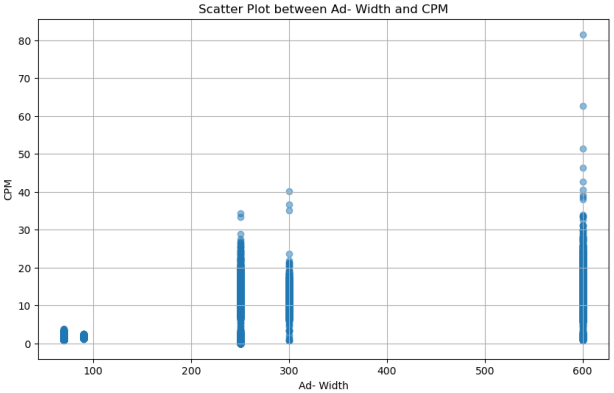


FIGURE 40

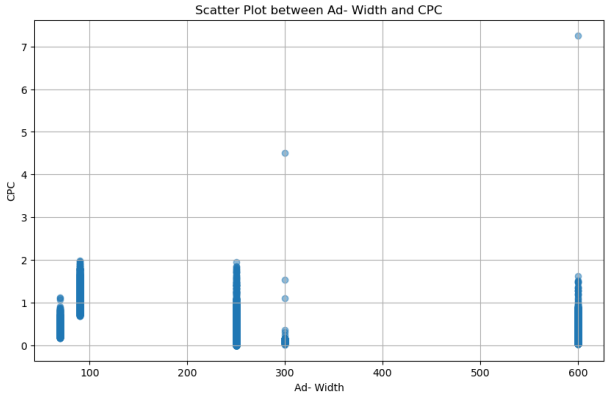


FIGURE 41

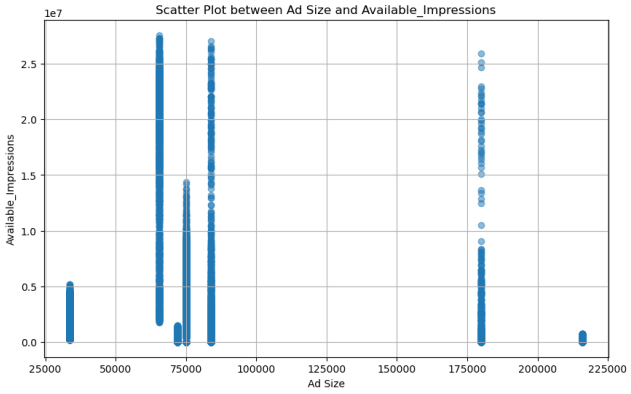


FIGURE 42

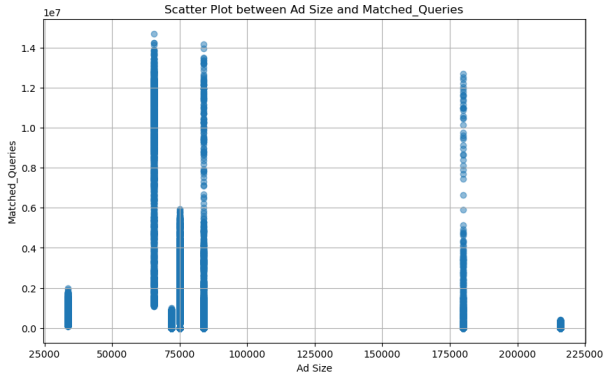


FIGURE 43

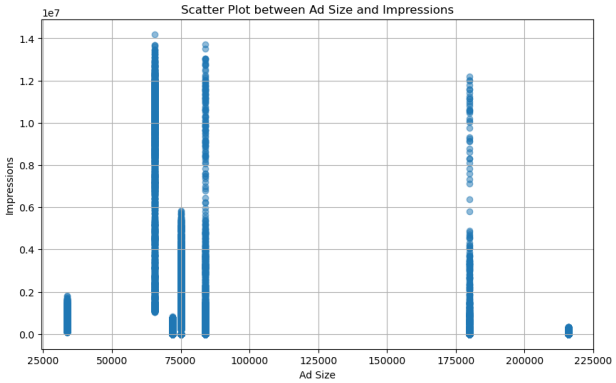


FIGURE 44

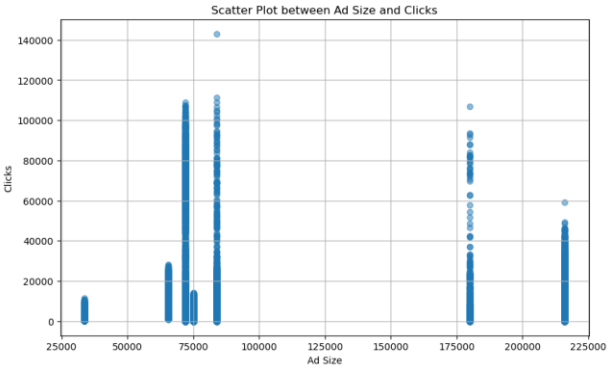


FIGURE 45

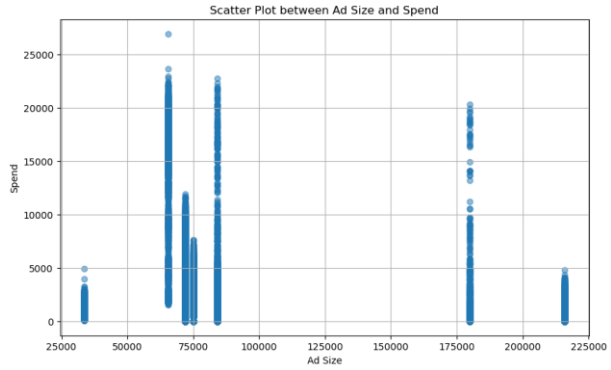


FIGURE 46

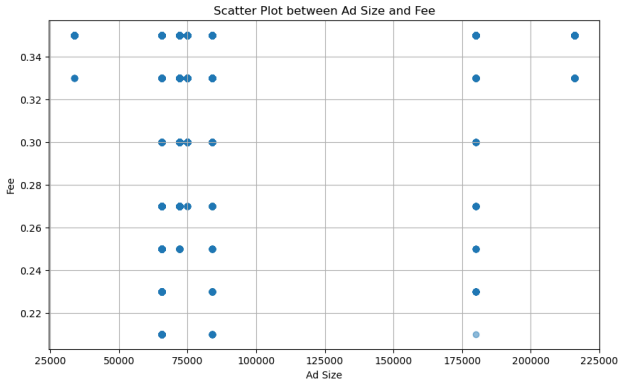


FIGURE 47

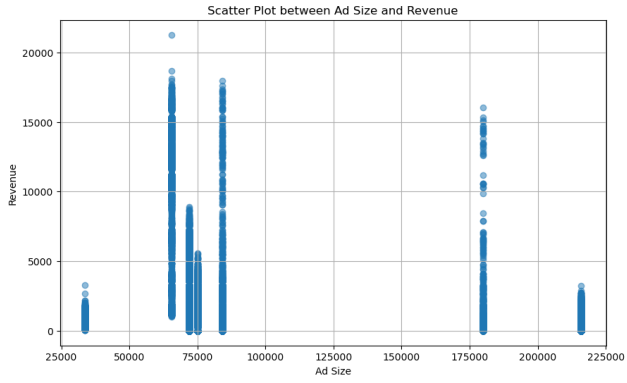


FIGURE 48

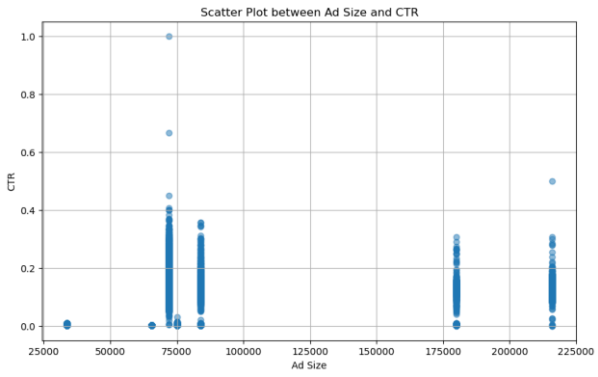


FIGURE 49

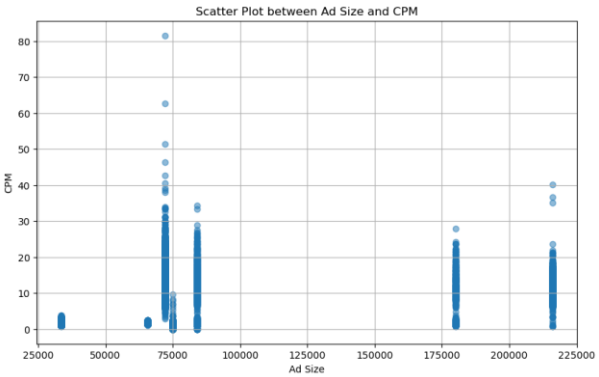


FIGURE 50

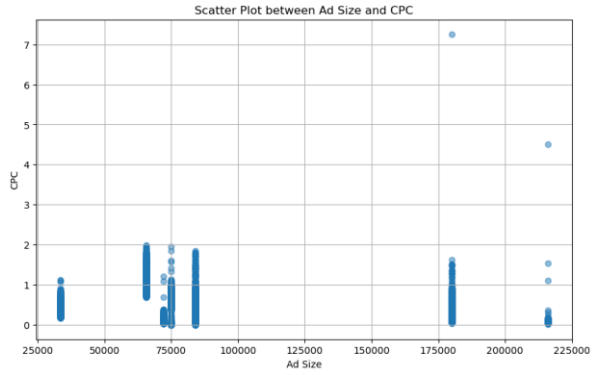


FIGURE 4

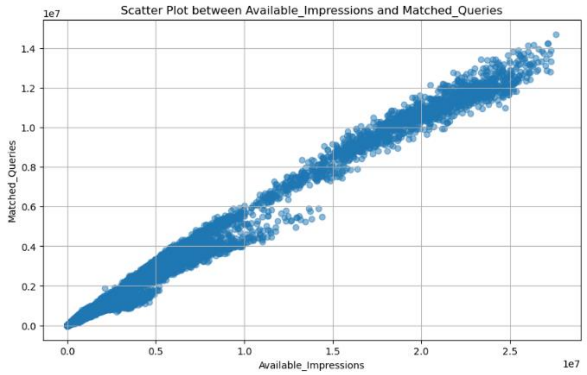


FIGURE 52

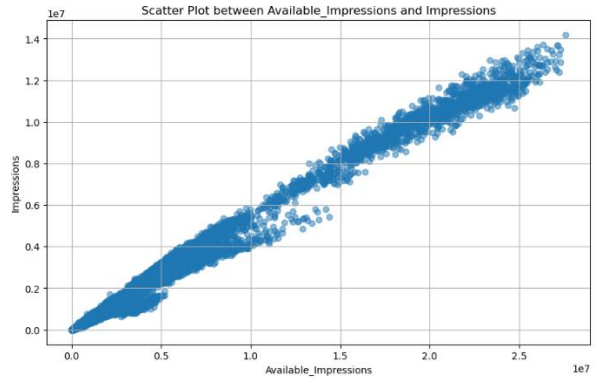


FIGURE 53

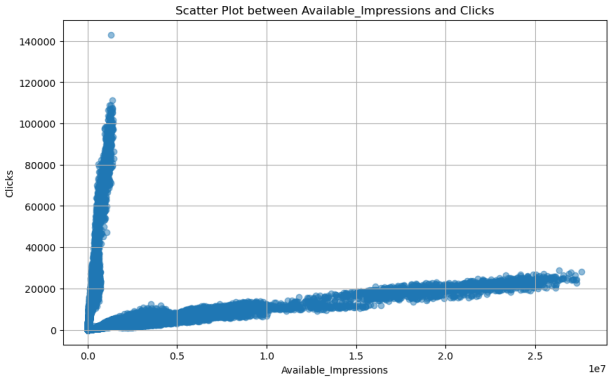


FIGURE 54

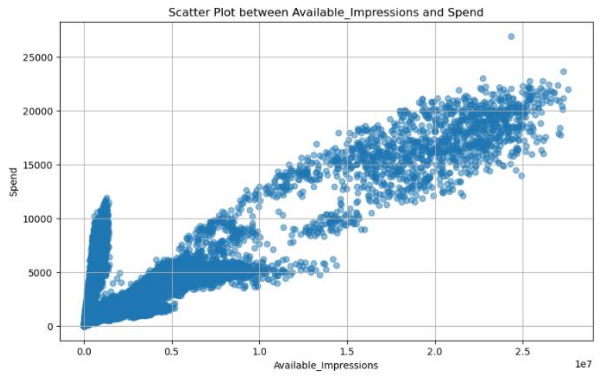


FIGURE 55

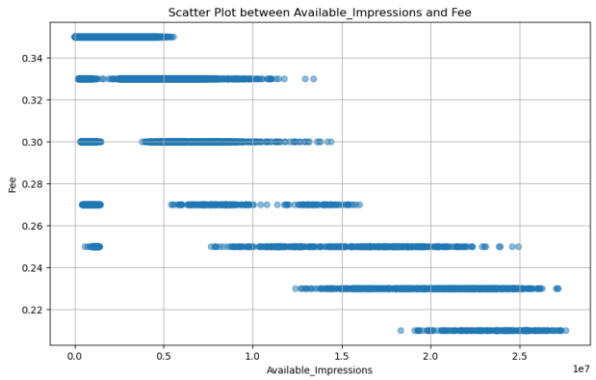


FIGURE 56

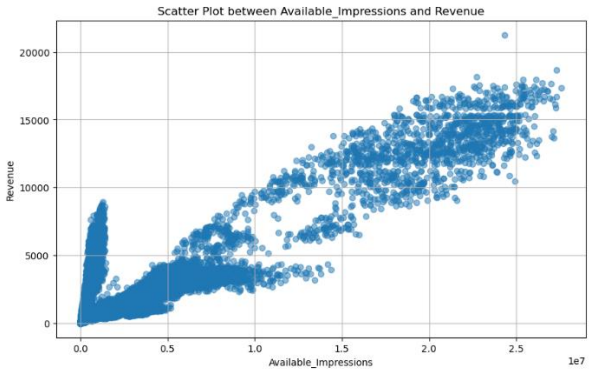


FIGURE 57

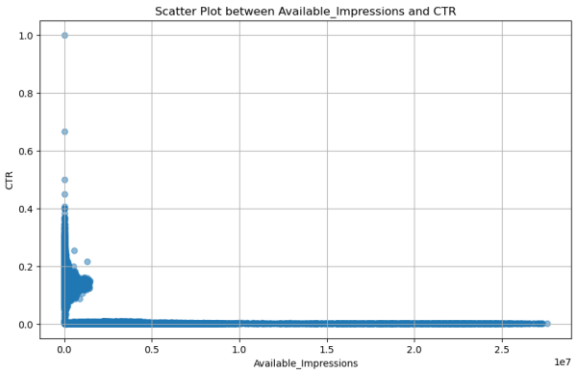


FIGURE 58

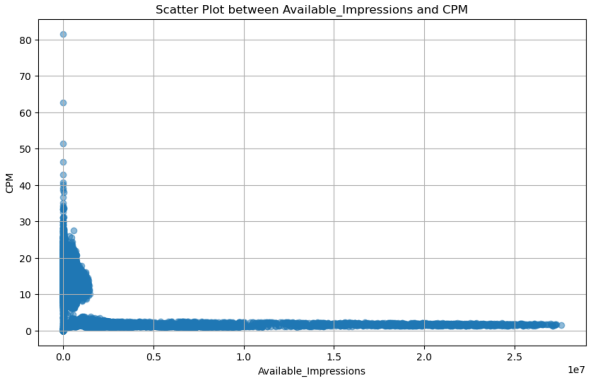


FIGURE 59

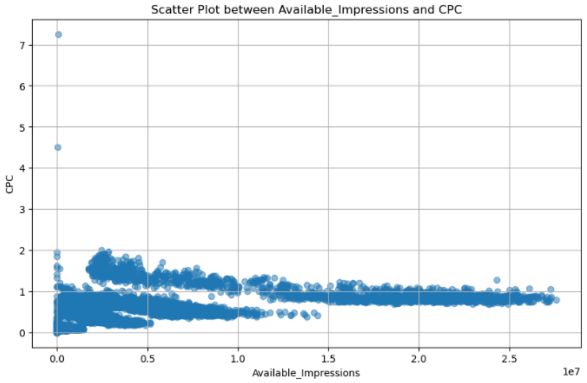


FIGURE 60

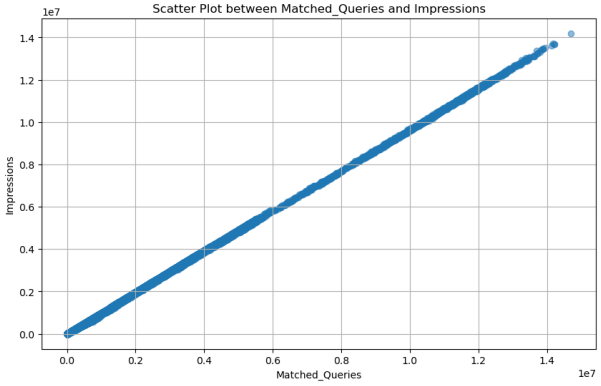


FIGURE 61

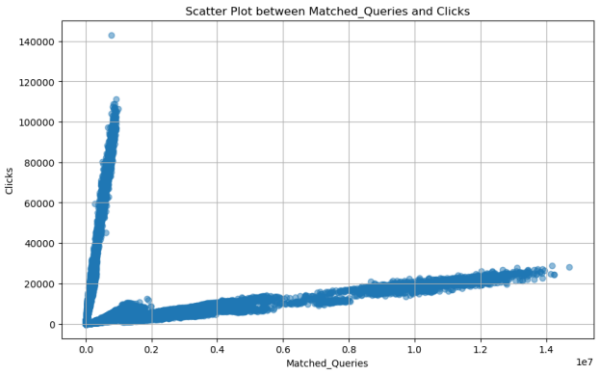


FIGURE 62

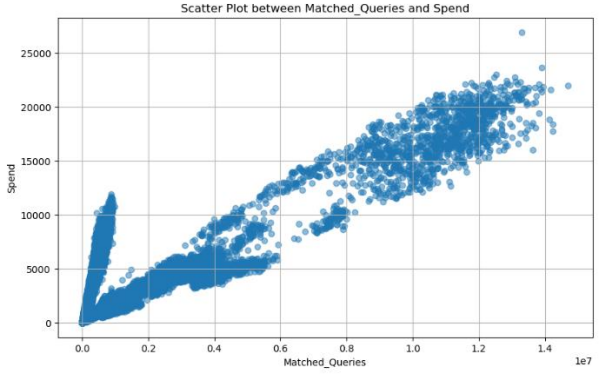


FIGURE 63

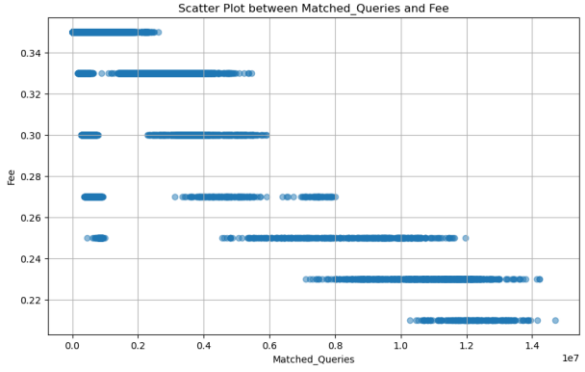


FIGURE 64

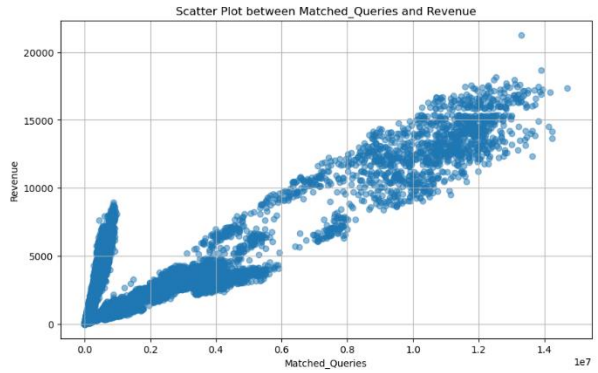


FIGURE 65

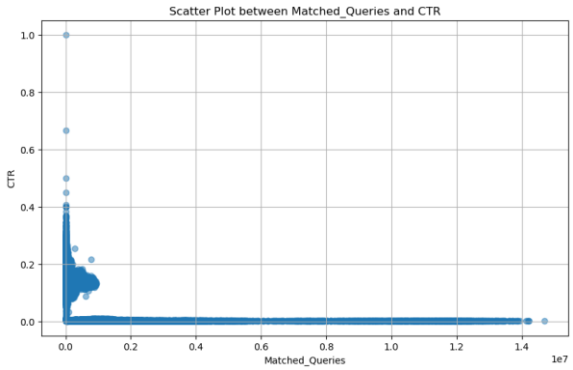


FIGURE 66

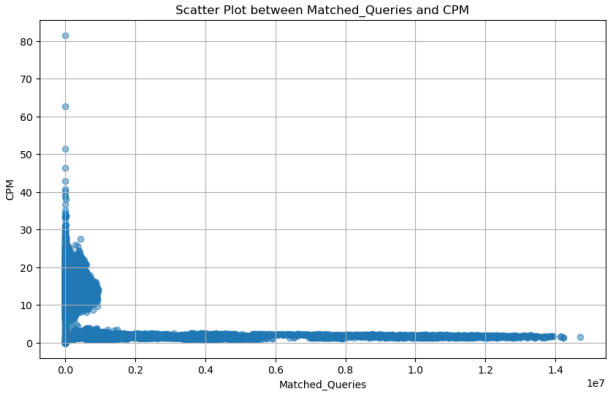


FIGURE 67

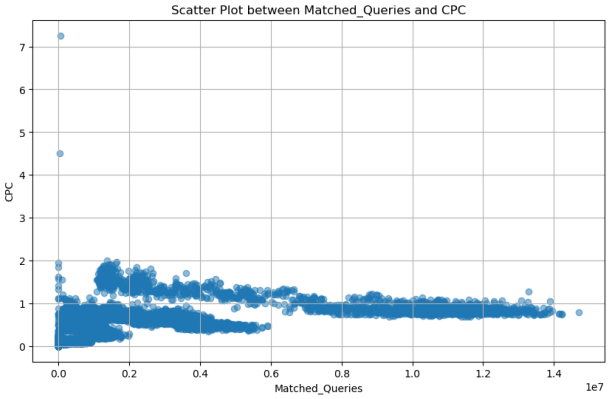


FIGURE 68

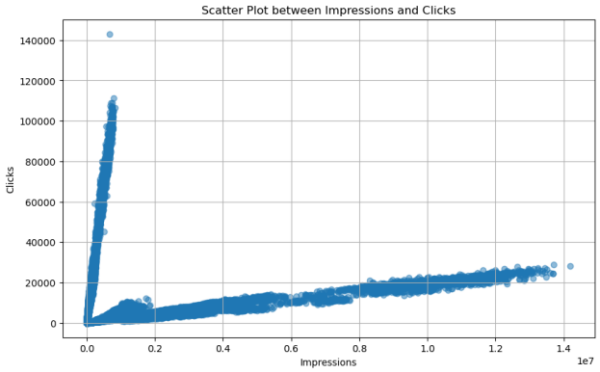


FIGURE 69

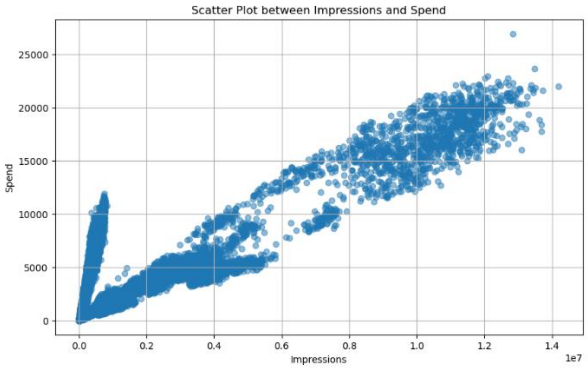


FIGURE 70

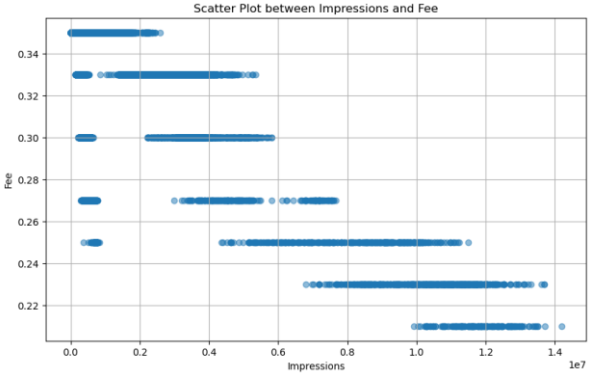


FIGURE 71

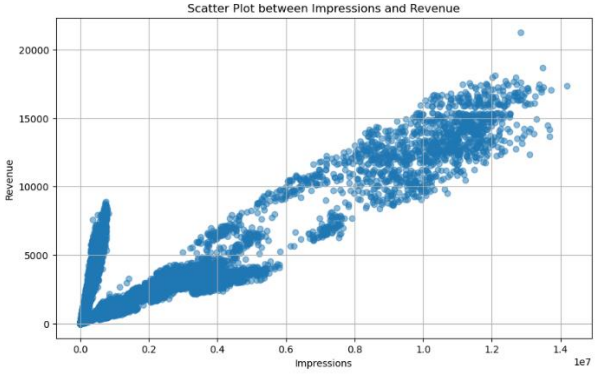


FIGURE 72

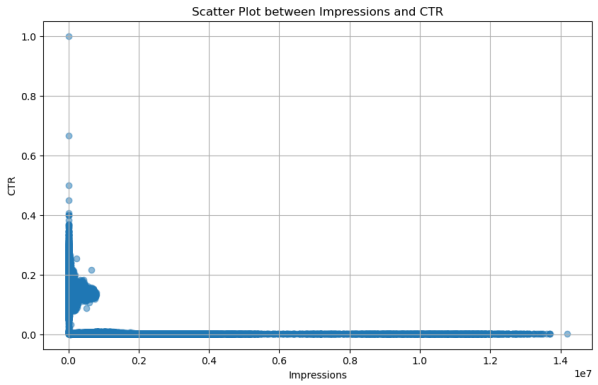


FIGURE 73

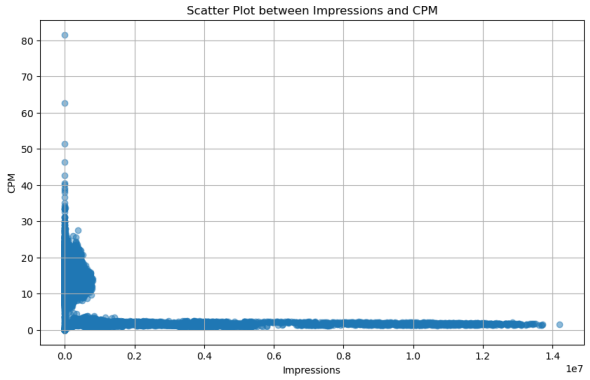


FIGURE 74

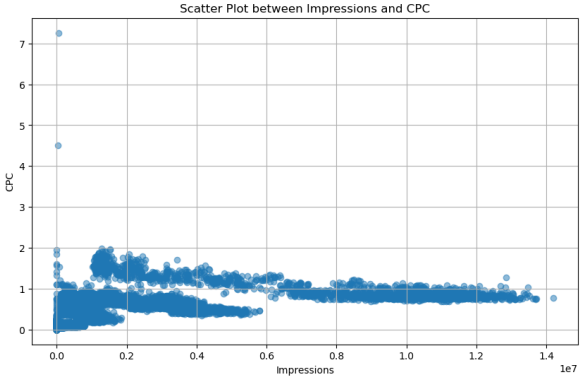


FIGURE 75

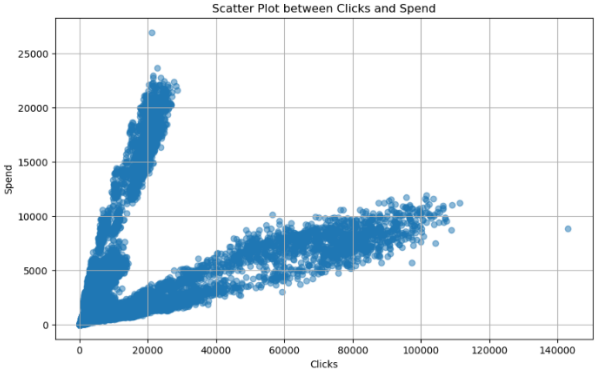


FIGURE 76

