Machine Learning 1

PROJECT REPORT

Gowdham S
PGPDSA.O.DEC23.A

Table of Contents

List of Figures

List of Equations

List of Tables

Data Dictionary of Problem 1

Column Name	Description
Timestamp	The Timestamp of the particular Advertisement.
Inventory Type	The Inventory Type of the particular Advertisement. Format 1 to 7. This is a Categorical Variable.
Ad - Length	The Length Dimension of the particular Advertisement.
Ad- Width	The Width Dimension of the particular Advertisement.
Ad Size	The Overall Size of the particular Advertisement. Length*Width.
Ad Type	The type of the particular Advertisement. This is a Categorical Variable.
Platform	The platform in which the particular Advertisement is displayed. Web, Video or App. This is a Categorical Variable.
Device Type	The type of the device which supports the particular Advertisement. This is a Categorical Variable.
Format	The Format in which the Advertisement is displayed. This is a Categorical Variable.
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.
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.
Impressions	The impression count of the particular Advertisement out of the total available impressions.
Clicks	It is a marketing metric that counts the number of times users have clicked on the particular advertisement to reach an online property.
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.
Fee	The percentage of the Advertising Fees payable by Franchise Entities.
Revenue	It is the income that has been earned from the particular advertisement.
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.
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.
CPC	CPC stands for "Cost-per-click". Cost-per-click (CPC) bidding means that you pay for each click on your ads. The Formula used here is CPC = Total Cost (spend) / Number of Clicks. Note that the Total Cost (spend) refers to the 'Spend' Column and the Number of Clicks refers to the 'Clicks' Column.

Data Dictionary of Problem 2

Column Name	Description
State	State Code
District	District Code
Name	Name
TRU1	Area Name
No_HH	No of Household
TOT_M	Total population Male
TOT_F	Total population Female
M_06	Population in the age group 0-6 Male
F_06	Population in the age group 0-6 Female
M_SC	Scheduled Castes population Male
F_SC	Scheduled Castes population Female
M_ST	Scheduled Tribes population Male
F_ST	Scheduled Tribes population Female
M_LIT	Literates population Male
F_LIT	Literates population Female
M_ILL	Illiterate Male
F_ILL	Illiterate Female
TOT_WORK_M	Total Worker Population Male
TOT_WORK_F	Total Worker Population Female
MAINWORK_M	Main Working Population Male
MAINWORK_F	Main Working Population Female
MAIN_CL_M	Main Cultivator Population Male
MAIN_CL_F	Main Cultivator Population Female
MAIN_AL_M	Main Agricultural Labourers Population Male
MAIN_AL_F	Main Agricultural Labourers Population Female
MAIN_HH_M	Main Household Industries Population Male
MAIN_HH_F	Main Household Industries Population Female
MAIN_OT_M	Main Other Workers Population Male
MAIN_OT_F	Main Other Workers Population Female
MARGWORK_M	Marginal Worker Population Male
MARGWORK_F	Marginal Worker Population Female
MARG_CL_M	Marginal Cultivator Population Male
MARG_CL_F	Marginal Cultivator Population Female
MARG_AL_M	Marginal Agriculture Labourers Population Male
MARG_AL_F	Marginal Agriculture Labourers Population Female
MARG_HH_M	Marginal Household Industries Population Male
MARG_HH_F	Marginal Household Industries Population Female
MARG_OT_M	Marginal Other Workers Population Male
MARG_OT_F	Marginal Other Workers Population Female
MARGWORK_3_6_M	Marginal Worker Population 3-6 Male
MARGWORK_3_6_F	Marginal Worker Population 3-6 Female
MARG_CL_3_6_M	Marginal Cultivator Population 3-6 Male
MARG_CL_3_6_F	Marginal Cultivator Population 3-6 Female
MARG_AL_3_6_M	Marginal Agriculture Labourers Population 3-6 Male
MARG_AL_3_6_F	Marginal Agriculture Labourers Population 3-6 Female
MARG_HH_3_6_M	Marginal Household Industries Population 3-6 Male
MARG_HH_3_6_F	Marginal Household Industries Population 3-6 Female
MARG_OT_3_6_M	Marginal Other Workers Population Person 3-6 Male
MARG_OT_3_6_F	Marginal Other Workers Population Person 3-6 Female
MARGWORK_0_3_M	Marginal Worker Population 0-3 Male
MARGWORK_0_3_F	Marginal Worker Population 0-3 Female
MARG_CL_0_3_M	Marginal Cultivator Population 0-3 Male
MARG_CL_0_3_F	Marginal Cultivator Population 0-3 Female
MARG_AL_0_3_M	Marginal Agriculture Labourers Population 0-3 Male
MARG_AL_0_3_F	Marginal Agriculture Labourers Population 0-3 Female
MARG_HH_0_3_M	Marginal Household Industries Population 0-3 Male
MARG_HH_0_3_F	Marginal Household Industries Population 0-3 Female
MARG_OT_0_3_M	Marginal Other Workers Population 0-3 Male
MARG_OT_0_3_F	Marginal Other Workers Population 0-3 Female
NON_WORK_M	Non Working Population Male
NON_WORK_F	Non Working Population Female

Problem 1

Context

Digital Ads Data:

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.

The following three features are commonly used in digital marketing:

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

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

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

Perform the following in given order:

- Read the data and perform basic analysis such as printing a few rows (head and tail), info, data summary, null values duplicate values, etc.
- Treat missing values in CPC, CTR and CPM using the formula given. You have to basically create an user defined function and then call the function for imputing.
- Check if there are any outliers.
- Do you think treating outliers is necessary for K-Means clustering? Based on your judgement decide whether to treat outliers and if yes, which method to employ. (As an analyst your judgement may be different from another analyst).
- Perform z-score scaling and discuss how it affects the speed of the algorithm.
- Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance.
- Make Elbow plot (up to n=10) and identify optimum number of clusters for k-means algorithm.
- Print silhouette scores for up to 10 clusters and identify optimum number of clusters.
- Profile the ads based on optimum number of clusters using silhouette score and your domain understanding
  [Hint: Group the data by clusters and take sum or mean to identify trends in clicks, spend, revenue, CPM, CTR, & CPC based on Device Type. Make bar plots.
- Conclude the project by providing summary of your learnings.

Objective

The objective is to conduct Exploratory Data Analysis (EDA) on the dataset, identifying missing values, duplicates, and outliers. Following this, scaling techniques will be applied to ensure uniform feature contribution. Finally, both Hierarchical and K-means clustering methods will be employed to reveal inherent patterns and groupings within the data, facilitating deeper insights and decision-making.

Problem 1 - Define the problem and perform Exploratory Data Analysis

- Problem definition - Check shape, Data types, statistical summary - Univariate analysis - Bivariate analysis - Key meaningful observations on individual variables and the relationship between variables

Solution:

Ads24x7, a Digital Marketing firm with $10 Million seed funding, aims to utilize Marketing Analytics for segmentation of ad types based on CPM, CPC, and CTR. Initial steps involve data exploration, addressing missing values via predefined formulas, and outlier assessment. Subsequently, z-score scaling enhances clustering, followed by Hierarchical clustering and Elbow plot analysis to determine optimal cluster count for K-means. Silhouette scores refine cluster selection, offering insights crucial for Ads24x7's Marketing Analytics expansion.

- We used shape attribute of pandas library [df1.shape] to determine the number of rows and columns of the dataframe df,

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

Table 1 Shape of Ads Dataset

- To check the types of data, we used info() method of pandas library to list the basic information of the data such as rows count, columns count and datatype of the columns.