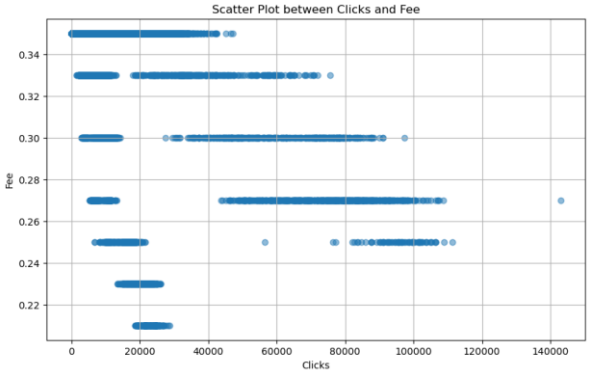


FIGURE 77

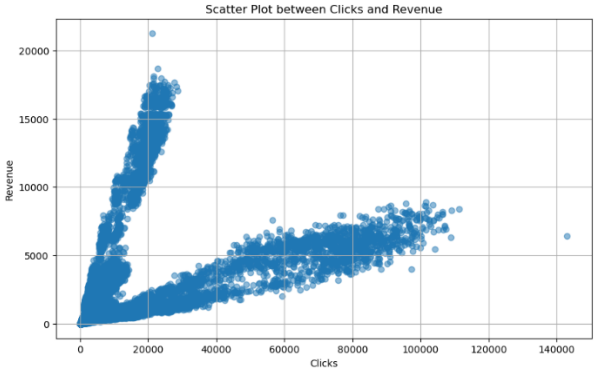


FIGURE 78

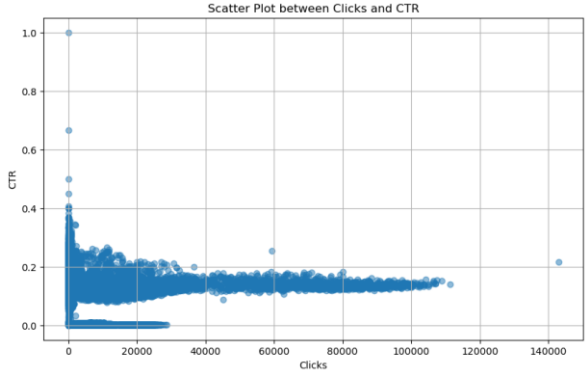


FIGURE 79

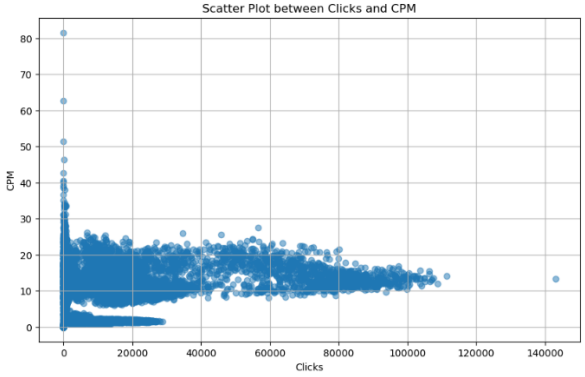


FIGURE 80

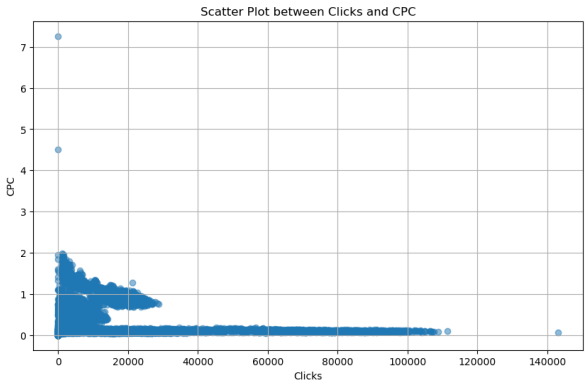


FIGURE 81

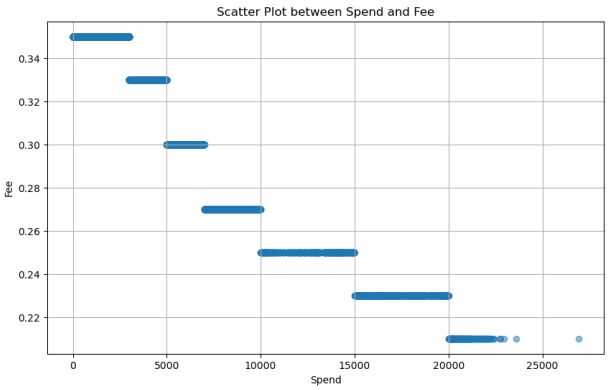


FIGURE 82

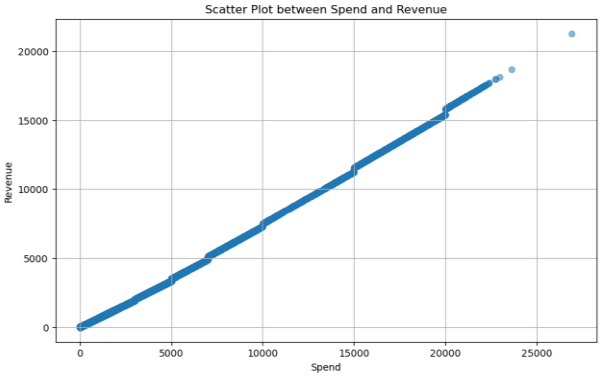


FIGURE 83

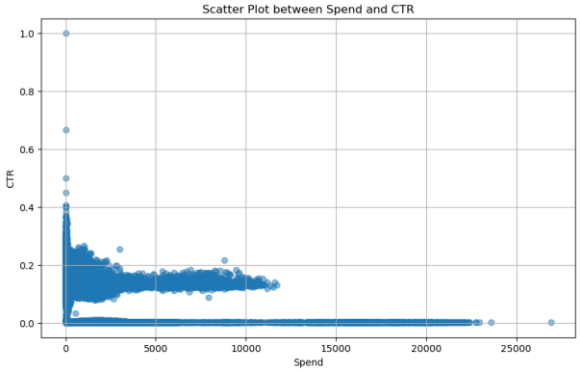


FIGURE 84

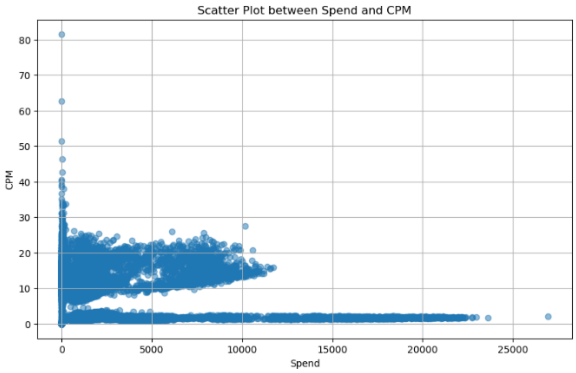


FIGURE 85

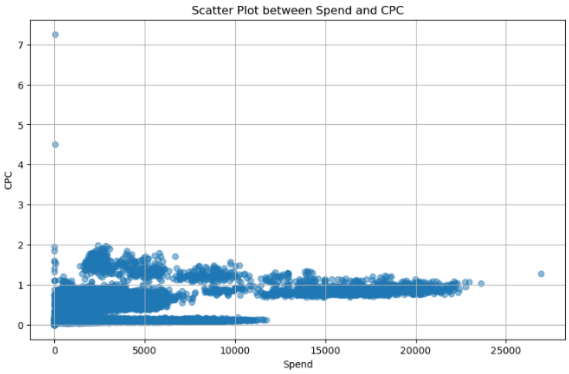


FIGURE 86

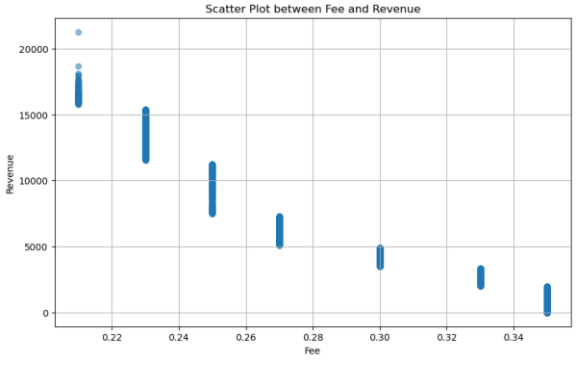


FIGURE 87

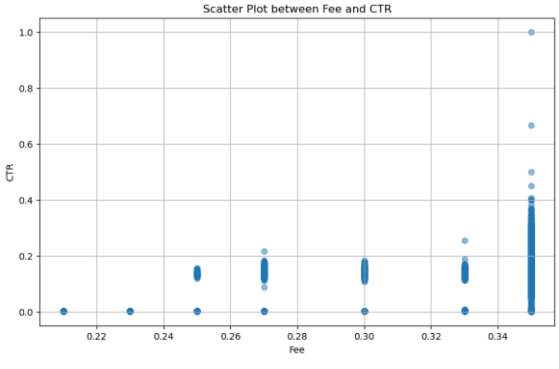


FIGURE 88

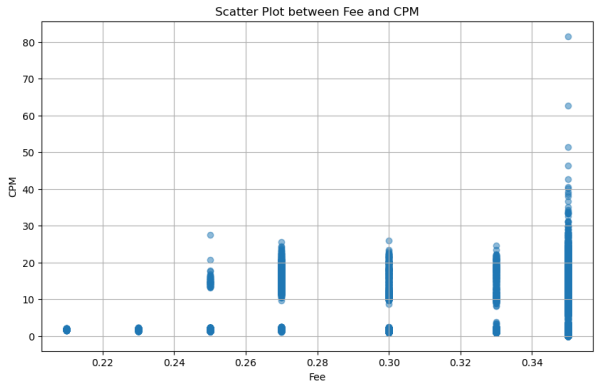


FIGURE 89

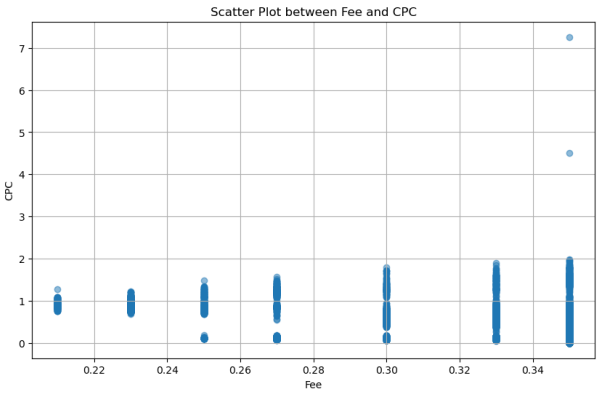


FIGURE 90

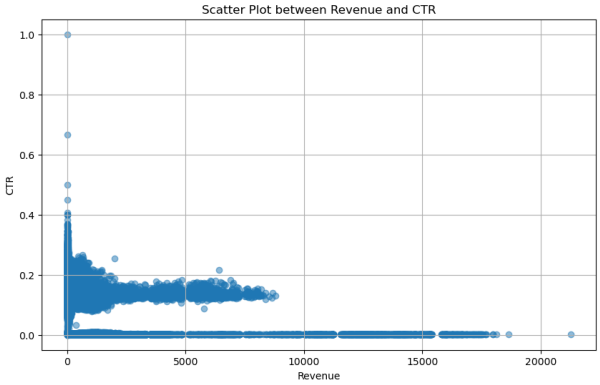


FIGURE 91

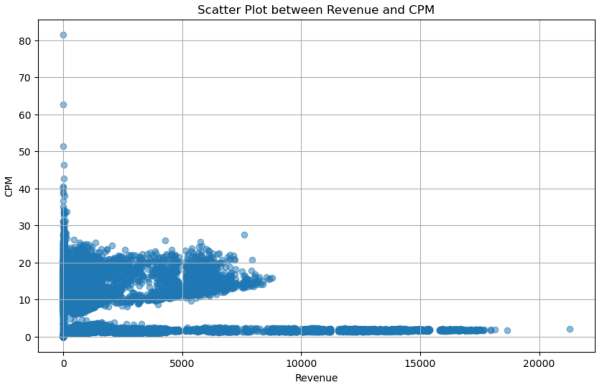


FIGURE 92

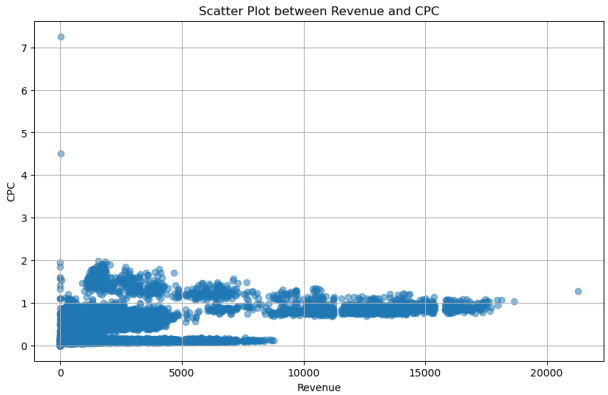


FIGURE 93

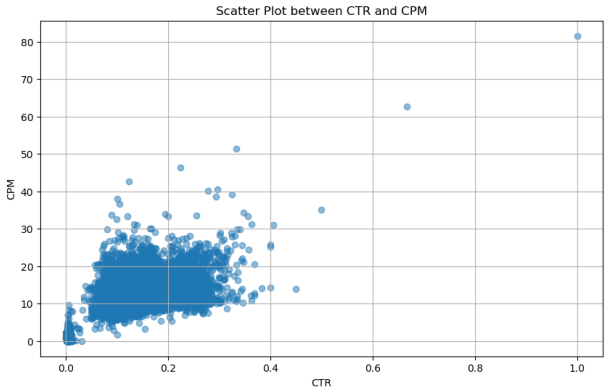


FIGURE 94

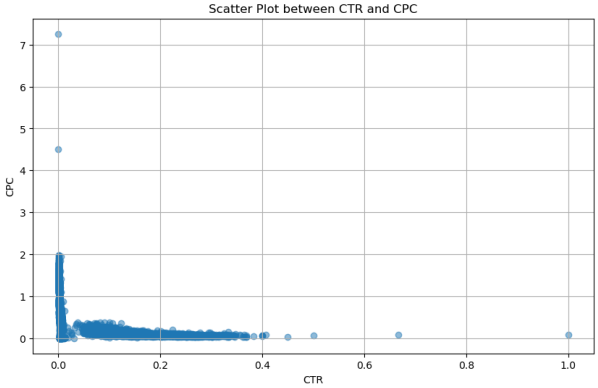


FIGURE 95

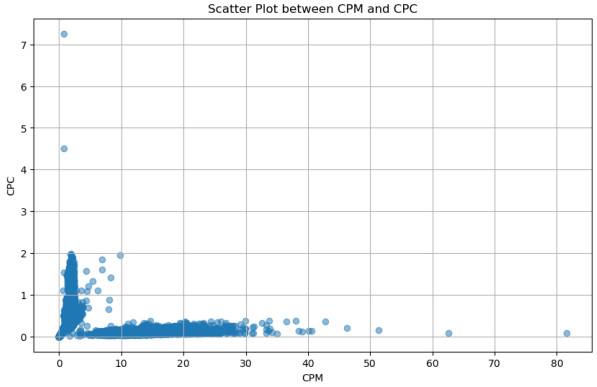


FIGURE 96

Insights:

Strong Positive Correlation: The scatter plots show a strong positive correlation between 'Available_Impressions', 'Matched_Queries', and 'Impressions'. This indicates that as the number of available impressions increases, the number of matched queries and actual impressions also increases.

Moderate Positive Correlation: There is a moderate positive correlation between 'Clicks' and 'Spend', as well as between 'Clicks' and 'Revenue'. This suggests that higher spending tends to result in more clicks, which in turn can lead to higher revenue.

Weak Correlation: The relationship between 'CTR', 'CPM', 'CPC' and other variables like 'Available_Impressions', 'Matched_Queries', 'Impressions', 'Clicks', 'Spend', and 'Revenue' appears to be weaker. This could mean that these metrics are influenced by factors other than just the volume of impressions or clicks.

Numerical Vs Categorical Data: We are going to plot box plot.

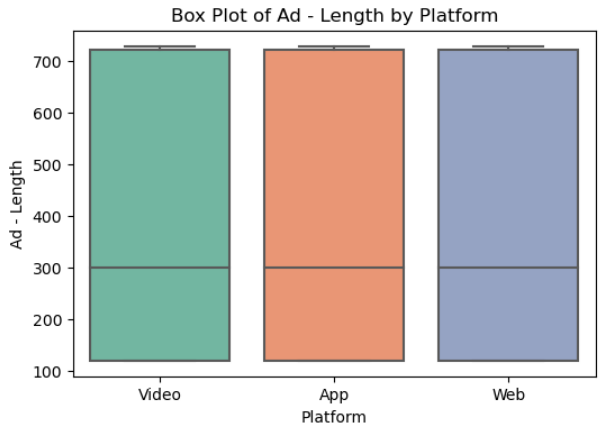


FIGURE 97

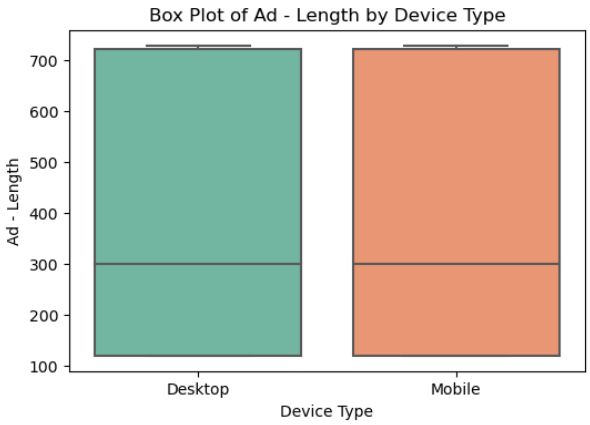


FIGURE 98

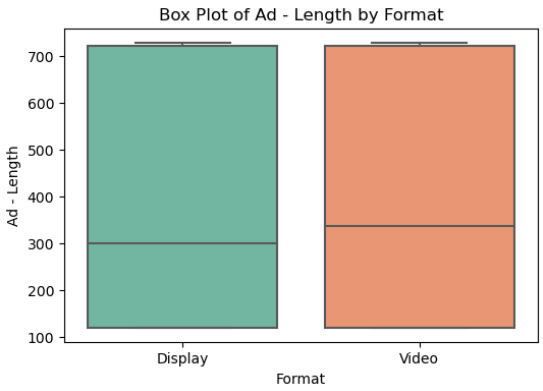


FIGURE 99

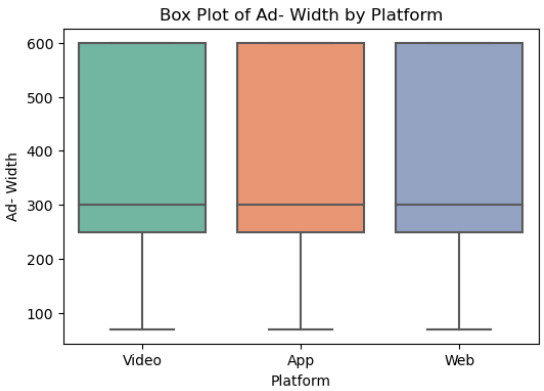


FIGURE 100

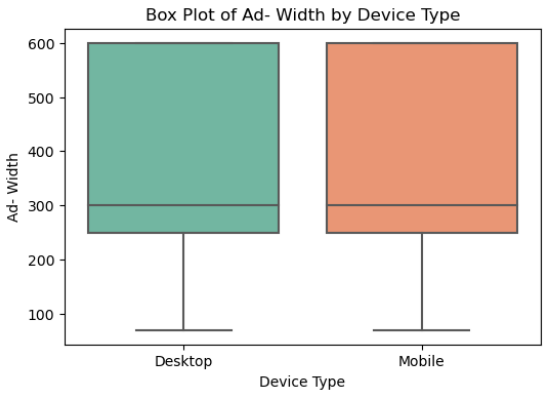


FIGURE 101

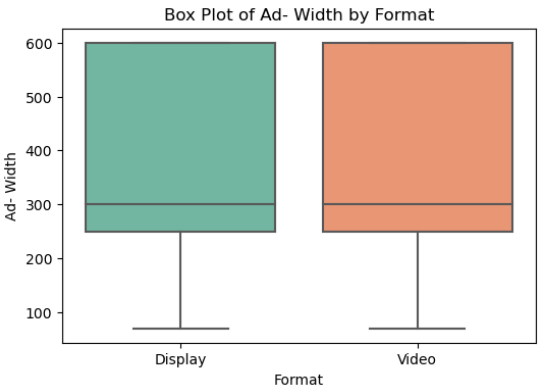


FIGURE 102

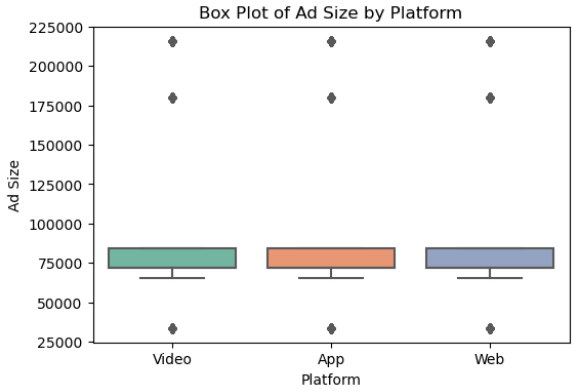


FIGURE 103

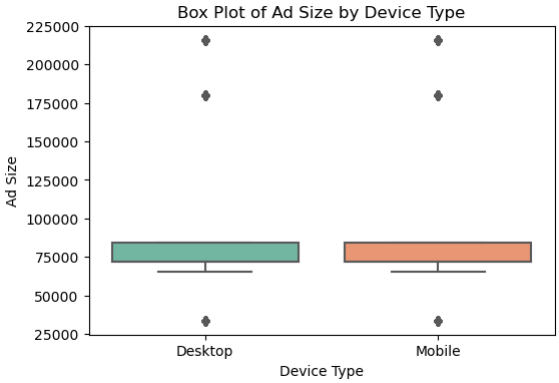


FIGURE 104

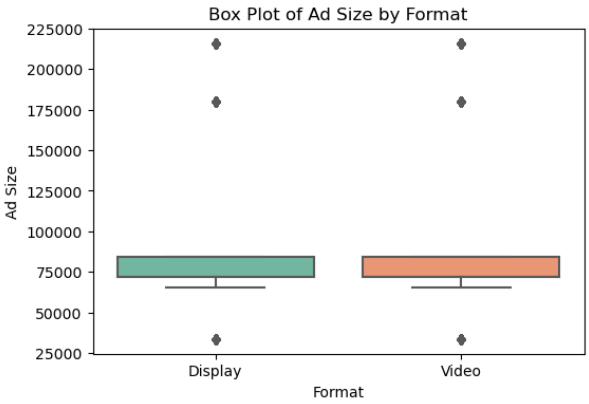


FIGURE 105

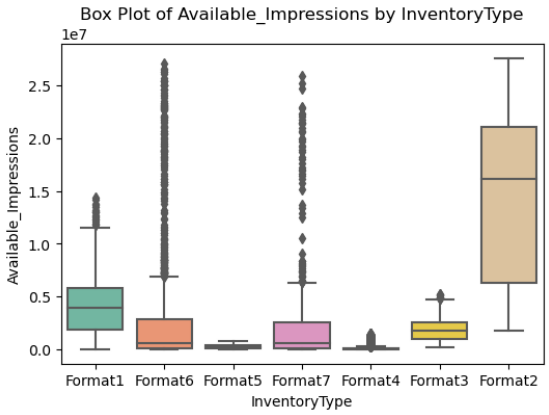


FIGURE 106

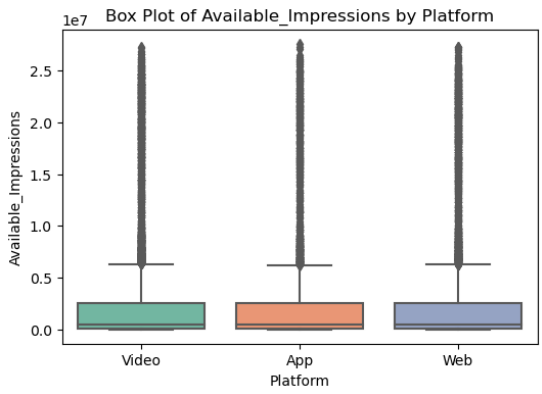


FIGURE 107

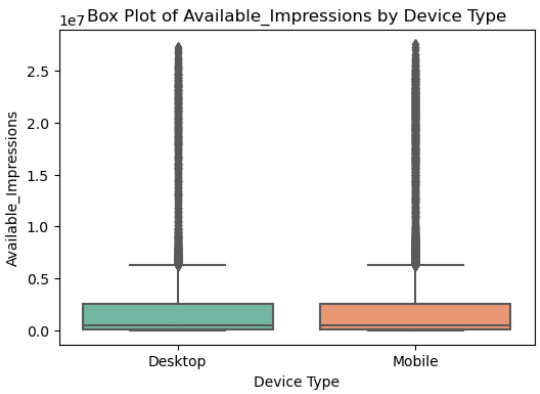


FIGURE 108

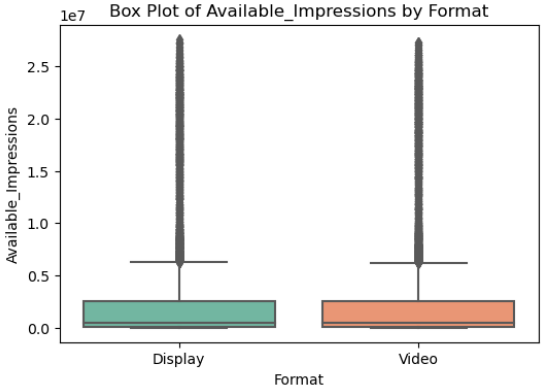


FIGURE 109

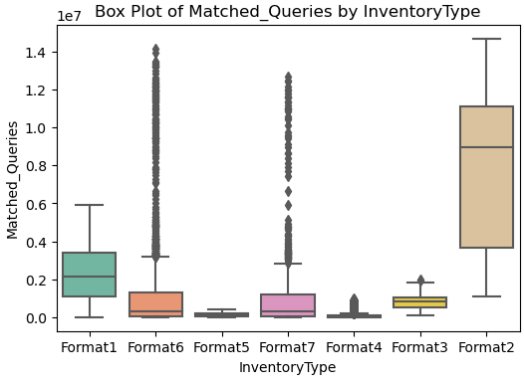


FIGURE 110

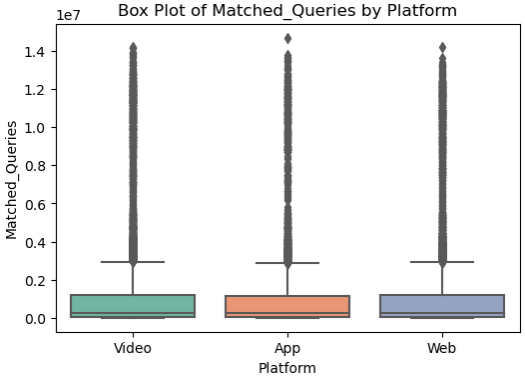


FIGURE 111

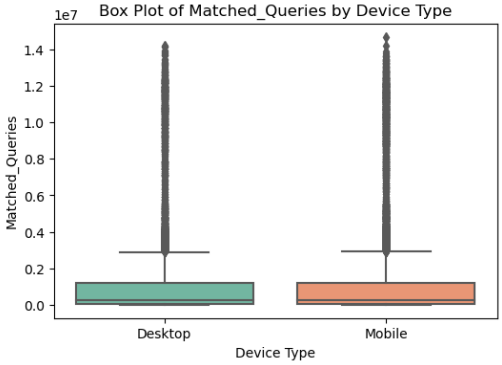


FIGURE 112

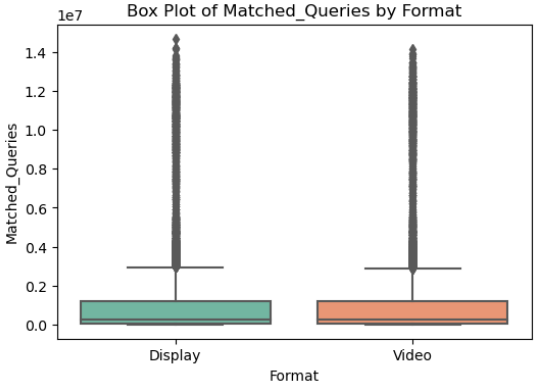


FIGURE 113

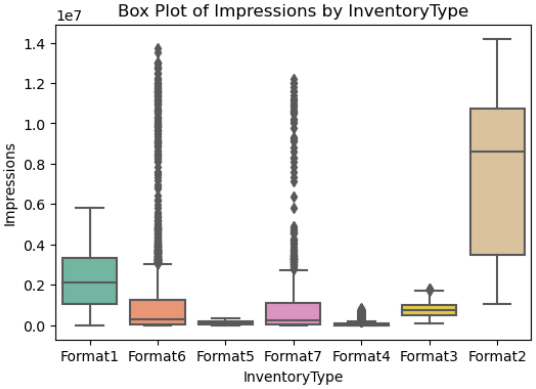


FIGURE 114

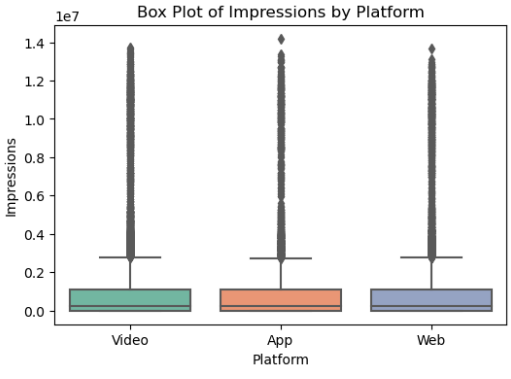


FIGURE 115

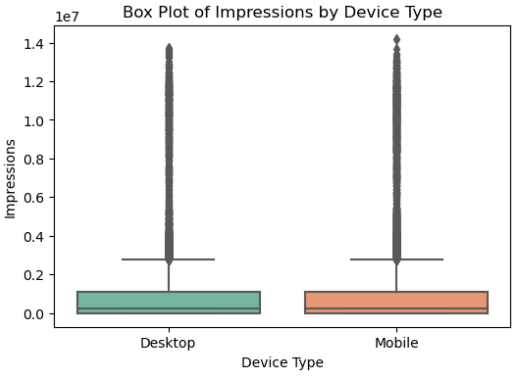


FIGURE 116

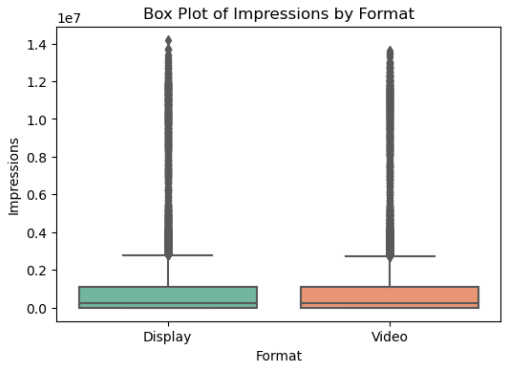


FIGURE 117

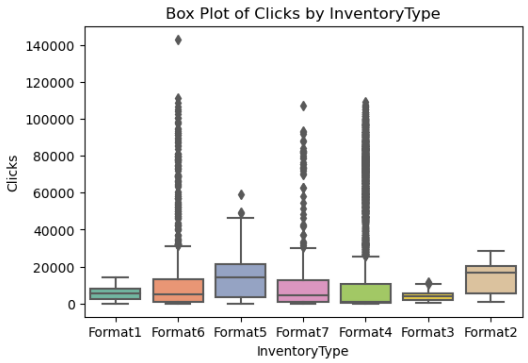


FIGURE 118

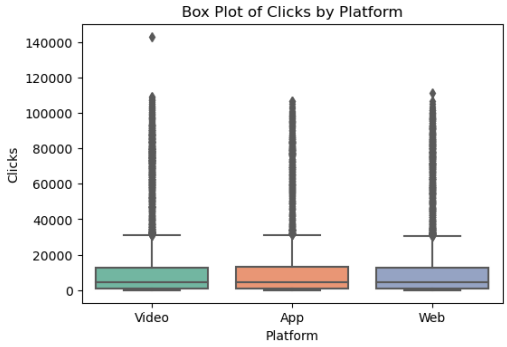


FIGURE 119

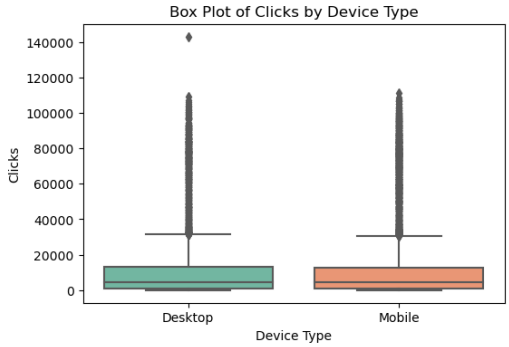


FIGURE 120

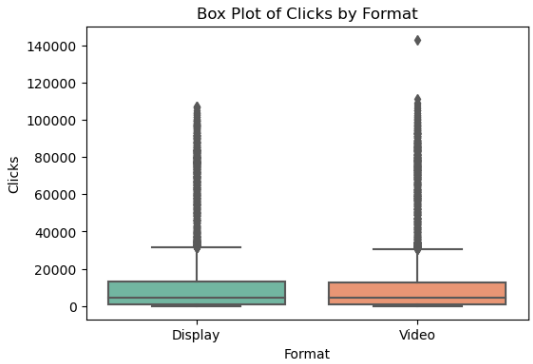


FIGURE 121

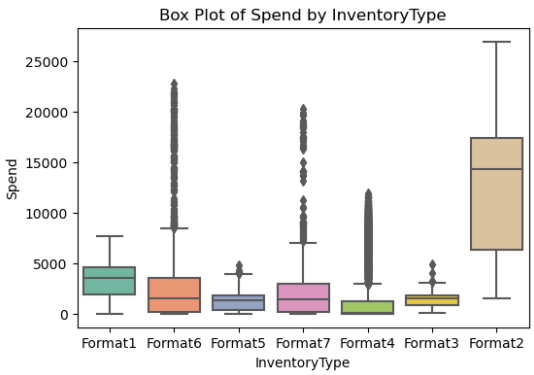


FIGURE 122

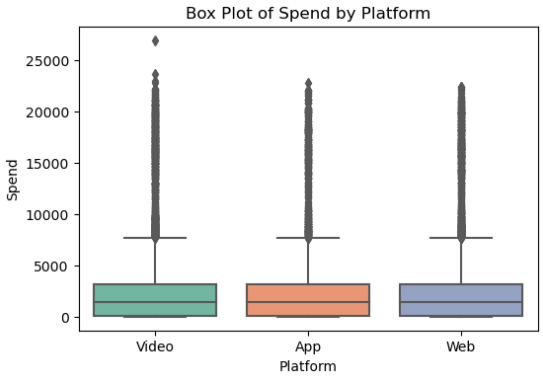


FIGURE 123

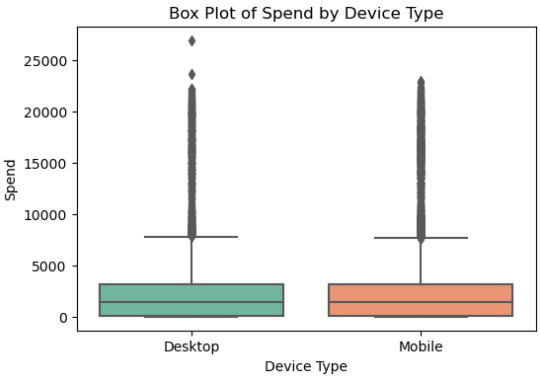


FIGURE 124

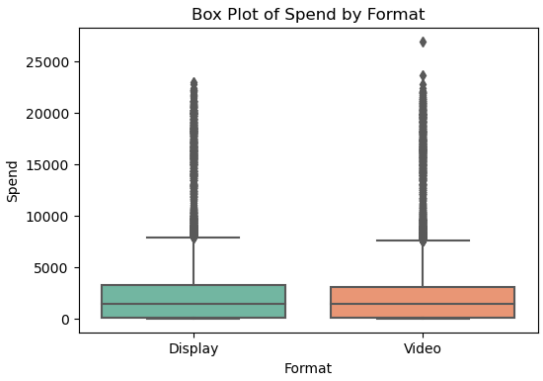


FIGURE 125