Column	Non-Null Count	Data Type
Timestamp	23066 non-null	object
Inventory Type	23066 non-null	object
Ad Length	23066 non-null	int64
Ad Width	23066 non-null	int64
Ad Size	23066 non-null	int64
Ad Type	23066 non-null	object
Platform	23066 non-null	object
Device Type	23066 non-null	object
Format	23066 non-null	object
Available_Impressions	23066 non-null	int64
Matched_Queries	23066 non-null	int64
Impressions	23066 non-null	int64
Clicks	23066 non-null	int64
Spend	23066 non-null	float64
Fee	23066 non-null	float64
Revenue	23066 non-null	float64
CTR	18330 non-null	float64
CPM	18330 non-null	float64
CPC	18330 non-null	float64

Table 2 Basic Information of Ads Dataset

- From the above table, We have **6** object, **7** int and **6** float data types in the dataset. Also we can see that there are **4736** null rows in the CTR, CPM and CPC columns.
- To check if there are any duplicate columns, we used df1.duplicated().sum(). We can see that there are no duplicate rows in the dataset.
- Once the above is done, we found the statistical summary of the dataset using describe method()

Index	Count	Mean	STD	Min	25%	50%	75%	Max
Ad Length	23066	385.16	233.65	120	120	300	720	728
Ad Width	23066	337.9	203.09	70	250	300	600	600
Ad Size	23066	96674.47	61538.33	33600	72000	72000	84000	216000
Available_Impressions	23066	2432044	4742888	1	33672.25	483771	2527712	27592861
Matched_Queries	23066	1295099	2512970	1	18282.5	258087.5	1180700	14702025
Impressions	23066	1241520	2429400	1	7990.5	225290	1112429	14194774
Clicks	23066	10678.52	17353.41	1	710	4425	12793.75	143049
Spend	23066	2706.63	4067.93	0	85.18	1425.12	3121.4	26931.87
Fee	23066	0.34	0.03	0.21	0.33	0.35	0.35	0.35
Revenue	23066	1924.25	3105.24	0	55.37	926.34	2091.34	21276.18
CTR	18330	0.07	0.08	0	0	0.08	0.13	1
CPM	18330	7.67	6.48	0	1.71	7.66	12.51	81.56
CPC	18330	0.35	0.34	0	0.09	0.16	0.57	7.26

Table 3 Statistical Summary of Ads Dataset

- Before proceeding with the Univariate and Bivariate analysis, we will first treat the missing values in CTR, CPM and CPC columns with formula given in the problem statement. We created three functions which will impute the missing values in CTR, CPM and CPC and we applied these functions to the ads data.

CPM = (Total Campaign Spend / Number of Impressions) * 1,000

CPC = Total Cost (spend) / Number of Clicks

CTR = Total Measured Clicks / Total Measured Ad Impressions x 100

Equation 1 Formulae of CPM, CPC & CTR

- After applying the custom functions for imputing missing values we can see that there are no missing values in the CPM, CPC and CTR columns and we verified this using isnull() method.
- Since Timestamp is of no use for our analysis, we can drop the Timestamp column using drop() method.
- First, we selected the numerical and categorical columns from the dataframe and assigned them to the variables using select_dtypes(). We then proceeded with the univariate and bivariate analysis.

Variable Name	Columns
Numerical Cols	'Ad Length', 'Ad Width', 'Ad Size', 'Available_Impressions', 'Matched_Queries', 'Impressions', 'Clicks', 'Spend', 'Fee', 'Revenue', 'CTR', 'CPM', 'CPC'
Categorical Cols	'InventoryType', 'Ad Type', 'Platform', 'Device Type', 'Format'

Table 4 Numerical & Categorical Columns of Ads Data

- After separating the numerical and categorical columns, we created boxplots for each of the numerical columns using seaborn library
- Then we plotted a histogram for the numerical columns followed by the distribution of categorical columns.

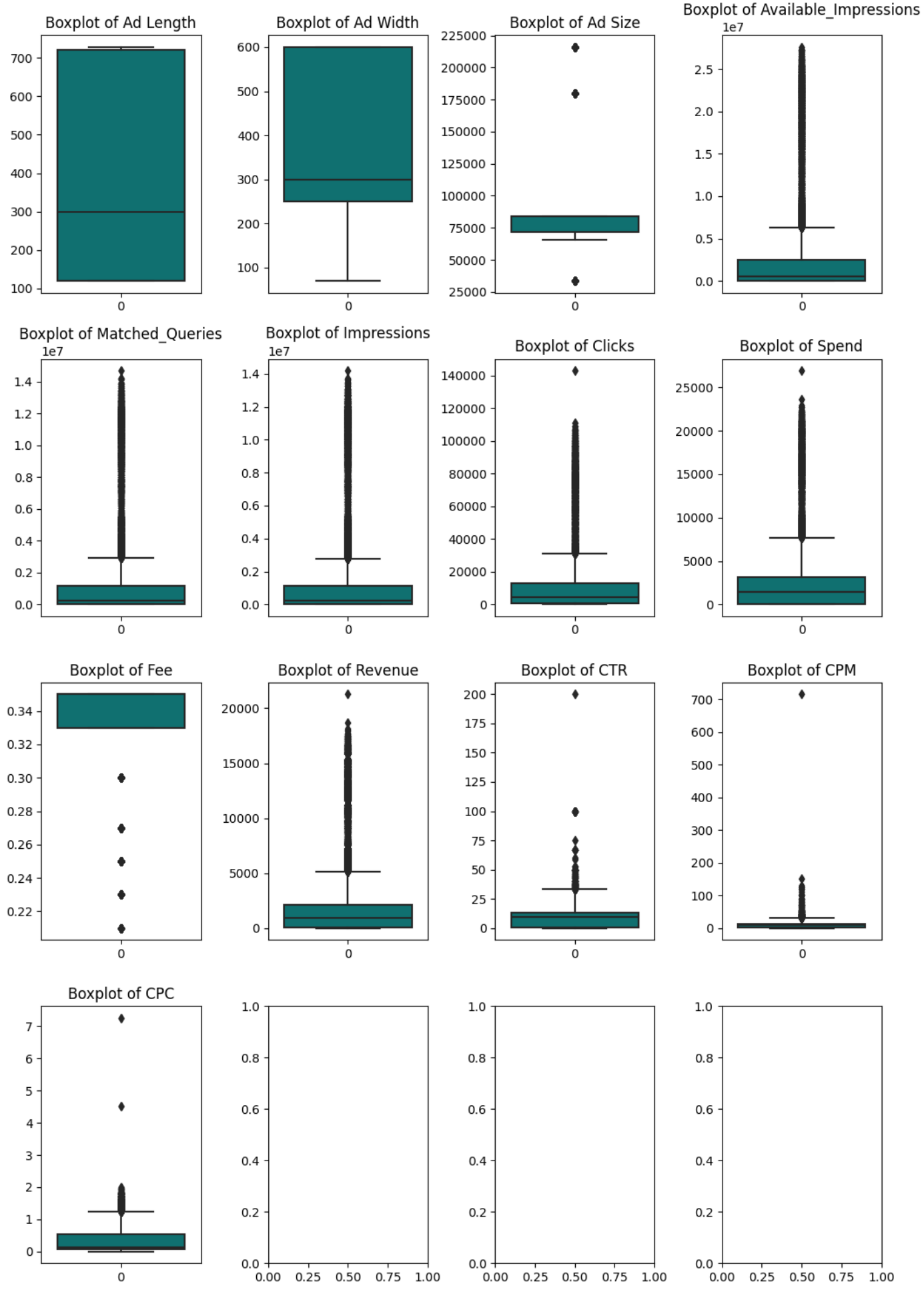


Figure 1 Box Plot of Numerical Columns

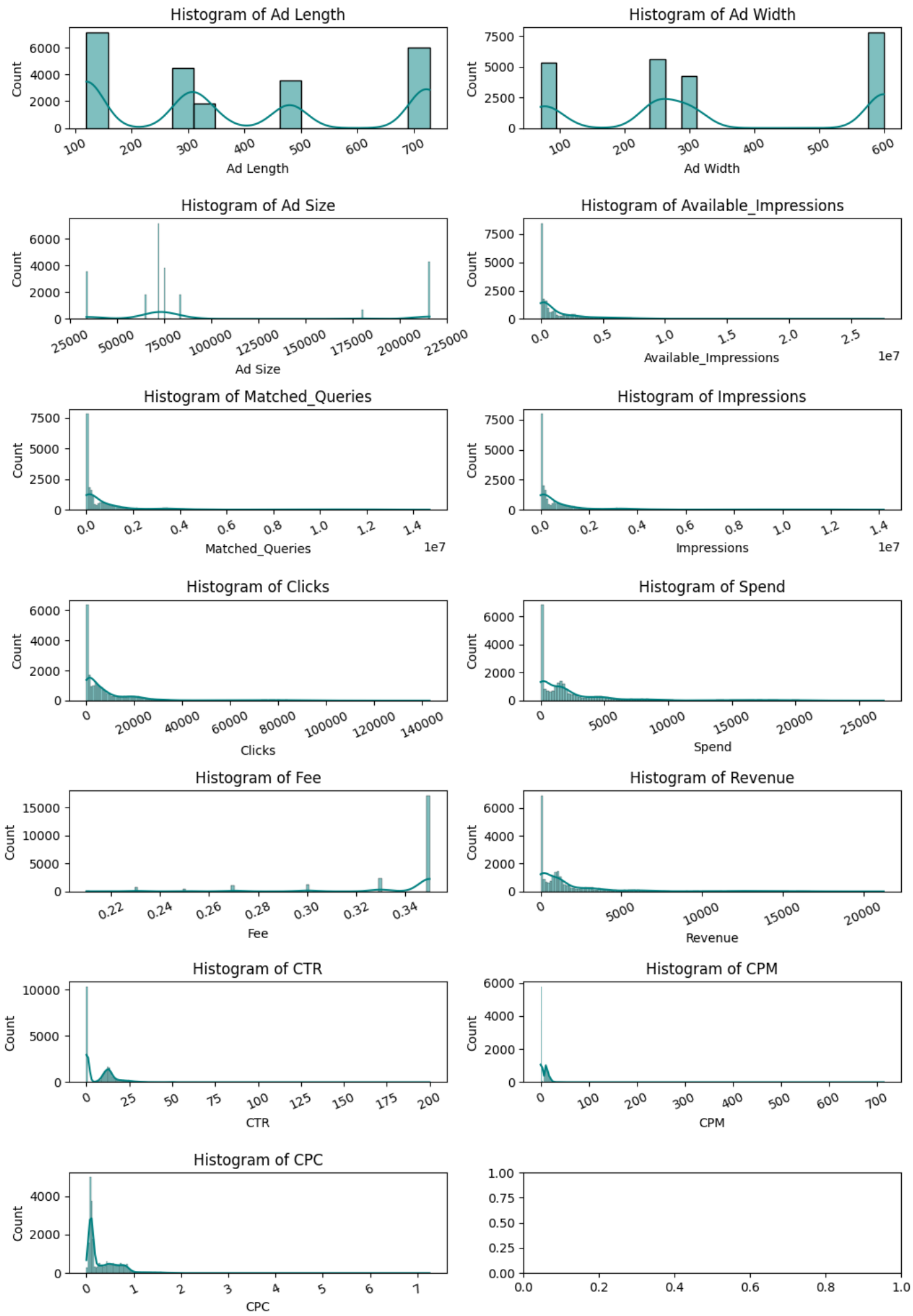


Figure 2 Histogram of Numerical Columns

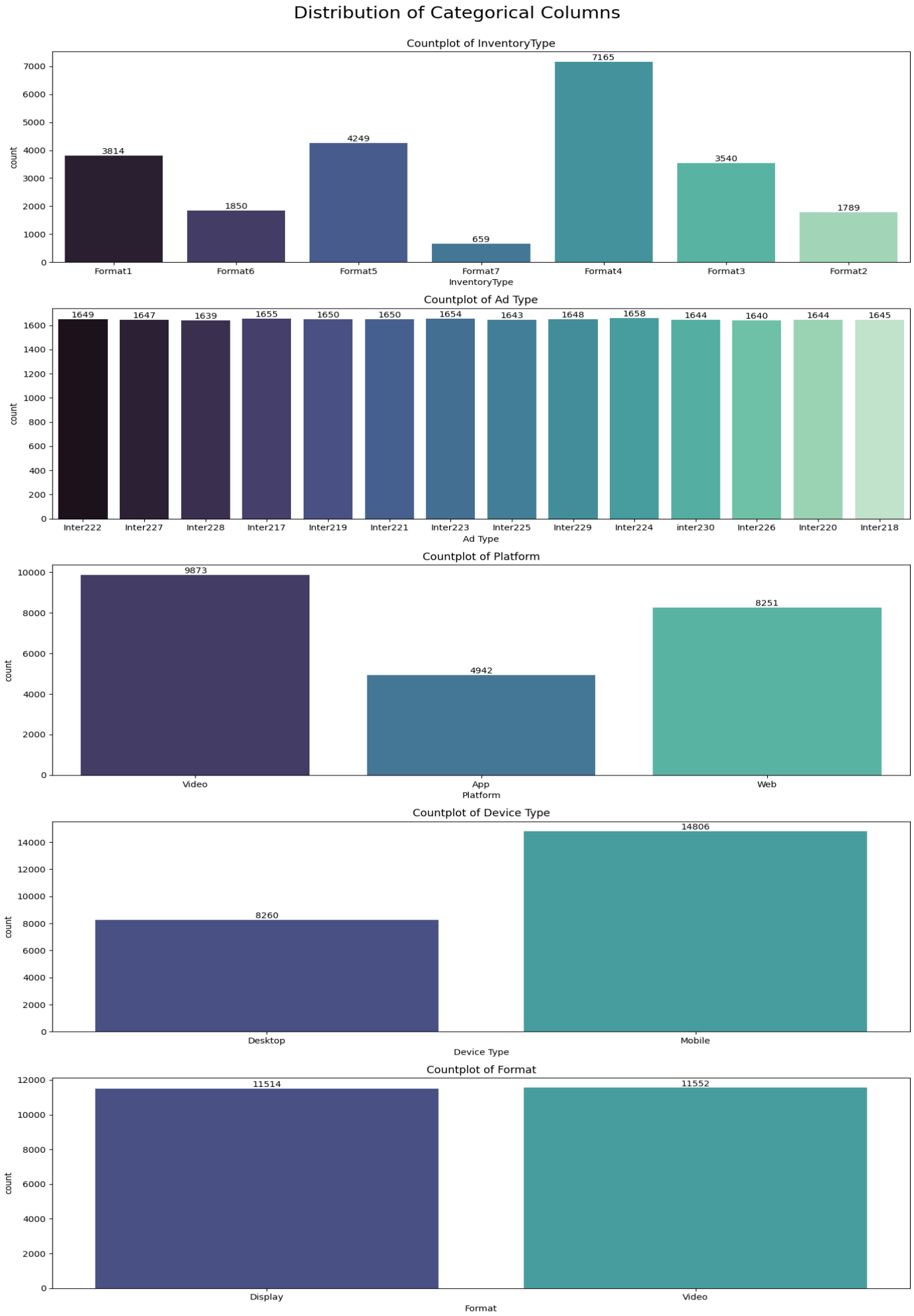


Figure 3 Count Plot of Categorical Columns

- Once the univariate analysis is done, we then proceeded with the bivariate analysis. This step involves examining how pairs of variables interact with each other to uncover relationships within the data.
- First, we plotted a heatmap showing the correlation between the variables. The heatmap provides a visual representation of the strength and direction of relationships between variables, aiding in identifying potential patterns or dependencies.
- Then, we further explored the relationships by plotting box plots of each categorical variable with all the numerical variables. This technique allows us to understand how the distribution of numerical variables varies across different categories, providing deeper insights into potential associations within the data.