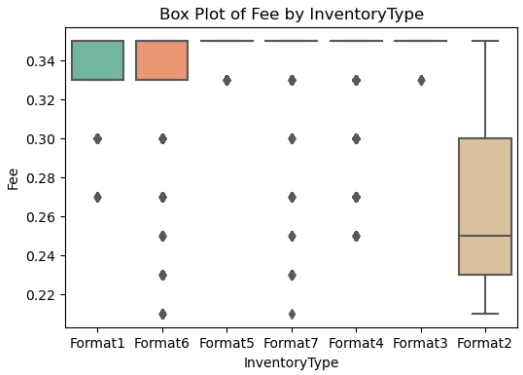


FIGURE 126

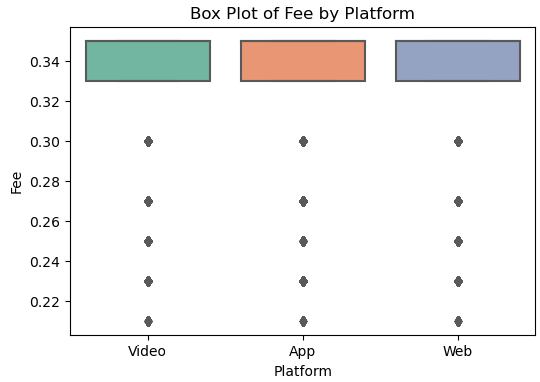


FIGURE 127

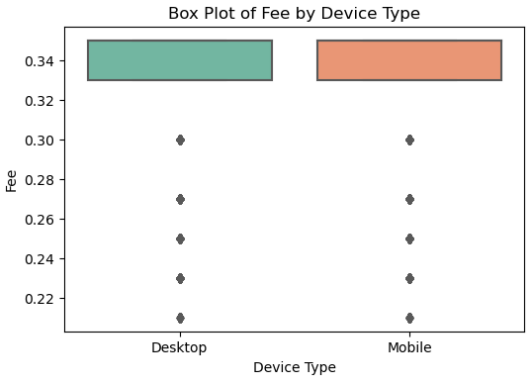


FIGURE 128

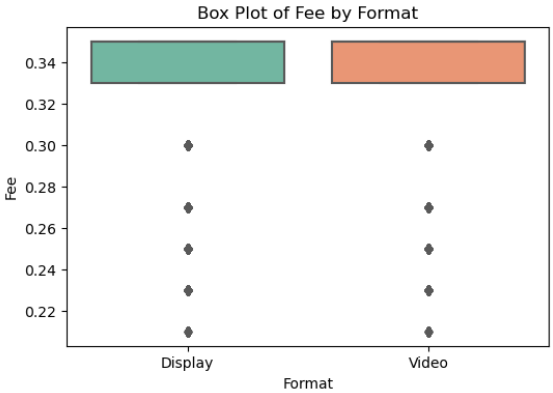


FIGURE 129

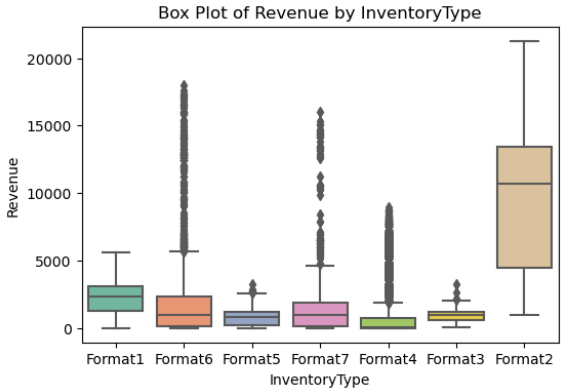


FIGURE 130

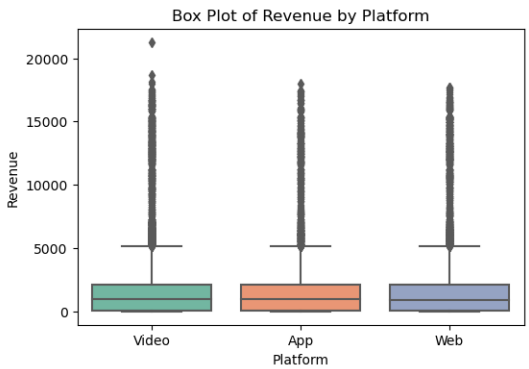


FIGURE 131

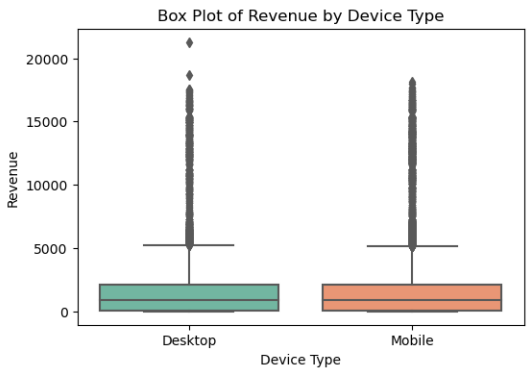


FIGURE 132

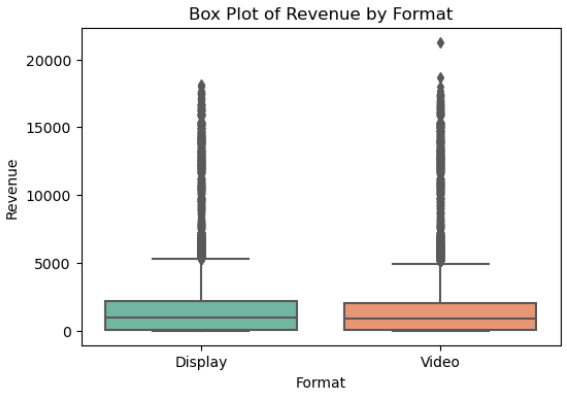


FIGURE 133

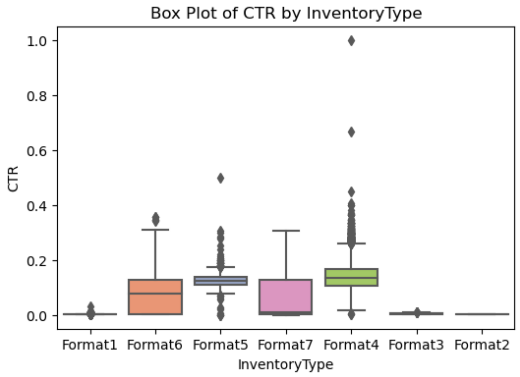


FIGURE 134

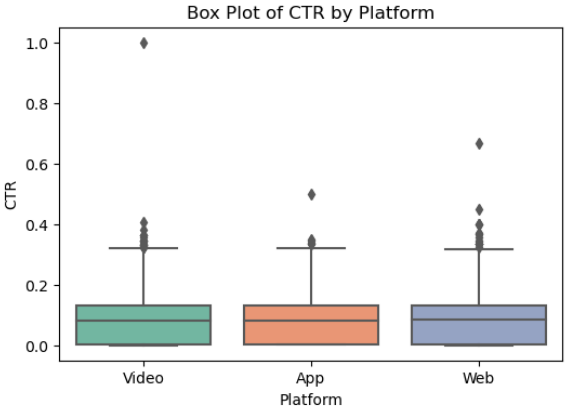


FIGURE 135

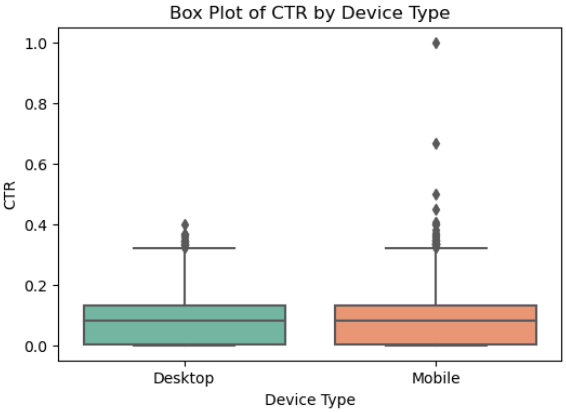


FIGURE 136

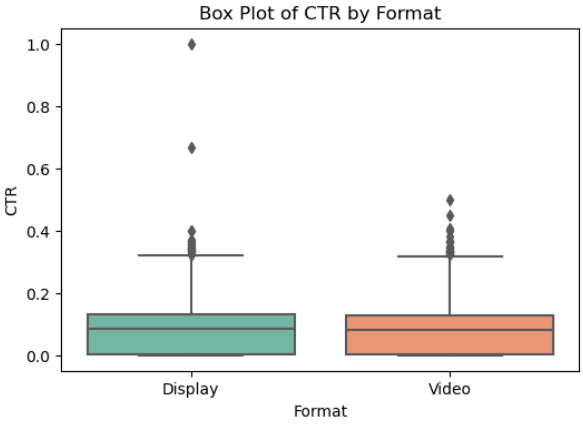


FIGURE 137

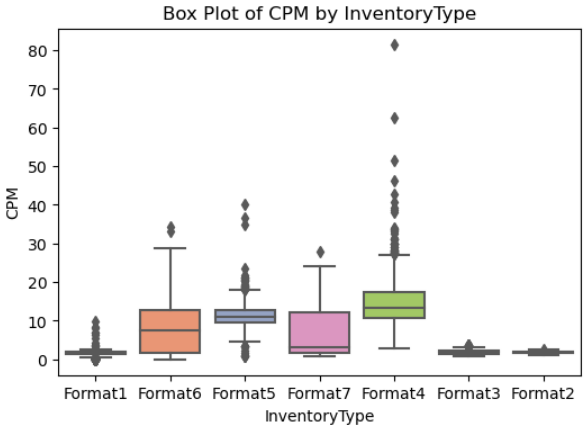
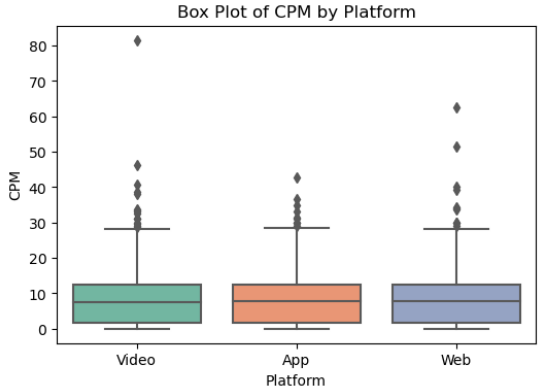


FIGURE 138



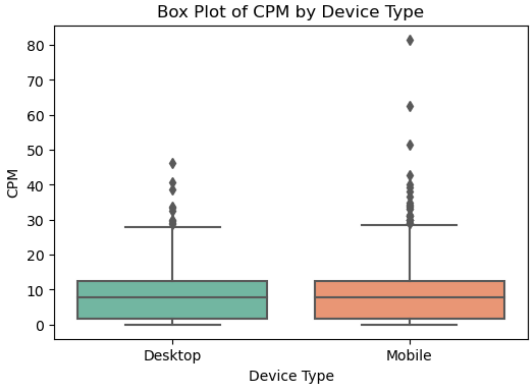


FIGURE 139

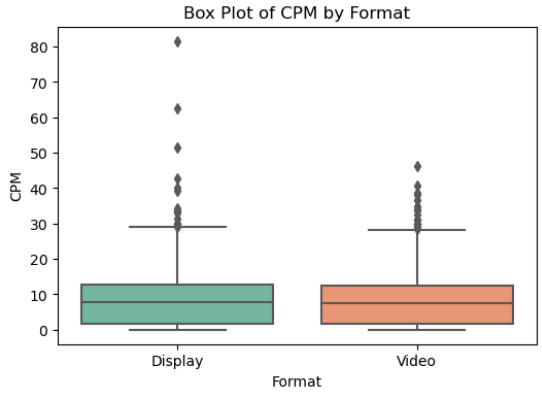


FIGURE 140

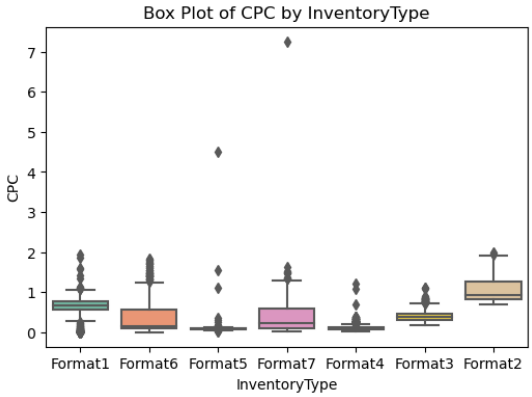


FIGURE 141

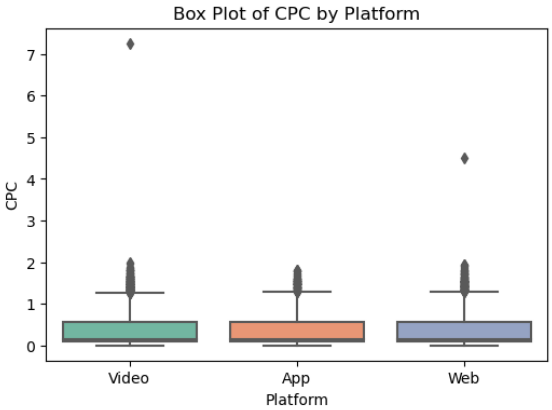


FIGURE 142

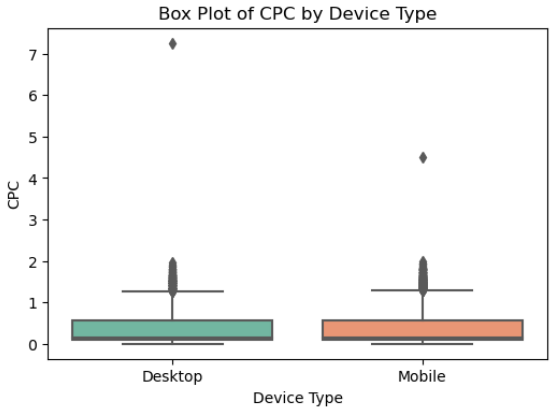


FIGURE 143

FIGURE 144

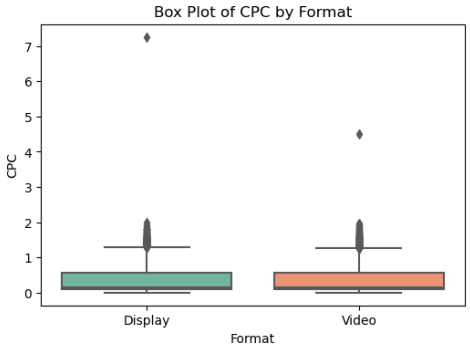


FIGURE 145

Insights:

Variability and Outliers:

The box plots reveal significant variability in numerical data across different categories. For instance, 'Ad Size' varies widely when categorized by 'InventoryType', with some categories showing a larger spread and potential outliers. Outliers are present in most numerical categories, as indicated by the points beyond the whiskers of the box plots. This is particularly noticeable in 'Clicks', 'Spend', and 'Revenue' across various categories.