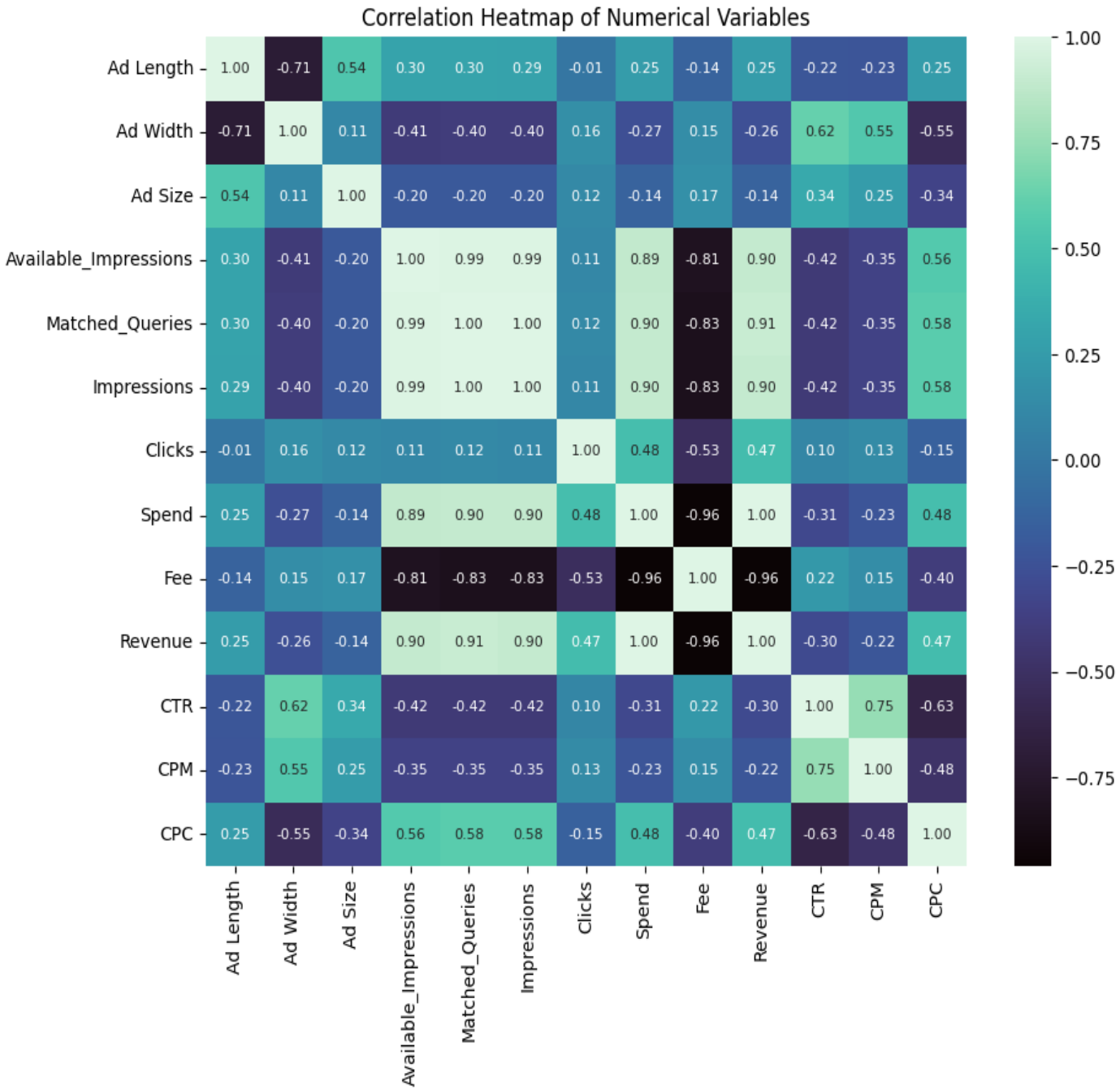


Figure 4 Correlation Heatmap of Numerical Variables

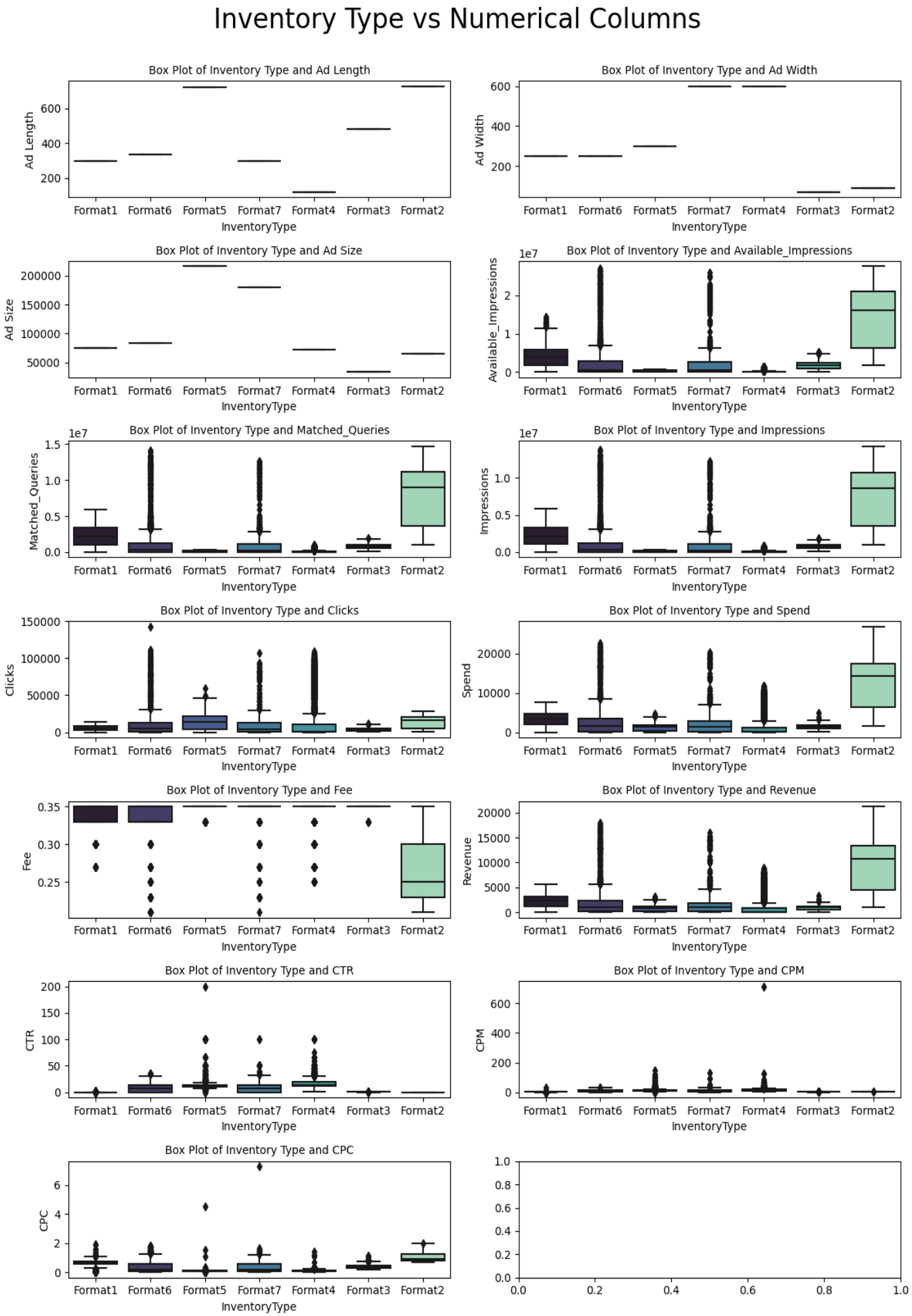


Figure 5 Boxplots of Inventory Type vs Numerical Columns

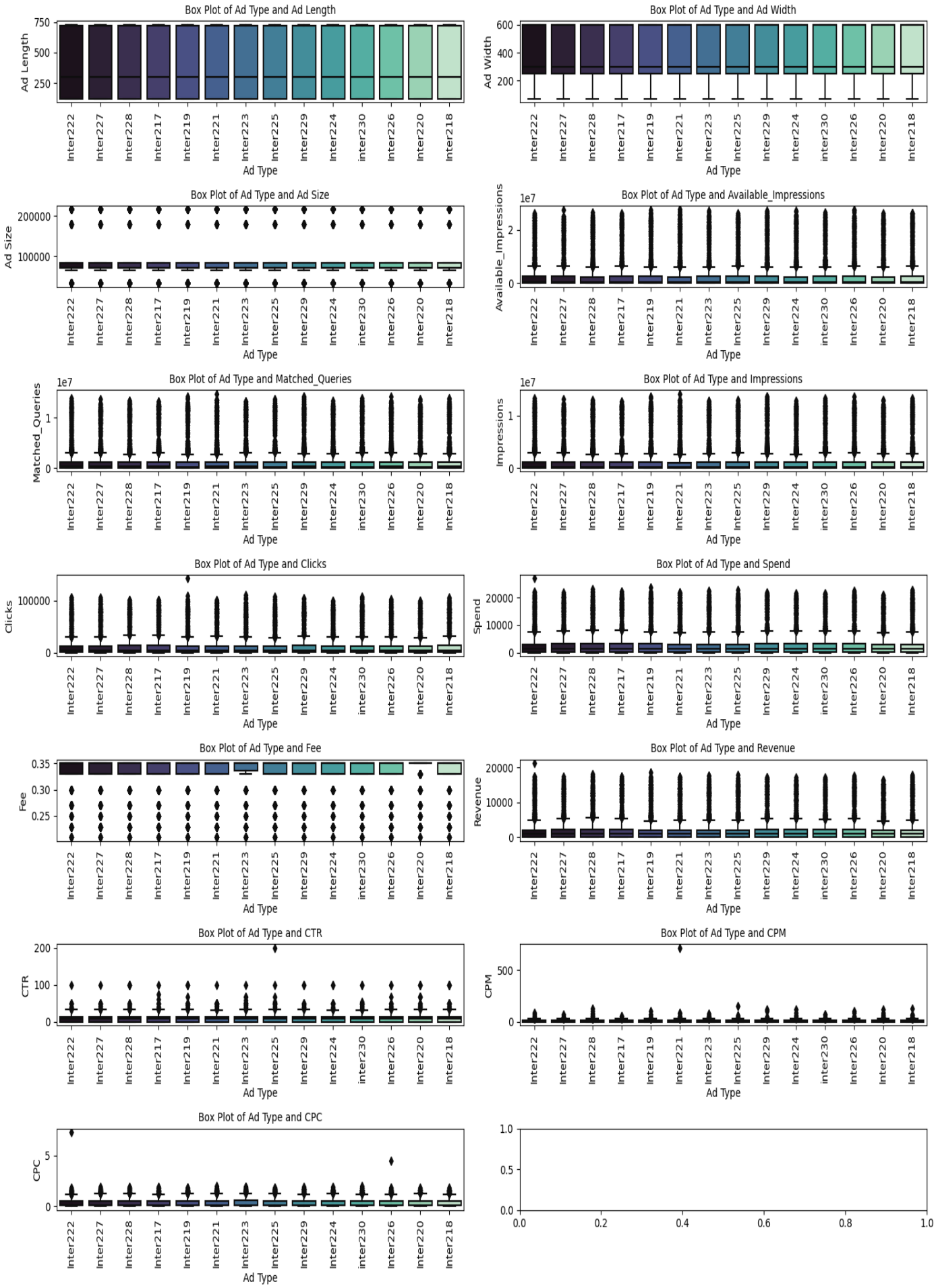


Figure 6 Boxplots of Ad Type vs Numerical Columns

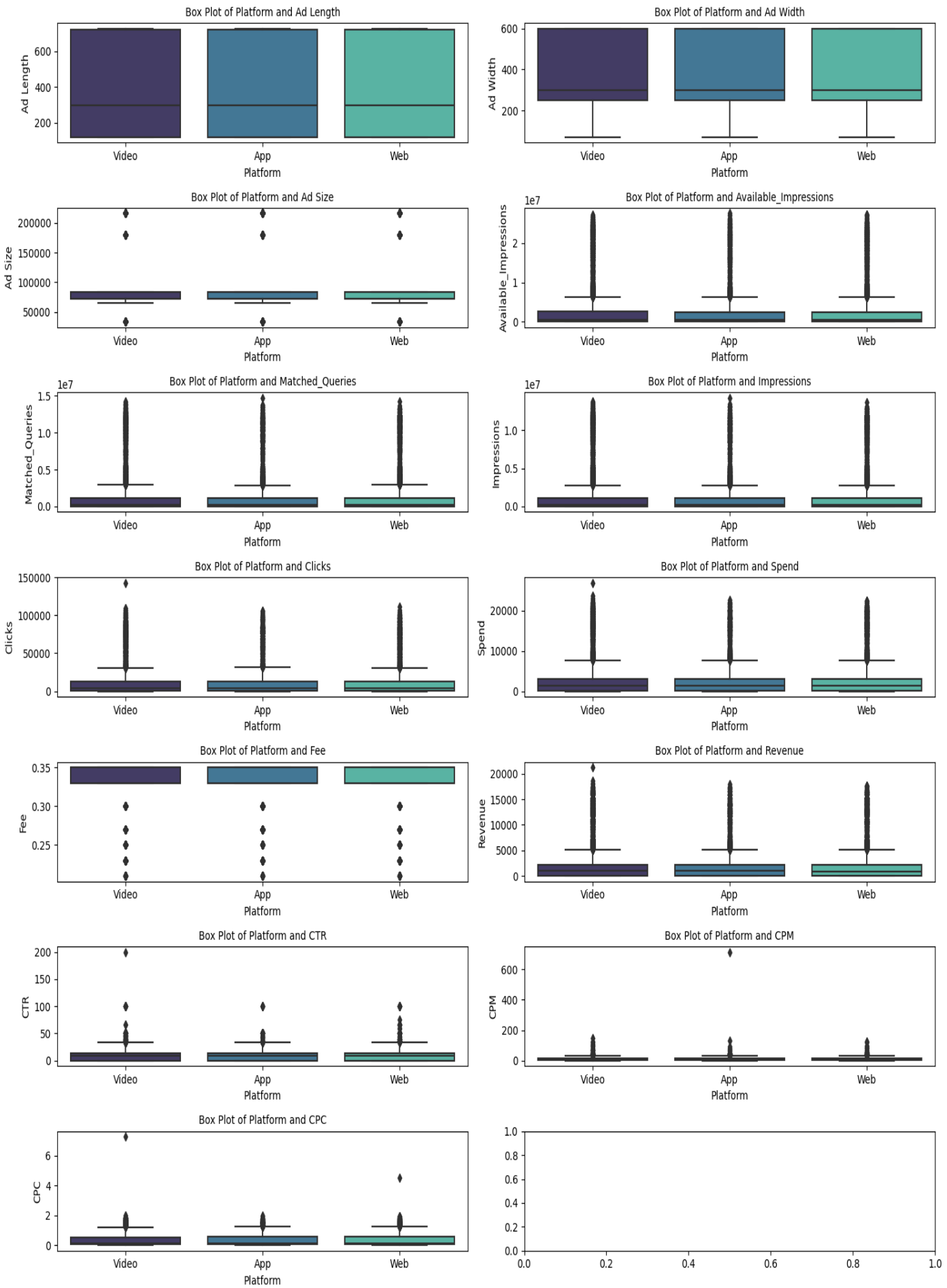


Figure 7 Boxplots of Platform Type vs Numerical Columns

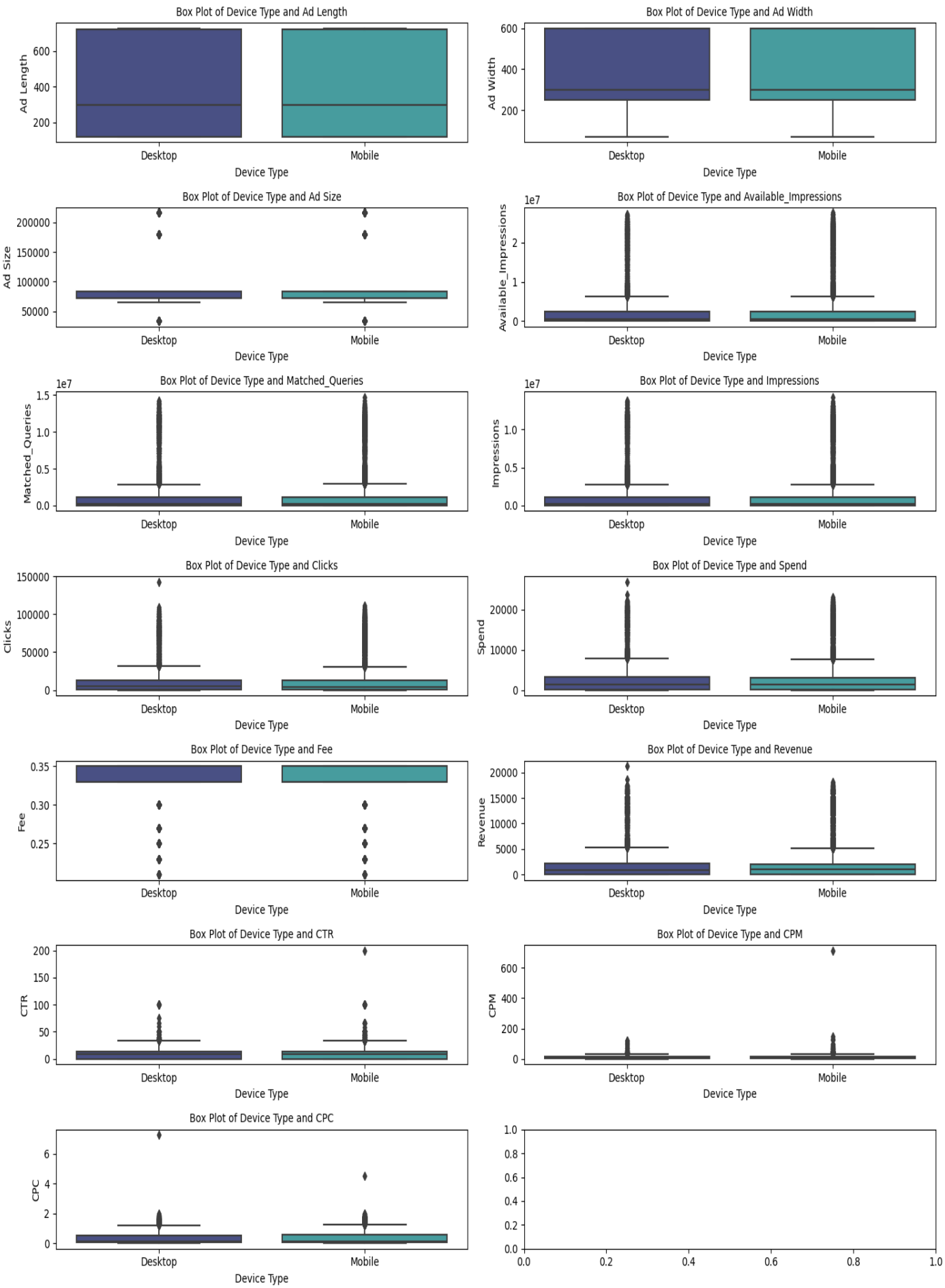


Figure 8 Boxplots of Device Type vs Numerical Columns

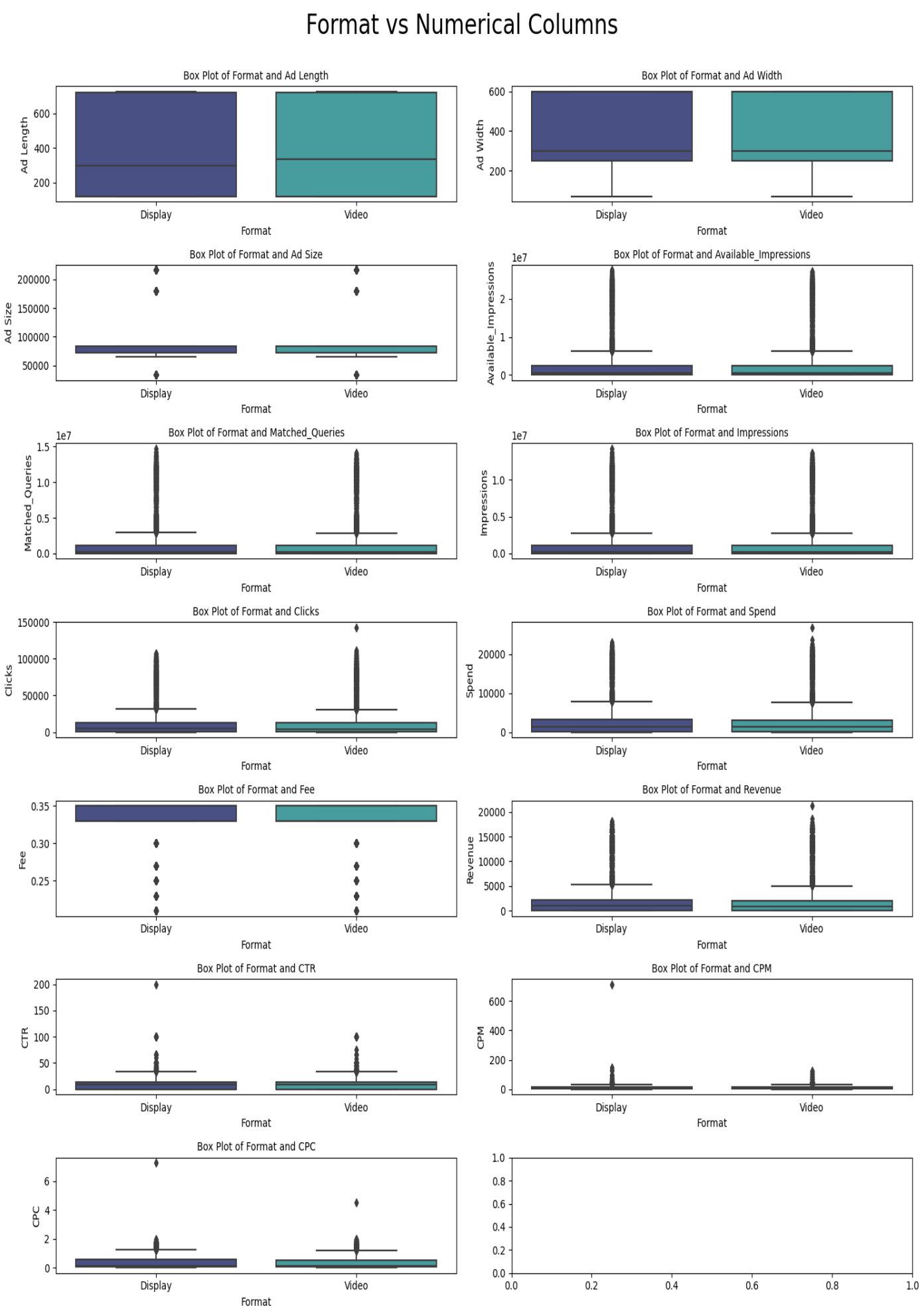


Figure 9 Boxplots of Format vs Numerical Columns

Observations:

- On our analysis, we can see that there are outliers in all the numerical columns except Ad length and Ad width.
- Available Impressions, Matched Queries, Impressions, Clicks, Spend, Revenue are highly correlated. This means that these variables tend to move together. For example, if there are more Available Impressions, there are also likely to be more Matched Queries, Impressions, Clicks, Spend, and Revenue.