Median Comparison:

The median values, represented by the horizontal line within the boxes, differ between categories. For example, the median 'Spend' appears to be higher for certain 'Ad Types' compared to others. 'CTR' medians are relatively low across all categories, suggesting that high CTR are not common.

Interquartile Range (IQR):

The IQR, represented by the height of the boxes, indicates the middle 50% of the data. Categories with taller boxes, such as 'Impressions' by 'Platform', suggest greater variability within the middle 50% of observations. Some categories, like 'Device Type', show a more consistent IQR across numerical data, indicating similar distribution patterns within these categories.

Category Specific Insights:

'Platform' and 'Device Type' categories show distinct patterns in 'Ad Size' and 'Available_Impressions', which could be indicative of platform specific or device specific advertising strategies. 'Format' seems to have a consistent influence on 'CPM' and 'CPC', with 'Video' format generally showing higher values.

Categorical Vs Categorical Data: We are going to plot a Heat Map.

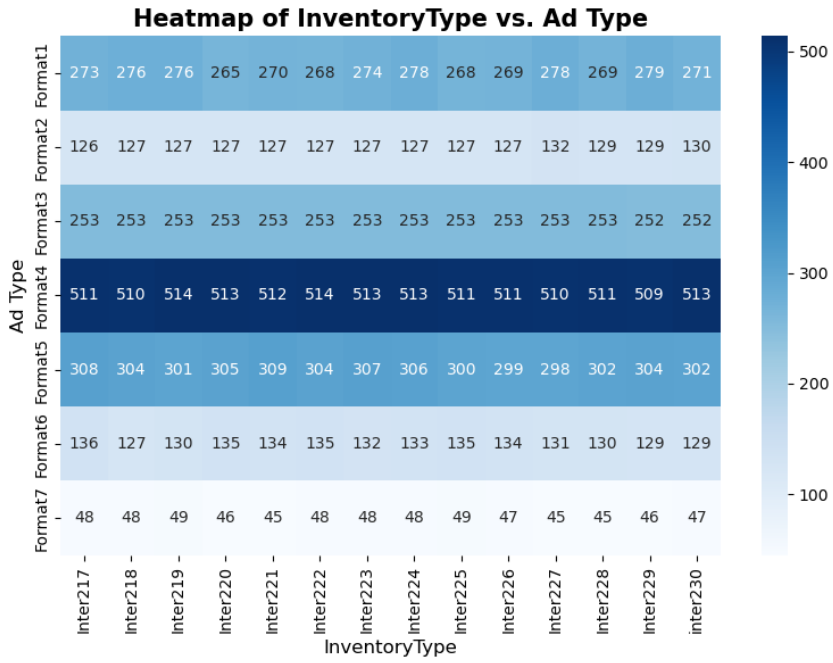


FIGURE 146

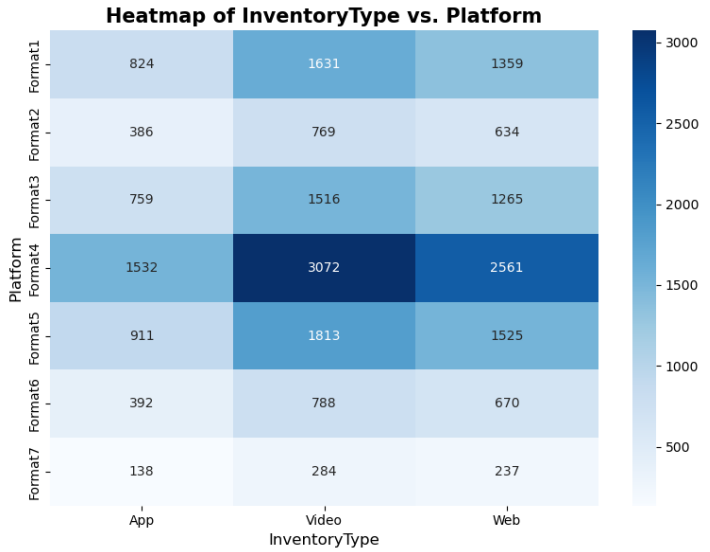


FIGURE 147

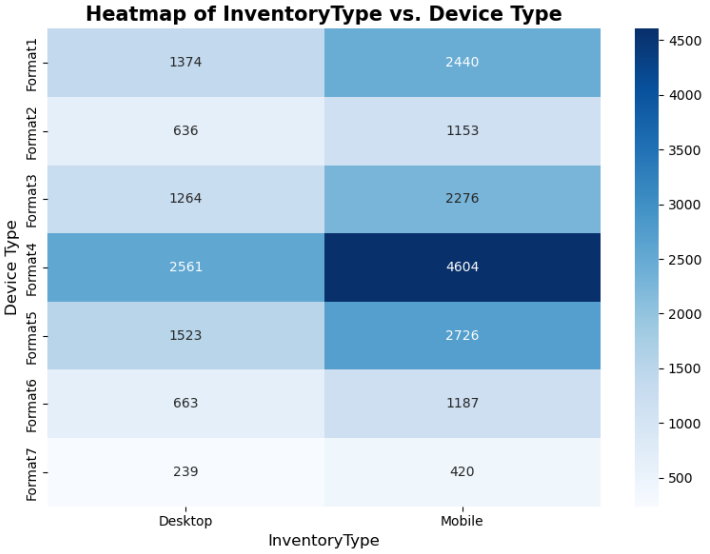


FIGURE 148

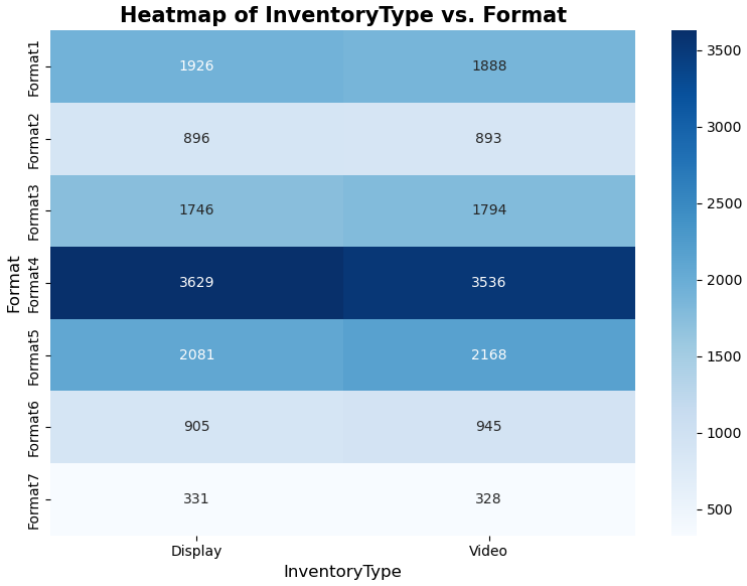


FIGURE 149

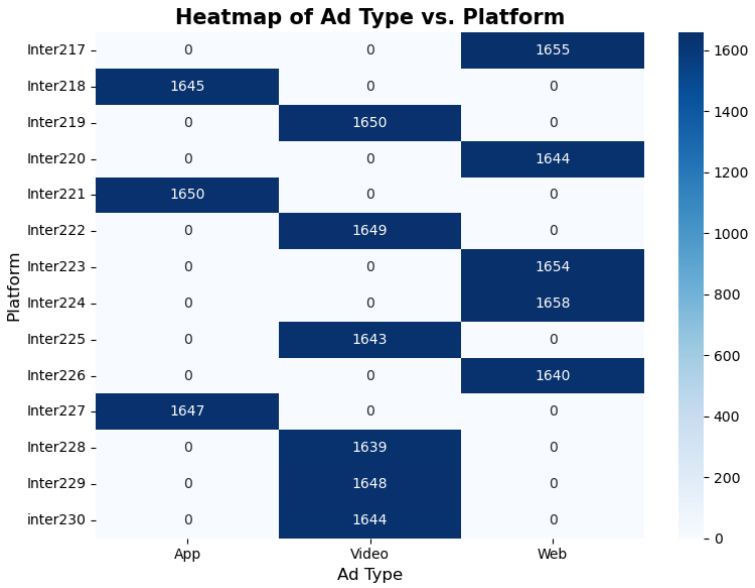


FIGURE 150

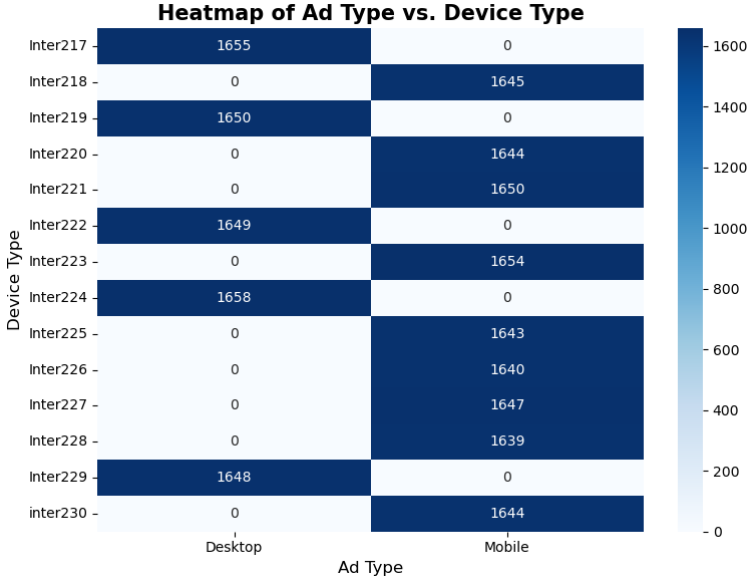


FIGURE 151

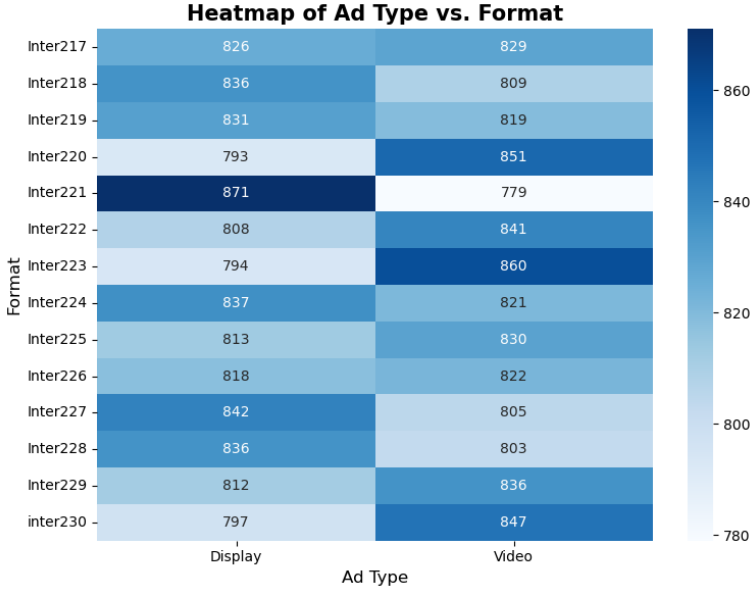


FIGURE 152

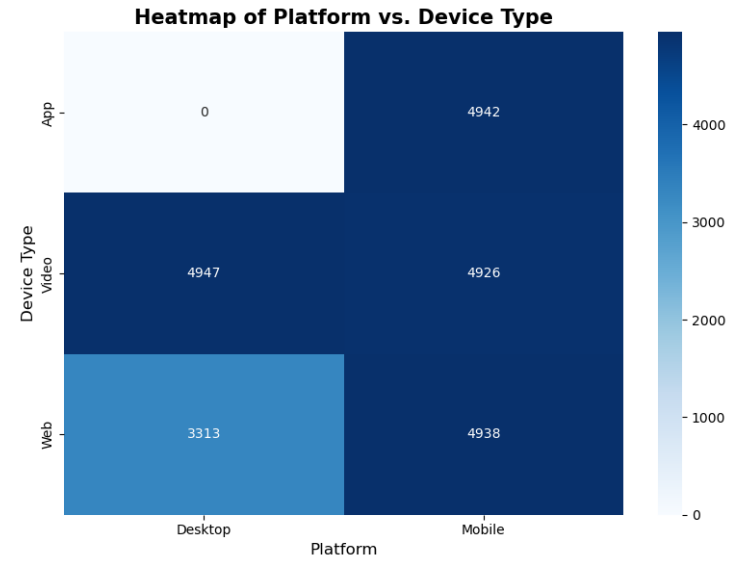


FIGURE 153

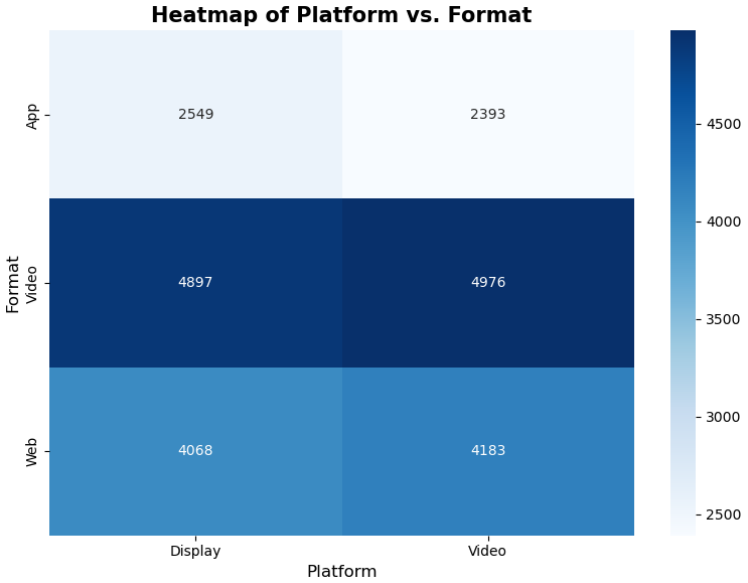


FIGURE 154

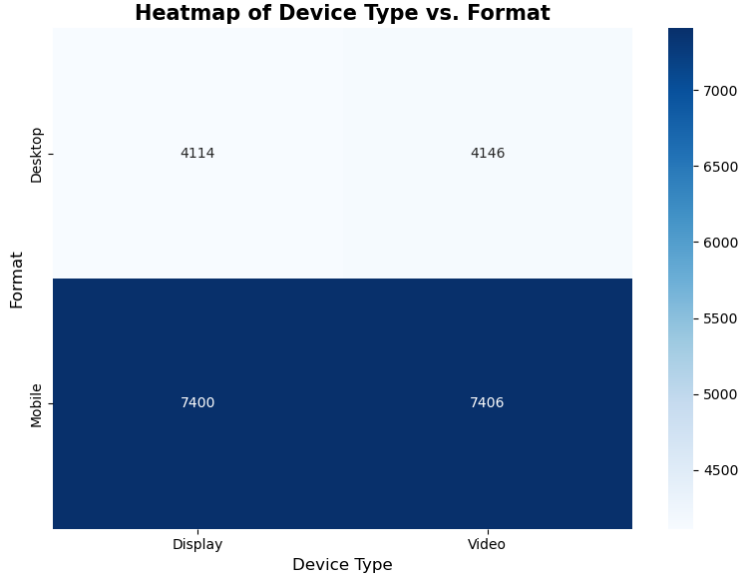


FIGURE 155

Insights:

Inventory Type and Ad Dimensions: The heatmap suggests that certain inventory types may be associated with specific ad dimensions, which could be useful for targeted advertising strategies.

Potential for Segmentation: The varying degrees of correlation between different pairs of variables suggest the potential for market segmentation based on these categorical attributes.

The insights from the heatmap can guide data driven decision-making in advertising campaigns, such as which inventory types to pair with certain ad sizes for optimal engagement.

PROBLEM 1.2 DATA PREPROCESSING:

Problem 1.2.1 Missing value check and treatment:

There are Few variables that has data missing.

```
Timestamp                  0
InventoryType              0
Ad - Length                0
Ad- Width                  0
Ad Size                    0
Ad Type                    0
Platform                   0
Device Type                0
Format                     0
Available_Impressions      0
Matched_Queries            0
Impressions                0
Clicks                     0
Spend                      0
Fee                        0
Revenue                    0
CTR                     4736
CPM                     4736
CPC                     4736
dtype: int64
```

We are going to treat the missing values with following formula:

CPM = (Spend / Impressions) x 1,000.

CPC = Spend / Clicks.

CTR = (Clicks / Impressions) x 100.

We Created a user defined function that does calculation as per the formula and populate the missing values.

Problem 1.2.2 Outlier Treatment:

We Checked Outlier and following observations are made:

1. No outliers were detected in the 'Ad - Length' and 'Ad - Width' columns within the calculated bounds.
2. Outliers were identified in 'Ad Size', 'Available_Impressions', and 'Matched_Queries' columns, indicating the presence of values that are significantly higher or lower than the rest of the data points.
3. Given the high number of outliers in the dataset, it is recommended to treat outliers before proceeding with K-Means clustering.

For treatment of outlier i going with Standardization Method using Z-score scaling Standardize the data to have a mean of 0 and a standard deviation of 1, which can reduce the impact of outliers.

Problem 1.2.3 Z-Score Scaling:

We need to import zscore from Scipy library. The columns have been standardized with a mean of 0 and a standard deviation of 1.