- Ad Width and Clicks have a weak positive correlation. This means that there is a slight tendency for ads that are wider to have more clicks.
- Ad Length and Clicks have a very weak negative correlation. This means that there is a very slight tendency for longer ads to have fewer clicks.
- Fee and Revenue are highly correlated. This means that there is a strong tendency for higher fees to be associated with higher revenue.
- CTR, CPM, and CPC are moderately correlated. This means that there is a moderate tendency for campaigns with a higher CTR to also have a higher CPM and CPC.

Problem 1 - Data Preprocessing

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

Solution:

- After treating missing values using the provided formula, our dataset is now devoid of any missing values, ensuring the integrity of our data for further analysis.
- Acknowledging the sensitivity of distance-based techniques to outliers, we prioritize outlier treatment before model training. By identifying and addressing outliers, we aim to enhance the robustness and effectiveness of our model.
- In preparation for outlier treatment, we streamline our dataset by dropping unnecessary columns. This includes removing object variables as they may not contribute to our clustering analysis.
- To systematically handle outliers, we have devised a function called **handle_outliers**. This function calculates upper and lower bound values using a predetermined equation. These bounds serve as thresholds beyond which data points are considered outliers and are subsequently treated to mitigate their impact on our analysis.

$$Upper\ Bound = Q3 + 1.5*IQR$$
$$Lower\ Bound = Q1 - 1.5*IQR$$

Equation 3 Upper and Lower Bound Calculation

- We can see the below Box plots of numerical variables before and after the treatment of outliers.
- We can clearly see that there are no outliers after the treatment.