PROBLEM 1.3 HIERARCHICAL CLUSTERING:

Problem 1.3.1 Construct a dendrogram using Ward linkage and Euclidean distance:

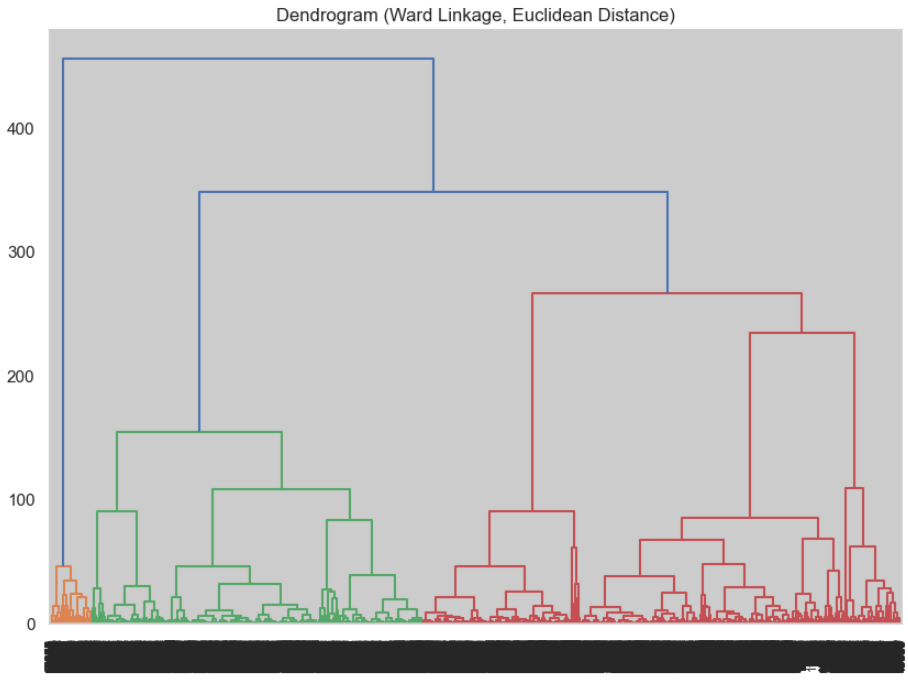


FIGURE 156

Problem 1.3.2 Identify the optimum number of Clusters:

Using the dendrogram, we can see that the optimal number of clusters is 3. This is because there are three main branches in the dendrogram, each representing a distinct cluster.

PROBLEM 1.4 K - MEANS CLUSTERING:

Problem 1.4.1 Apply K-means Clustering:

By Importing KMeans from sklearn library we applied K- Means clustering.

Problem 1.4.2 Plot the Elbow curve:

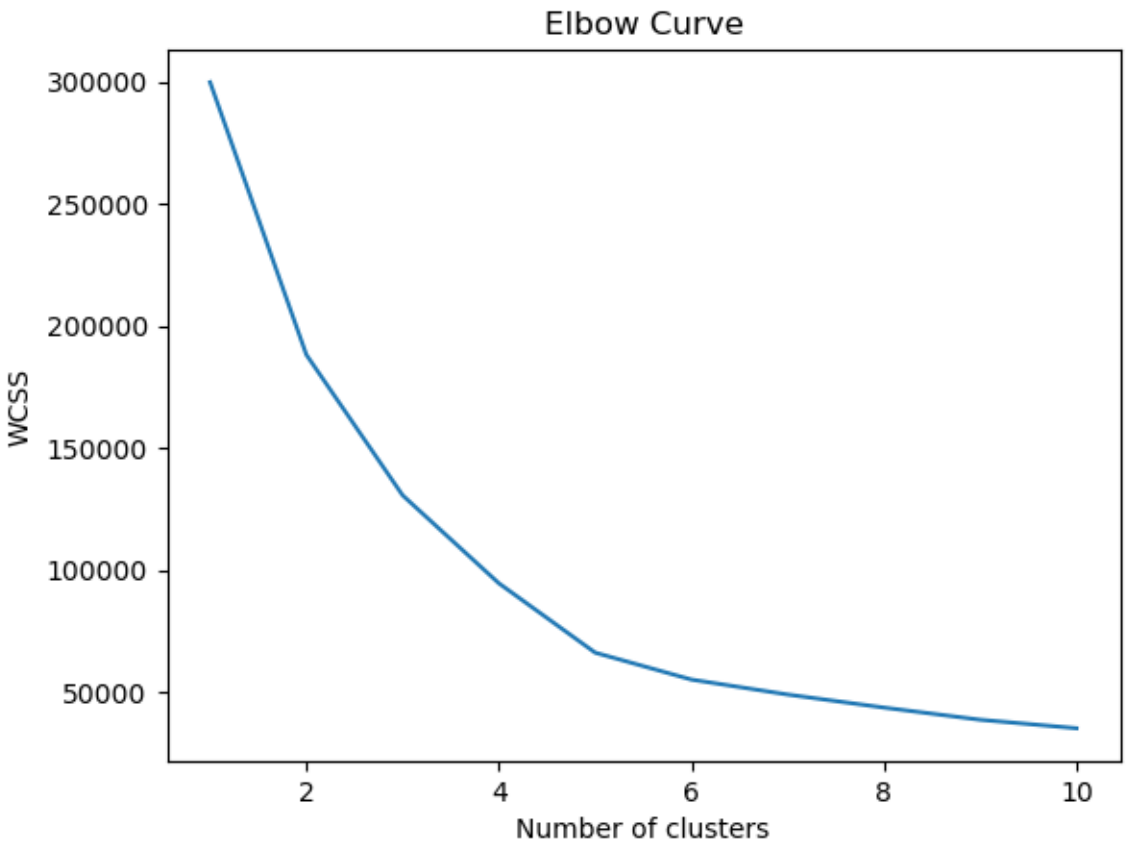


FIGURE 157

Problem 1.4.3 Check Silhouette Scores:

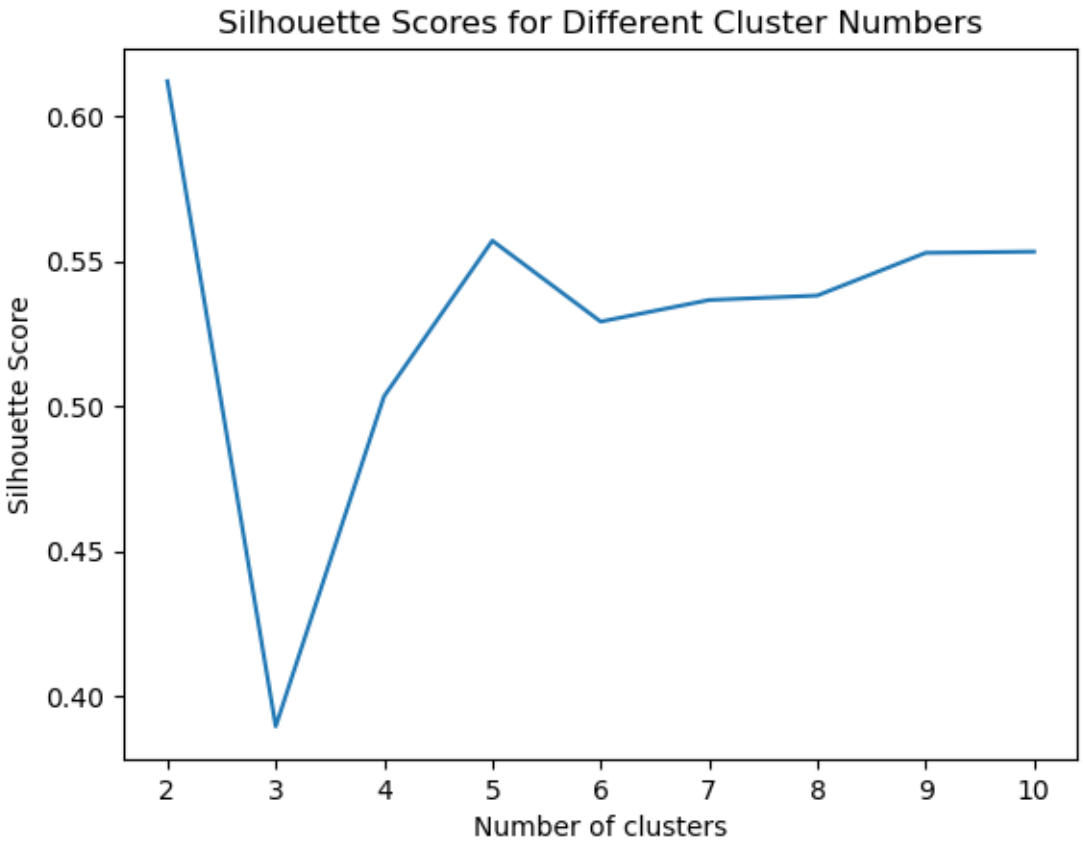


FIGURE 158

Problem 1.4.4 Figure out the appropriate number of clusters:

The elbow curve and silhouette scores suggest that the optimal number of clusters for the data is 3.

Elbow Curve:

1. The WCSS decreases rapidly as the number of clusters increases from 1 to 3.

2. After 3 clusters, the rate of decrease in WCSS slows down, indicating that the optimal number of clusters is likely around 3.

3. Beyond 3 clusters, the WCSS continues to decrease but at a much slower rate, suggesting diminishing returns in terms of clustering quality.

Silhouette Scores:

1. The Silhouette Score is highest for 2 clusters, indicating that the data is best clustered into 2 groups.

2. The score drops significantly for 3 clusters, suggesting that 3 clusters do not fit the data well.

3. The score increases again for 4 clusters and remains relatively stable for 5 to 10 clusters, with slight variations.

Therefore, based on both the elbow curve and silhouette scores, the appropriate number of clusters for the data is 3.

Problem 1.4.5 Cluster Profiling:

```
Cluster 0 Profile:
- Size: 8855
- CTR Mean: -0.8701653579285658
- CPM Mean: -0.7384432265892631
- CPC Mean: 0.7515200364263053
- CTR Standard Deviation: 0.01918687437468589
- CPM Standard Deviation: 0.06858786672336069
- CPC Standard Deviation: 0.948952785795959

Cluster 1 Profile:
- Size: 1361
- CTR Mean: -0.88745386120812
- CPM Mean: -0.7415230248588176
- CPC Mean: 1.6084132982832413
- CTR Standard Deviation: 0.002010287960275988
- CPM Standard Deviation: 0.027531339762669162
- CPC Standard Deviation: 0.3801042053776553

Cluster 2 Profile:
- Size: 12850
- CTR Mean: 0.6936294902382645
- CPM Mean: 0.5874029267144572
- CPC Mean: -0.6882303829975428
- CTR Standard Deviation: 0.8417251728534236
- CPM Standard Deviation: 1.006344307769157
- CPC Standard Deviation: 0.14965912855904195

Cluster 3 Profile:
- Size: 0
- CTR Mean: nan
- CPM Mean: nan
- CPC Mean: nan
- CTR Standard Deviation: nan
- CPM Standard Deviation: nan
- CPC Standard Deviation: nan

Cluster 4 Profile:
- Size: 0
- CTR Mean: nan
- CPM Mean: nan
- CPC Mean: nan
- CTR Standard Deviation: nan
```

- CPM Standard Deviation: nan
- CPC Standard Deviation: nan

Cluster 5 Profile:
- Size: 0
- CTR Mean: nan
- CPM Mean: nan
- CPC Mean: nan
- CTR Standard Deviation: nan
- CPM Standard Deviation: nan
- CPC Standard Deviation: nan

Cluster 6 Profile:
- Size: 0
- CTR Mean: nan
- CPM Mean: nan
- CPC Mean: nan
- CTR Standard Deviation: nan
- CPM Standard Deviation: nan
- CPC Standard Deviation: nan

Cluster 7 Profile:
- Size: 0
- CTR Mean: nan
- CPM Mean: nan
- CPC Mean: nan
- CTR Standard Deviation: nan
- CPM Standard Deviation: nan
- CPC Standard Deviation: nan

Cluster 8 Profile:
- Size: 0
- CTR Mean: nan
- CPM Mean: nan
- CPC Mean: nan
- CTR Standard Deviation: nan
- CPM Standard Deviation: nan
- CPC Standard Deviation: nan

Cluster 9 Profile:
- Size: 0
- CTR Mean: nan
- CPM Mean: nan
- CPC Mean: nan
- CTR Standard Deviation: nan
- CPM Standard Deviation: nan
- CPC Standard Deviation: nan

PROBLEM 1.5 ACTIONABLE INSIGHTS & RECOMMENDATIONS:

Problem 1.5.1 Extract meaningful insights (atleast 3) from the clusters to identify the most effective types of ads, target audiences, or marketing strategies that can be inferred from each segment:

1.High Click-Through Rate (CTR) Segment (Cluster 2)

Size: 12,850 (largest cluster)

CTR Mean: 0.6936

CPM Mean: 0.5874

CPC Mean: -0.6882

CTR Standard Deviation: 0.8417

CPM Standard Deviation: 1.0063

CPC Standard Deviation: 0.1497

Interpretation:

This cluster has the highest CTR mean, indicating that ads targeting this segment are the most engaging and receive the most clicks.

The negative CPC mean suggests that the cost per click is relatively lower compared to the other clusters, making this segment cost effective for driving traffic.

Given the high CTR and relatively lower CPC, marketing strategies should prioritize targeting this segment for ad campaigns aimed at maximizing engagement and minimizing costs.

2.Cost Effective but Lower Engagement Segment (Cluster 0)

Size: 8,855

CTR Mean: -0.8702

CPM Mean: -0.7384

CPC Mean: 0.7515

CTR Standard Deviation: 0.0192

CPM Standard Deviation: 0.0686

CPC Standard Deviation: 0.9490

Interpretation:

This cluster has a lower CTR mean, indicating that ads targeting this segment are less engaging and receive fewer clicks.

The CPM mean is also lower, suggesting that the cost to reach this segment is relatively low.

The high CPC mean indicates that although the cost to reach this audience is low, the cost per click is higher, making it less cost effective for driving clicks.

Marketing strategies should consider alternative approaches for this segment, such as improving ad relevance or content to boost engagement, or focusing on awareness campaigns rather than click-based conversions.

3.High Cost, Low Engagement Segment (Cluster 1)

Size: 1,361

CTR Mean: -0.8875

CPM Mean: -0.7415

CPC Mean: 1.6084

CTR Standard Deviation: 0.0020

CPM Standard Deviation: 0.0275

CPC Standard Deviation: 0.3801

Interpretation:

This cluster also has a low CTR mean, similar to Cluster 0, indicating low engagement.

The CPM mean is low, which means the cost to reach this segment is low.

However, the very high CPC mean indicates that the cost per click is significantly higher, making this segment the most expensive in terms of driving traffic.

Given the low engagement and high CPC, marketing strategies should carefully evaluate the ROI of targeting this segment. It might be more effective to either refine the ad content to better match this audience's interests or reallocate budget to more cost effective segments.

Summary of Strategic Recommendations:

1.Focus on Cluster 2 for Maximum ROI:

Prioritize ad spend on Cluster 2 due to its high engagement and lower cost per click. This cluster offers the best potential for effective and cost-efficient advertising.

2.Optimize Strategies for Cluster 0:

Consider improving the relevance and quality of ads targeting Cluster 0 to boost engagement. Alternatively, use this segment for brand awareness campaigns rather than direct response campaigns.

3.Reevaluate Investment in Cluster 1:

Due to high CPC and low engagement, carefully assess the value of targeting Cluster 1. Look for ways to better align ads with this audience's preferences or consider shifting budget to more effective segments like Cluster 2.

Problem 1.5.2 Based on the clustering analysis and key insights, provide actionable recommendations (atleast 3) to Ads24x7 on how to optimize their digital marketing efforts, allocate budgets efficiently, and tailor ad content to specific audience segments:

Actionable Recommendations:

1. Prioritize Investment in High-Engagement Segment (Cluster 2)

- Allocate a Larger Budget: Given the high CTR and lower CPC in Cluster 2, Ads24x7 should allocate a larger portion of their digital marketing budget to this segment. This will maximize ROI by leveraging the high engagement rates and cost-effectiveness of reaching this audience.

- Tailor Ad Content: Develop and test various creative ad content specifically designed for Cluster Since this segment is highly responsive, ensure the ad content is engaging, visually appealing, and contains strong calls-to-action (CTAs) to drive conversions.

2. Enhance Ad Relevance and Engagement in Low-Cost Segment (Cluster 0)

- Improve Ad Quality: For Cluster 0, which has lower engagement but lower CPM, focus on enhancing the relevance and quality of the ads. Conduct A/B testing to identify which ad elements resonate best with this audience and iterate based on performance data.

- Awareness Campaigns: Utilize this segment for brand awareness campaigns rather than direct response campaigns. Since the cost to reach this audience is low, it’s an opportunity to build brand recognition and familiarity without focusing solely on click-based conversions.

3. Reevaluate and Optimize Spending for High-Cost, Low-Engagement Segment (Cluster 1)

- Assess ROI: Carefully assess the return on investment for Cluster 1 due to its high CPC and low engagement. Determine if the high cost per click is justified by the quality or value of the conversions coming from this segment.

- Refine Targeting and Messaging: If continuing to target Cluster 1, refine the ad targeting and messaging to better align with this audience's interests and needs. Use insights from past campaign performance to create more personalized and compelling ad content.

- Budget Reallocation: Consider reallocating budget from Cluster 1 to more effective segments, such as Cluster 2, to ensure marketing funds are being utilized where they can achieve the greatest impact.

4. Leverage Data-Driven Insights for Continuous Improvement

- Regularly Monitor and Analyze Performance: Continuously monitor the performance of ad campaigns across all clusters. Use data-driven insights to make informed decisions on budget allocation, ad content, and targeting strategies.

- Adjust Strategies Based on Trends: Be agile in adjusting strategies based on the latest performance trends. If a particular cluster shows changes in engagement or cost metrics, be prepared to pivot and reallocate resources accordingly.

5. Implement Advanced Targeting and Personalization

- Utilize Advanced Targeting Techniques: Use advanced targeting options such as behavioral targeting, lookalike audiences, and retargeting to reach the most relevant audience segments. This will help in delivering more personalized and effective ads.

- Personalize Ad Content: Personalize ad content based on the specific preferences and behaviors of each cluster. For example, use different messaging, visuals, and offers for each segment to increase relevance and engagement.

By following these recommendations, Ads24x7 can optimize their digital marketing efforts, allocate budgets more efficiently, and tailor ad content to specific audience segments, ultimately driving better performance and higher ROI.

Problem 2:

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

PROBLEM 2.1 DEFINE THE PROBLEM AND PERFORM EXPLORATORY DATA ANALYSIS:

Problem 2.1.1 Problem Definition:

The task involves analyzing the Primary Census Abstract (PCA) data for female-headed households excluding institutional households from the 2011 Census of India. Given the complexity and volume of the data, it is challenging to extract meaningful insights without employing data science techniques. The goal is to perform detailed Exploratory Data Analysis (EDA) and use Principal Component Analysis (PCA) to identify the optimum principal components that explain the most variance in the dataset. This will help simplify the data while retaining the most important information for further analysis.

Problem 2.1.2 Check shape, Data types, statistical summary:

First, we import all the necessary libraries seaborn, numpy, pandas, sklearn, matplotlib etc. to perform our analysis

Next, we import the data set “PCA_Data”

Data Dictionary:

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

Shape of the data:

Shape of the dataset is 640 rows and 61 Columns.

Data Type:

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
```

There are 59 Numerical Data and 2 Categorical Data.

Statistical Summary:

	State Code	Dist.Code	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	...	MAI
count	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	...	
mean	17.114062	320.500000	51222.871875	79940.576563	122372.084375	12309.098438	11942.300000	13820.946875	20778.392188	6191.807813	...	
std	9.426486	184.896367	48135.405475	73384.511114	113600.717282	11500.906881	11326.294567	14426.373130	21727.887713	9912.668948	...	
min	1.000000	1.000000	350.000000	391.000000	698.000000	56.000000	56.000000	0.000000	0.000000	0.000000	...	
25%	9.000000	160.750000	19484.000000	30228.000000	46517.750000	4733.750000	4672.250000	3466.250000	5603.250000	293.750000	...	
50%	18.000000	320.500000	35837.000000	58339.000000	87724.500000	9159.000000	8663.000000	9591.500000	13709.000000	2333.500000	...	
75%	24.000000	480.250000	68892.000000	107918.500000	164251.750000	16520.250000	15902.250000	19429.750000	29180.000000	7658.000000	...	
max	35.000000	640.000000	310450.000000	485417.000000	750392.000000	96223.000000	95129.000000	103307.000000	156429.000000	96785.000000	...	

8 rows × 59 columns

Insights:

1. The dataset contains 640 entries (rows) and 61 features (columns).

2. We have 59 features that are numerical and 2 categorical type data.

Problem 2.1.3 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:

Selected Variable:

1. No_HH (Number of Households)
2. TOT_M (Total Male Population)
3. TOT_F (Total Female Population)
4. M_LIT (Male Literates)
5. F_LIT (Female Literates)

```
Summary Statistics:
                No_HH          TOT_M          TOT_F          M_LIT  \
count      640.000000     640.000000     640.000000     640.000000
mean     51222.871875   79940.576563  122372.084375   57967.979688
std      48135.405475   73384.511114  113600.717282   55910.282466
min        350.000000     391.000000     698.000000     286.000000
25%      19484.000000   30228.000000   46517.750000   21298.000000
50%      35837.000000   58339.000000   87724.500000   42693.500000
75%      68892.000000  107918.500000  164251.750000   77989.500000
max     310450.000000  485417.000000  750392.000000  403261.000000

               F_LIT
count     640.000000
mean    66359.565625
std     75037.860207
min       371.000000
25%     20932.000000
```

```
50%     43796.500000
75%     84799.750000
max    571140.000000


Missing Values:
No_HH    0
TOT_M    0
TOT_F    0
M_LIT    0
F_LIT    0
dtype: int64

Correlation Matrix:
           No_HH     TOT_M     TOT_F     M_LIT     F_LIT
No_HH   1.000000  0.916170  0.970590  0.931938  0.928087
TOT_M   0.916170  1.000000  0.982640  0.989312  0.931708
TOT_F   0.970590  0.982640  1.000000  0.985441  0.957012
M_LIT   0.931938  0.989312  0.985441  1.000000  0.967956
F_LIT   0.928087  0.931708  0.957012  0.967956  1.000000
```

Summary Statistics:

Number of Households (No_HH)	Mean: 51,222.87 Std Dev: 48,135.41 Min: 350 Max: 310,450
Total Male Population (TOT_M)	Mean: 79,940.58 Std Dev: 73,384.51 Min: 391 Max: 485,417
Total Female Population (TOT_F)	Mean: 122,372.08 Std Dev: 113,600.72 Min: 698 Max: 750,392
Male Literates (M_LIT)	Mean: 57,967.98 Std Dev: 55,910.28 Min: 286 Max: 403,261
Female Literates (F_LIT):	Mean: 66,359.57

Std Dev: 75,037.86

Min: 371

Max: 571,140

Missing Values: None of the selected columns have missing values.

Histogram Plot

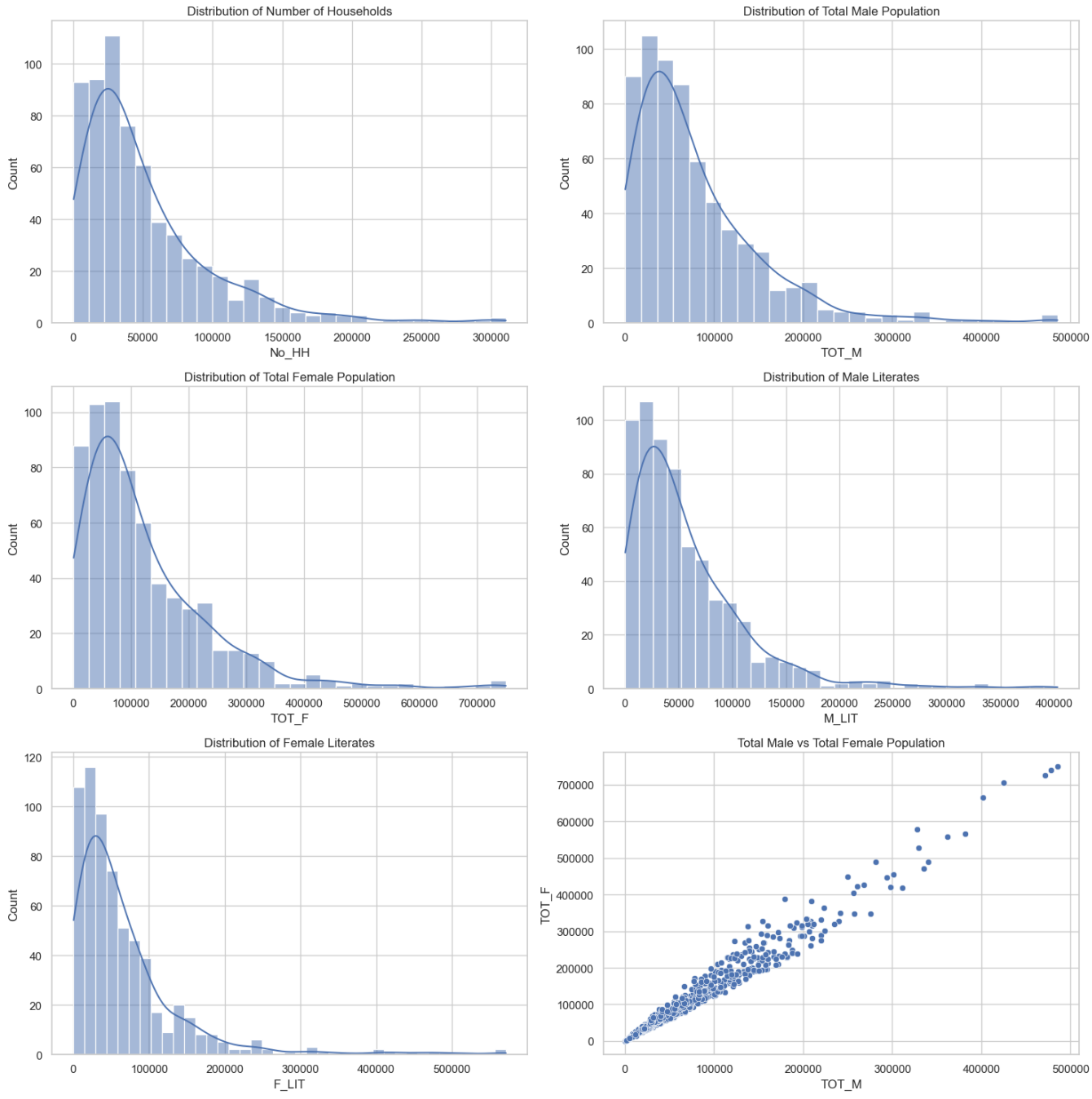


FIGURE 159

Correlation Matrix

Strong positive correlations exist between all pairs of selected variables.

1. No_HH and TOT_M: 0.916
2. No_HH and TOT_F: 0.971
3. No_HH and M_LIT: 0.932
4. No_HH and F_LIT: 0.928
5. TOT_M and TOT_F: 0.983
6. TOT_M and M_LIT: 0.989
7. TOT_M and F_LIT: 0.932
8. TOT_F and M_LIT: 0.985
9. TOT_F and F_LIT: 0.957
10. M_LIT and F_LIT: 0.968

Heat Map:

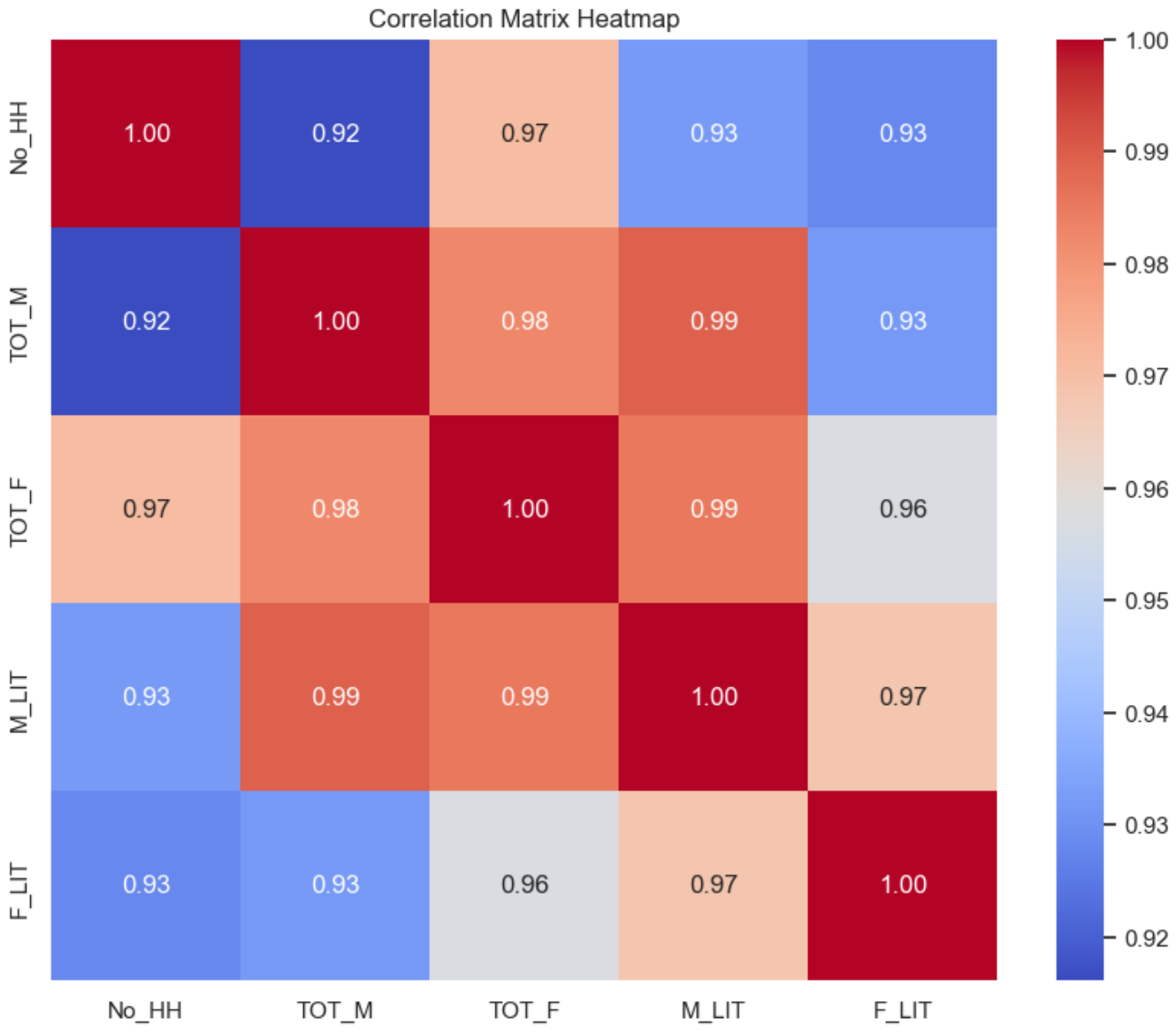


FIGURE 160

Insights:

Household Size and Population: There is a very high correlation between the number of households and the total male and female population, indicating that areas with more households tend to have larger populations.

Literacy Rates: Both male and female literacy rates are highly correlated with each other and with the total population metrics. This suggests that regions with higher populations also tend to have higher numbers of literate individuals.

Gender Distribution: The high correlation between TOT_M and TOT_F indicates a balanced gender distribution in the regions.

Education Trends: The literacy rates for males and females are strongly correlated, implying that efforts to improve literacy in a region tend to benefit both genders similarly.

Problem 2.1.4 Question & Answer:

1. Is there a relationship between the number of households and the total number of males and females?

To analyze the relationship between the number of households and the total number of males and females, we can create a scatter plot with the number of households on the x-axis and the total number of males and females on the y-axis. We can also calculate the correlation coefficient between these two variables.

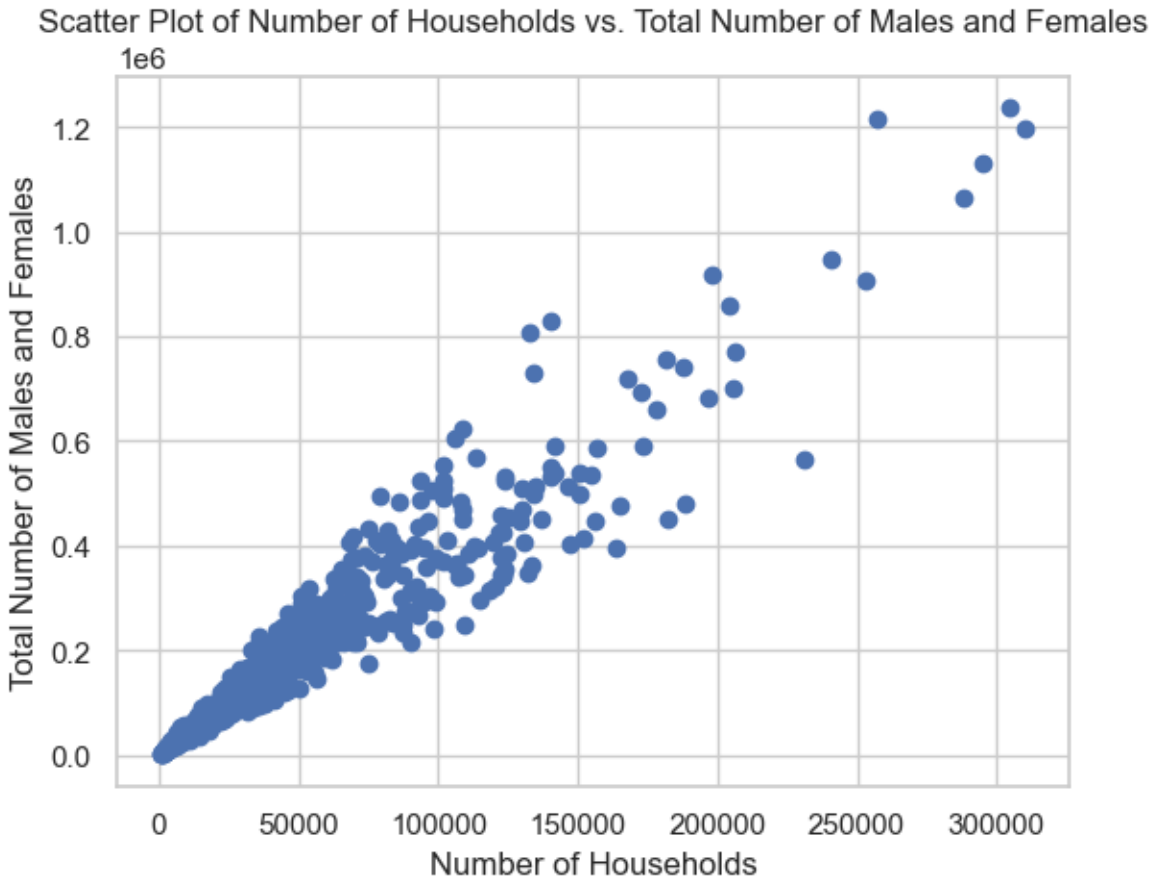


FIGURE 161

There is a very high correlation between the number of households and the total male and female population, indicating that areas with more households tend to have larger populations.

2. Is there a relationship between the number of household members and the number of literate males and females?

To analyze the relationship between the number of households and the total number of literate males and females. We can also calculate the correlation coefficient between these two variables.

```
Correlation coefficient between No_HH and TOT_M + TOT_F: 0.9531856335491954
```

There is a positive relationship between the number of household members and the number of literate males and females. This means that as the number of household members increases, the number of literate males and females also tends to increase. In other words, larger households are more likely to have more literate members. This could be due to a number of factors, such as increased access to education, resources, and support within larger households.

3. What are the most important factors that contribute to household literacy?

Following factors appear to be the most important contributors to household literacy:

Household Size: Larger households tend to have more literate members. This could be due to several reasons, such as increased access to educational resources, peer influence, and shared knowledge within the household.

Number of Literate Males and Females: The number of literate males and females in a household has a positive impact on the overall household literacy rate. This suggests that having literate role models and support within the household can encourage and facilitate literacy among other members.

Literacy Rate of Parents: The literacy rate of parents, particularly mothers, has a significant influence on the literacy outcomes of their children. Households with literate parents are more likely to prioritize education and create a supportive environment for learning.

Access to Educational Resources: The availability of educational resources, such as books, libraries, and schools, within the community plays a crucial role in promoting household literacy. Households with better access to these resources are more likely to have literate members.

Socioeconomic Status: Socioeconomic status is often associated with household literacy. Households with higher socioeconomic status tend to have better access to education, healthcare, and other resources that can support literacy development.

Cultural and Community Norms: Cultural and community norms related to education and literacy can influence household literacy rates. Communities that value education and prioritize literacy are more likely to have households with high literacy rates.

4. Is there a difference in the literacy rates of males and females?