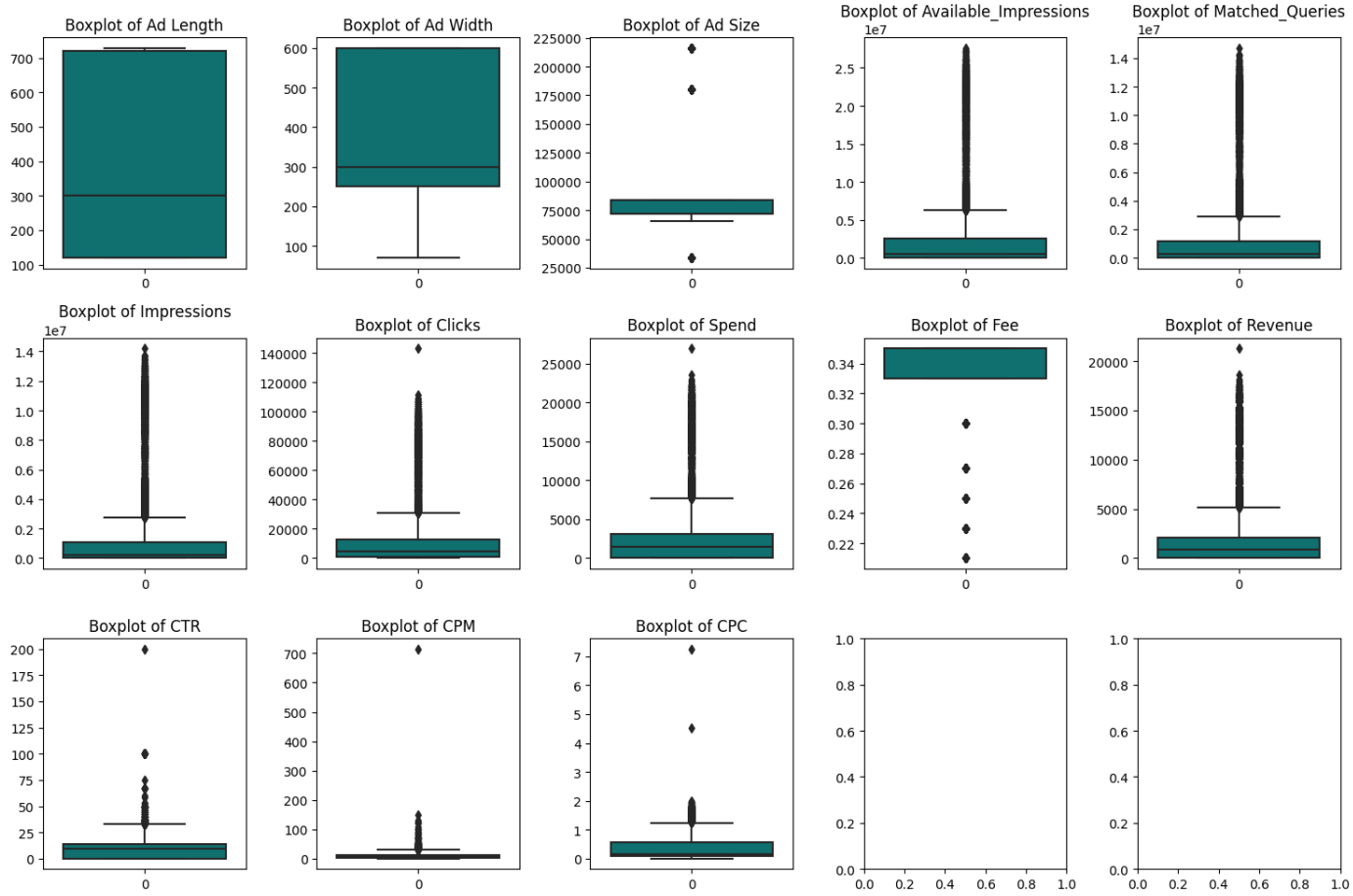


Figure 10 Box Plot of Numerical Variables Before Outlier Treatment

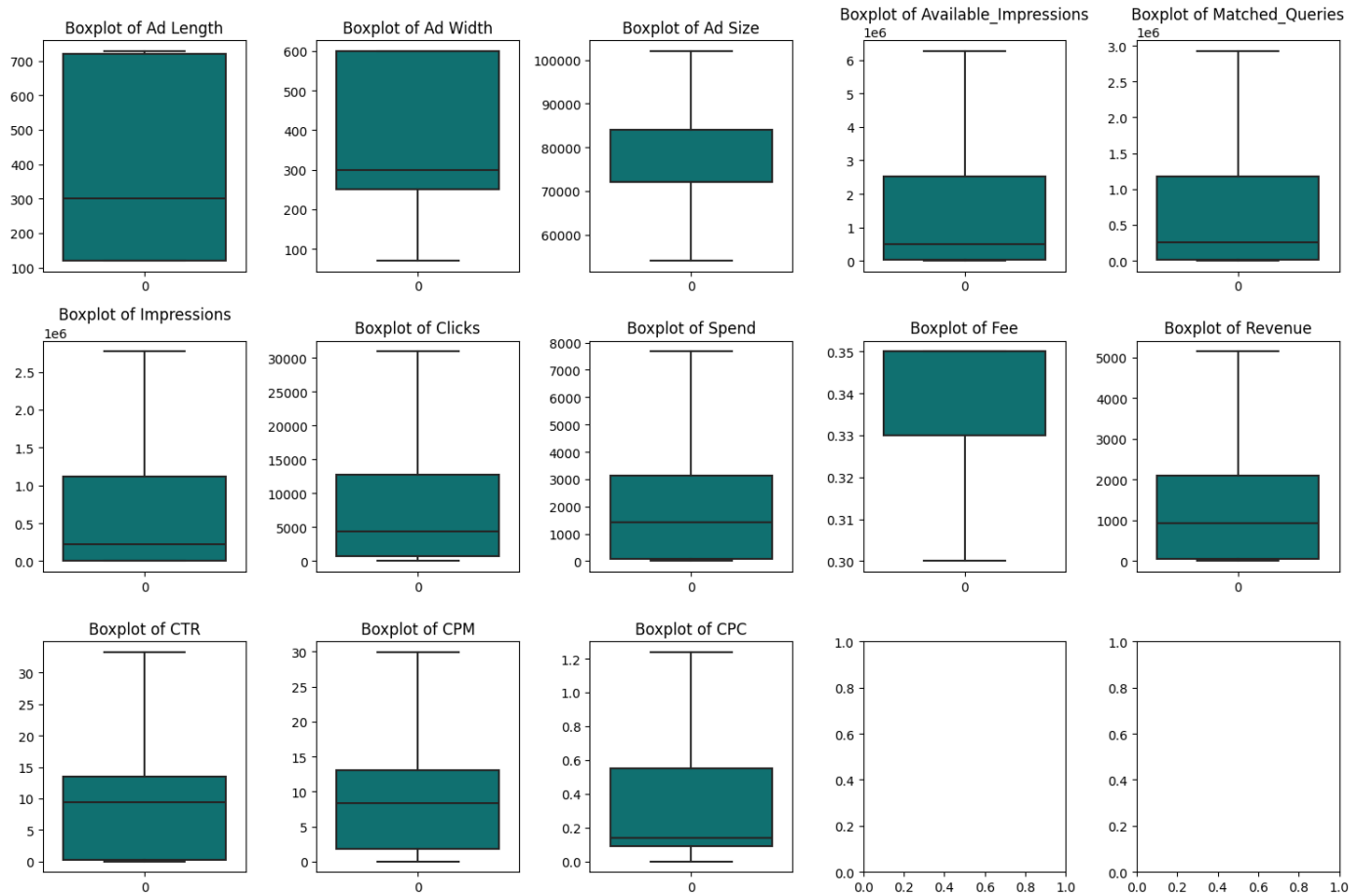


Figure 11 Box Plot of Numerical Variables After Outlier Treatment

- Once outlier treatment was done, We proceeded with scaling the data as scaling is necessary for distance based algorithms. We used StandardScaler() from scikit learn library to scale the data.

Ad Length	Ad Width	Ad Size	Available Impressions	Matched Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
-0.36	-0.43	-0.10	-0.76	-0.78	-0.77	-0.87	-0.89	0.54	-0.88	-0.96	-1.19	-1.04
-0.36	-0.43	-0.10	-0.76	-0.78	-0.77	-0.87	-0.89	0.54	-0.88	-0.95	-1.19	-1.04
-0.36	-0.43	-0.10	-0.75	-0.78	-0.77	-0.87	-0.89	0.54	-0.88	-0.96	-1.19	-1.04
-0.36	-0.43	-0.10	-0.76	-0.78	-0.77	-0.87	-0.89	0.54	-0.88	-0.97	-1.19	-1.04
-0.36	-0.43	-0.10	-0.76	-0.78	-0.77	-0.87	-0.89	0.54	-0.88	-0.95	-1.19	-1.04

Table 5 Sample Data after Scaling

- Z-score scaling ensures equal contribution of all features to distance calculations in both K-means and hierarchical clustering algorithms.
- Without Scaling data, the algorithm may be biased towards higher value.
- It facilitates faster convergence by stabilizing the optimization process, leading to reduced computational time.
- Clustering algorithms become more robust to variations in feature scales across datasets, promoting scalability and generalizability.

Problem 1 - Hierarchical Clustering

- Construct a dendrogram using Ward linkage and Euclidean distance - Identify the optimum number of Clusters

Solution:

Please find below Dendrogram performed for Hierarchical using WARD and Euclidean Distance on the Scaled Data such as “df1_scaled”. In this Dendrogram, value of P = 80, which means that only the last 80 merged clusters are shown.