```
Males have a higher literacy rate than females.
```

5. How can we use the insights from our EDA to improve literacy rates in the community?

Based on the insights from our EDA, we can implement the following strategies to improve literacy rates in the community:

1. Focus on early childhood education:
    - Provide access to quality early childhood education programs that emphasize language and literacy development.
    - Train early childhood educators on effective literacy teaching methods.
2. Improve access to books and reading materials:
    - Establish community libraries and book clubs.
    - Partner with local schools and organizations to distribute books to children in need.
    - Encourage parents to read to their children regularly.
3. Address socioeconomic factors:
    - Provide financial assistance to families struggling to afford educational resources.
    - Offer after-school and summer programs that provide academic support and enrichment activities.
    - Collaborate with community organizations to address social and economic barriers to literacy.
4. Empower parents and caregivers:
    - Conduct workshops and trainings for parents and caregivers on how to support their children's literacy development.
    - Provide resources and information on the importance of early literacy.
5. Monitor and evaluate progress:
    - Regularly assess literacy rates in the community.
    - Track the progress of individual students and families.
    - Use data to inform and adjust our strategies as needed.

PROBLEM 2.1 DATA PROCESSING:

Problem 2.1.1 Missing value check and treatment:

There are no missing values in the dataset.

Problem 2.1.2 Check for and treat (if needed) data irregularities:

We have to:

1. Check for Duplicate value
2. Check for Outliers

There are no missing values in the dataset.

By help of Box plot we can see so outliers are present in the data set (Pls. Refer Jupyter Notebook for all Plot image).

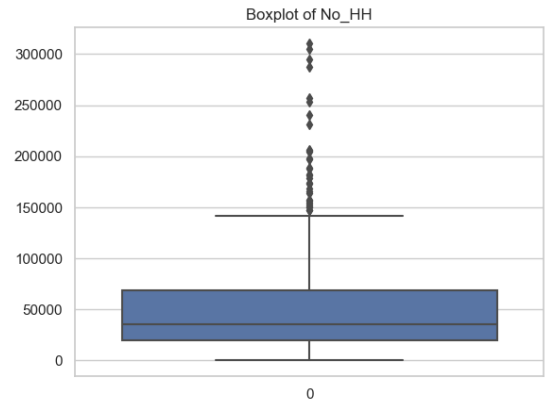


FIGURE 162

So, We are going to treat outlier using IQR Method.

(Pls. Refer Jupyter Notebook for Plot image after treatment)

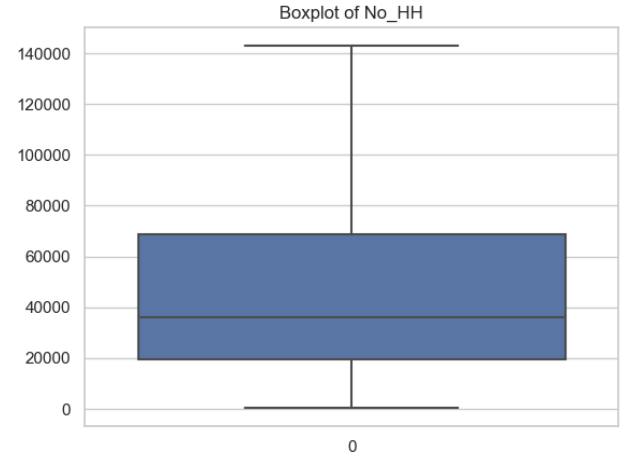


FIGURE 163

Problem 2.1.3 Scale the Data using the z-score method:

We need to import StandardScaler from Sklearn library.

	State Code	Dist.Code	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	...	MARG_CL_0_3_M	MARG_CL_0_3_F	MARG_AI
0	-1.710782	-1.729347	-1.038986	-0.874837	-0.937027	-0.624685	-0.561282	-1.080201	-1.079963	-0.510440	...	-0.093587	-0.860882	-(
1	-1.710782	-1.723934	-1.076896	-0.938023	-1.009723	-0.773932	-0.835657	-1.079873	-1.079635	-0.771833	...	-0.719169	-0.877096	-(
2	-1.710782	-1.718521	-1.121858	-1.154665	-1.141539	-1.141642	-1.138104	-1.080201	-1.079635	0.122588	...	-1.130551	-1.128423	-(
3	-1.710782	-1.713109	-1.201599	-1.217171	-1.214930	-1.197772	-1.176091	-1.080447	-1.079963	-0.399531	...	-1.050477	-1.100286	-(
4	-1.710782	-1.707696	-0.938495	-0.921309	-0.935018	-0.700931	-0.740523	-1.078807	-1.078160	0.432534	...	-0.369844	-0.298617	

5 rows × 59 columns

Problem 2.1.4 Visualize the data before and after scaling and comment on the impact on outliers:

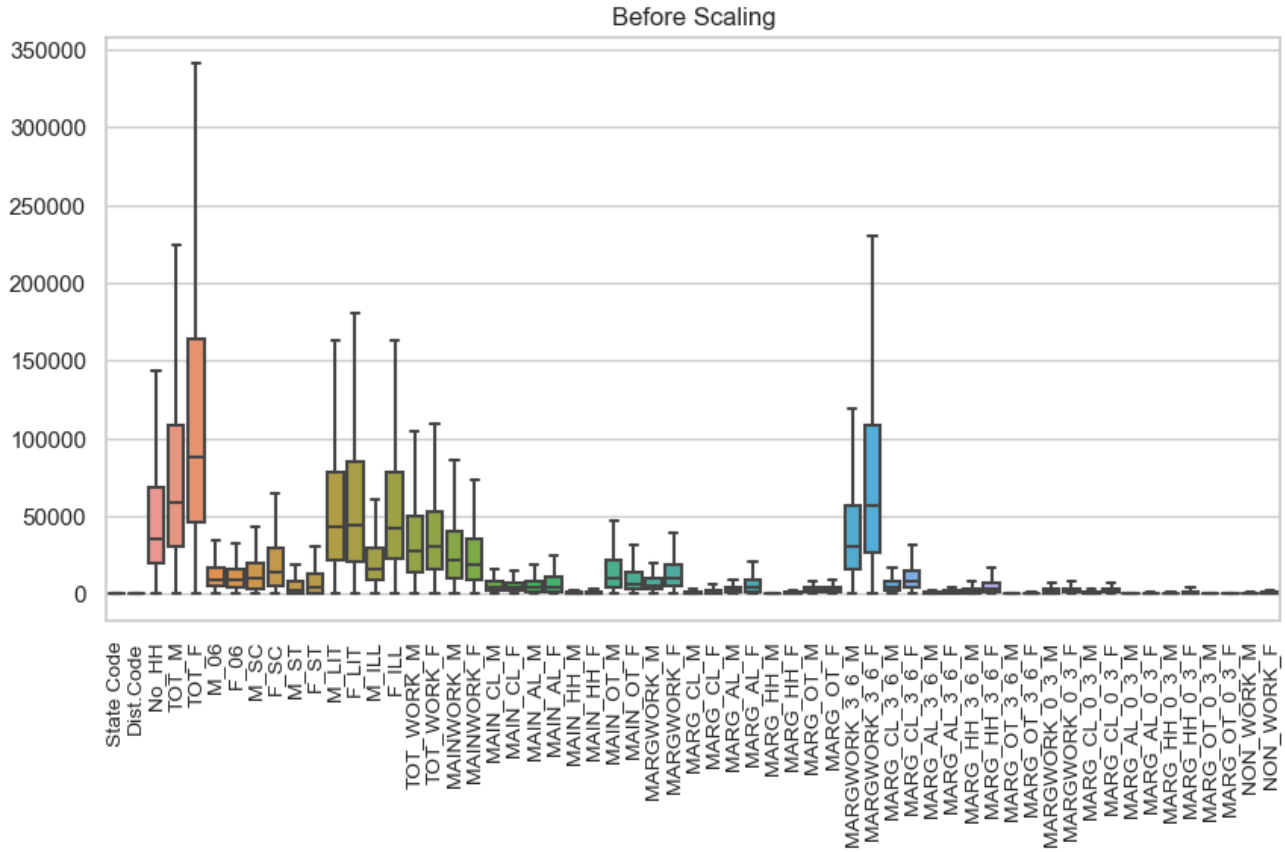


FIGURE 164

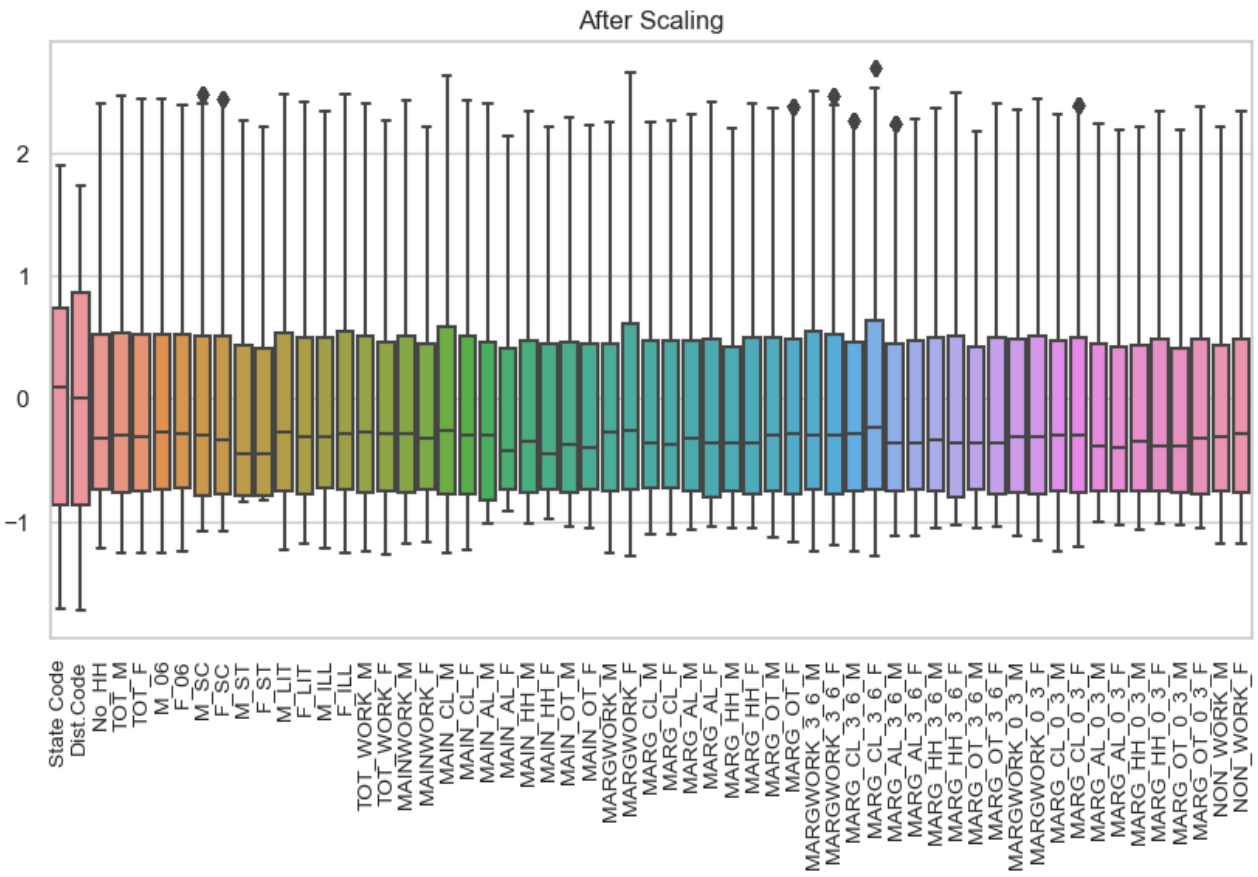


FIGURE 165

Impact of scaling on outlier:

- Outliers are less pronounced after scaling.
- The distribution of each variable is more normal after scaling.
- This suggests that scaling can be an effective way to reduce the impact of outliers on downstream analyses.

PROBLEM 2.3 DATA PROCESSING:

Problem 2.3.1 Create the covariance matrix:

Covariance Matric has been created.

Problem 2.3.2 Get eigen values and eigen vectors:

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


Eigen Vectors:
[[-0.03036889 -0.17219661 -0.30134696 ...  0.02156452  0.00500425
   0.00171484]
 [-0.03034403 -0.1690169  -0.30641795 ... -0.02250001 -0.00665917
  -0.00213945]
 [-0.14957826 -0.11969182 -0.06824342 ... -0.02710355  0.01252643
   0.02384027]
 ...
 [-0.14106815  0.04130293  0.07350673 ...  0.03029966  0.02554542
   0.02740047]
 [-0.14743391 -0.03753818  0.11220488 ... -0.01073912 -0.00174554
   0.01358457]
 [-0.14214317 -0.03868258  0.01794257 ...  0.00069812 -0.01051117
  -0.0014845 ]]
```

Problem 2.3.3 Identify the optimum number of PCs:

```
Number of PCs to explain at least 90% of the variance: 6
```

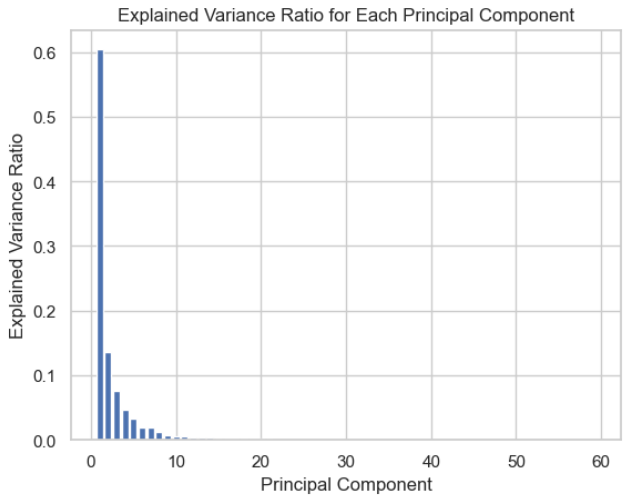


FIGURE 166

Problem 2.3.4 Show Scree plot:

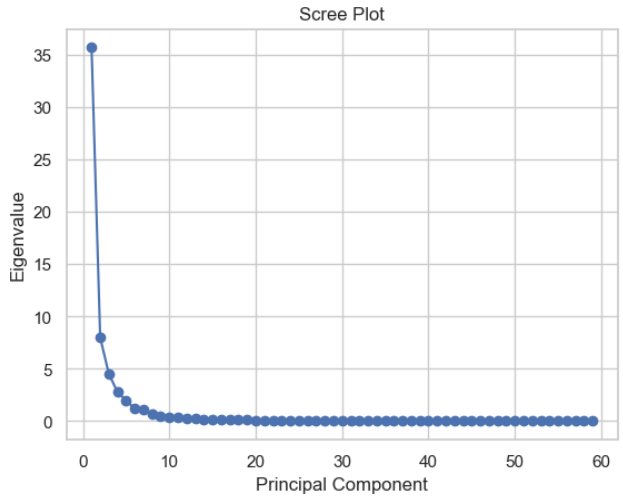


FIGURE 167

Problem 2.3.5 Compare PCs with Actual Columns and identify which is explaining most variance:

```
PC1 explains the most variance for:
- F_ST
PC2 explains the most variance for:
- MARG_AL_0_3_F
PC3 explains the most variance for:
- MAIN_HH_M
PC4 explains the most variance for:
- MARG_AL_M
PC5 explains the most variance for:
- M_ST
PC6 explains the most variance for:
- MARG_OT_3_6_F


PC1 is explaining the most variance overall.
```

Problem 2.3.6 Write inferences about all the PCs in terms of actual variables:

PC1:

- This component has negative loadings for most variables, indicating a general negative association.
- Variables with relatively higher negative loadings include No_HH, TOT_M, TOT_F, M_06, F_06, M_SC, F_SC, 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, MARGWORK_M, MARGWORK_F, MARG_CL_M, MARG_CL_F, MARG_AL_M, MARG_AL_F, MARG_HH_M, MARG_HH_F, MARG_OT_M, MARG_OT_F, MARGWORK_3_6_M, MARGWORK_3_6_F, MARG_CL_3_6_M, MARG_CL_3_6_F, MARG_AL_3_6_M, MARG_AL_3_6_F, MARG_HH_3_6_M, MARG_HH_3_6_F, MARG_OT_3_6_M, MARG_OT_3_6_F, MARGWORK_0_3_M, MARGWORK_0_3_F, MARG_CL_0_3_M, MARG_CL_0_3_F, MARG_AL_0_3_M, MARG_AL_0_3_F, MARG_HH_0_3_M, MARG_HH_0_3_F, MARG_OT_0_3_M, MARG_OT_0_3_F, NON_WORK_M, and NON_WORK_F.

- The negative loadings suggest that as this component increases, the values of these variables tend to decrease together.

PC2:

- This component also has negative loadings for most variables, but less extreme compared to PC1.
- Variables with relatively higher negative loadings include State Code, Dist.Code, No_HH, TOT_M, TOT_F, M_06, F_06, M_SC, F_SC, 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, MARGWORK_M, MARGWORK_F, MARG_CL_M, MARG_CL_F, MARG_AL_M, MARG_AL_F, MARG_HH_M, MARG_HH_F, MARG_OT_M, MARG_OT_F, MARGWORK_3_6_M, MARGWORK_3_6_F, MARG_CL_3_6_M, MARG_CL_3_6_F, MARG_AL_3_6_M, MARG_AL_3_6_F, MARG_HH_3_6_M, MARG_HH_3_6_F, MARG_OT_3_6_M, MARG_OT_3_6_F, MARGWORK_0_3_M, MARGWORK_0_3_F, MARG_CL_0_3_M, MARG_CL_0_3_F, MARG_AL_0_3_M, MARG_AL_0_3_F, MARG_HH_0_3_M, MARG_HH_0_3_F, MARG_OT_0_3_M, MARG_OT_0_3_F, NON_WORK_M, and NON_WORK_F.
- Similarly to PC1, the negative loadings indicate a tendency for variables to decrease together as this component increases.

PC3:

- This component shows a mix of positive and negative loadings.
- Variables with relatively higher loadings include M_ST, F_ST, MARG_AL_0_3_M, and MARG_AL_0_3_F.
- Positive loadings for M_ST and F_ST suggest a positive association between these variables and PC3, while negative loadings for MARG_AL_0_3_M and MARG_AL_0_3_F suggest a negative association.
- Positive loadings for M_ST and F_ST suggest that areas with higher proportions of scheduled tribe (ST) populations (both male and female) are positively associated with this component.
- Negative loadings for MARG_AL_0_3_M and MARG_AL_0_3_F suggest that areas with lower proportions of marginalized populations aged 0-3 (both male and female) are negatively associated with this component.

PC4:

- This component shows a mix of positive and negative loadings.
- Variables with relatively higher loadings include State Code, Dist.Code, MAIN_CL_M, MAIN_CL_F, MAIN_AL_M, MAIN_AL_F, MARG_AL_M, MARG_AL_F, MARG_HH_M, MARG_HH_F, MARG_OT_M, and MARG_OT_F.
- Higher values on this component are associated with areas characterized by higher proportions of marginalized and non-marginalized workers across various sectors (male and female), along with higher proportions of households engaged in other types of work (male and female). Conversely, areas with lower values on this component are associated with lower proportions of these characteristics.

PC5:

- This component shows a mix of positive and negative loadings.
- Variables with relatively higher loadings include M_ST, F_ST, and MARG_HH_0_3_F.
- Higher values on this component are associated with areas with higher proportions of scheduled tribe (ST) populations (both male and female) and a higher proportion of households with children aged 0-3 (female). Conversely, areas with lower values on this component are associated with lower proportions of these characteristics.

PC6:

- This component shows a mix of positive and negative loadings.
- Variables with relatively higher loadings include PC1, PC2, TOT_F, and MAIN_OT_F.
- This component is influenced by multiple factors, including those captured in PC1, PC2, and specific variables related to total female population and main workers in other occupations (female). Higher values on this component may represent areas with a combination of characteristics reflected in these variables, while lower values may represent areas with fewer such characteristics.

Problem 2.3.7 Write linear equation for first PC:

PC1 = - 0.030 * State Code- 0.030 * Dist.Code- 0.150 * No_HH- 0.159 * TOT_M- 0.158 * TOT_F- 0.156 * M_06- 0.156 * F_06- 0.143 * M_SC- 0.143 * F_SC- 0.019 * M_ST- 0.018 * F_ST- 0.155 * M_LIT- 0.146 * F_LIT- 0.154 * M_ILL- 0.158 * F_ILL- 0.154 * TOT_WORK_M- 0.143 * TOT_WORK_F- 0.142 * MAINWORK_M- 0.126 * MAINWORK_F- 0.111 * MAIN_CL_M- 0.083 * MAIN_CL_F- 0.120 * MAIN_AL_M- 0.091 * MAIN_AL_F- 0.142 * MAIN_HH_M- 0.134 * MAIN_HH_F- 0.123 * MAIN_OT_M- 0.117 * MAIN_OT_F- 0.156 * MARGWORK_M- 0.149 * MARGWORK_F- 0.087 * MARG_CL_M- 0.064 * MARG_CL_F- 0.127 * MARG_AL_M- 0.116 * MARG_AL_F- 0.145 * MARG_HH_M- 0.142 * MARG_HH_F- 0.151 * MARG_OT_M- 0.148 * MARG_OT_F- 0.158 * MARGWORK_3_6_M- 0.156 * MARGWORK_3_6_F- 0.157 * MARG_CL_3_6_M- 0.149 * MARG_CL_3_6_F- 0.094 * MARG_AL_3_6_M- 0.066 * MARG_AL_3_6_F- 0.128 * MARG_HH_3_6_M- 0.114 * MARG_HH_3_6_F- 0.145 * MARG_OT_3_6_M- 0.141 * MARG_OT_3_6_F- 0.151 * MARGWORK_0_3_M- 0.148 * MARGWORK_0_3_F- 0.142 * MARG_CL_0_3_M- 0.133 * MARG_CL_0_3_F- 0.062 * MARG_AL_0_3_M- 0.056 * MARG_AL_0_3_F- 0.119 * MARG_HH_0_3_M- 0.113 * MARG_HH_0_3_F- 0.142 * MARG_OT_0_3_M- 0.141 * MARG_OT_0_3_F- 0.147 * NON_WORK_M- 0.142 * NON_WORK_F