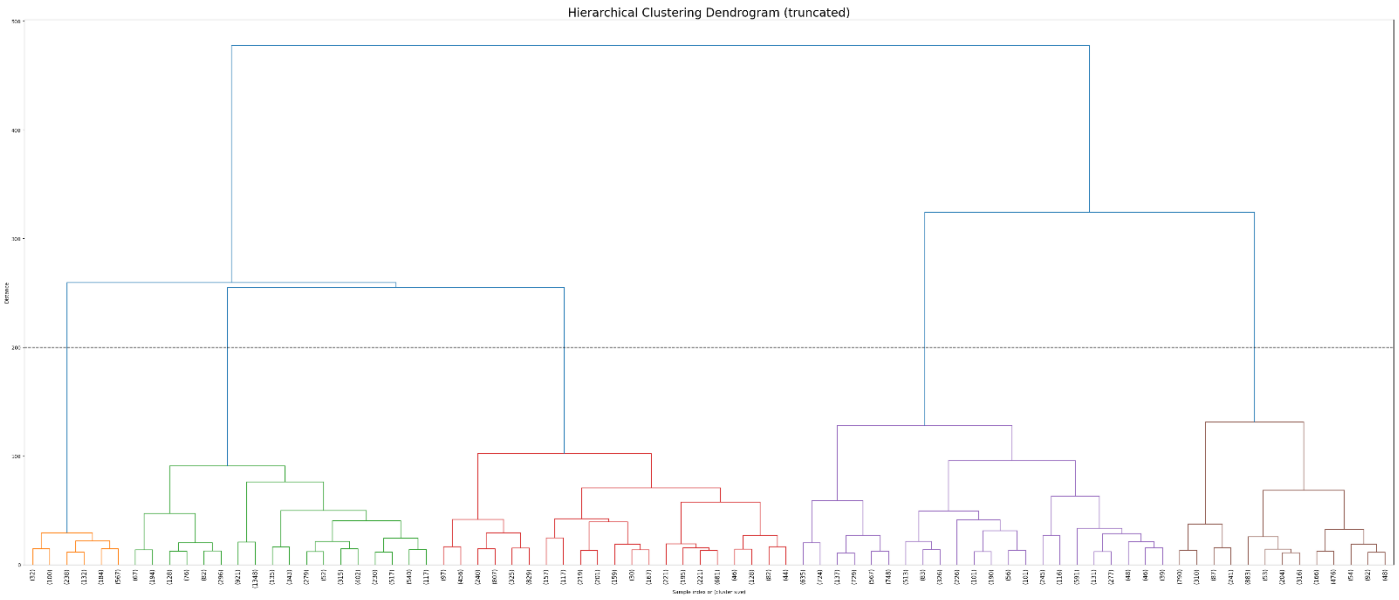


Figure 12 Dendrogram for P=80

- For better readability & visualization we have added the dendrogram with p value equals 20.

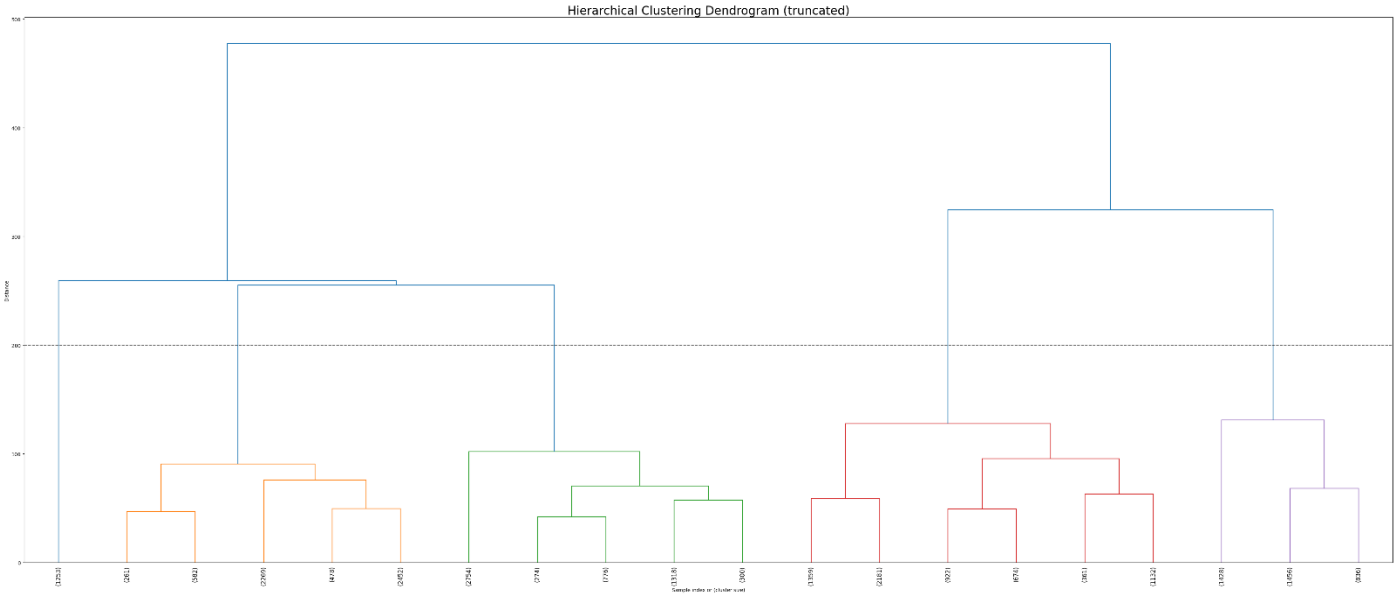


Figure 13 Dendrogram for P=20

- Based on our understanding and analysis of the data, we chose the optimum number of clusters as 5.

Problem 1 - K-means Clustering

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

Solution:

- We have plotted Elbow plot for k values upto 10 to identify the optimum number of clusters.

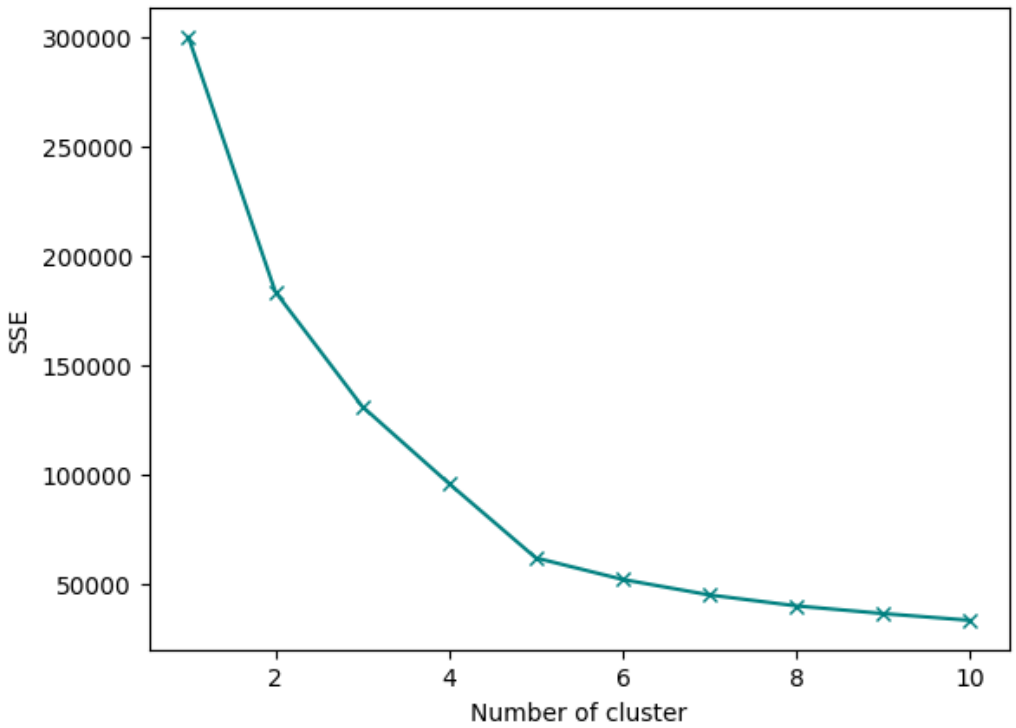


Figure 14 Elbow Plot

- Then we proceeded with calculating the silhouette score for upto 10 cluster to find the optimum number of clusters

No. of Clusters	Silhouette Score
2	0.39
3	0.38
4	0.45
5	0.52
6	0.52
7	0.52
8	0.48
9	0.44
10	0.44

Table 6 Silhouette Score

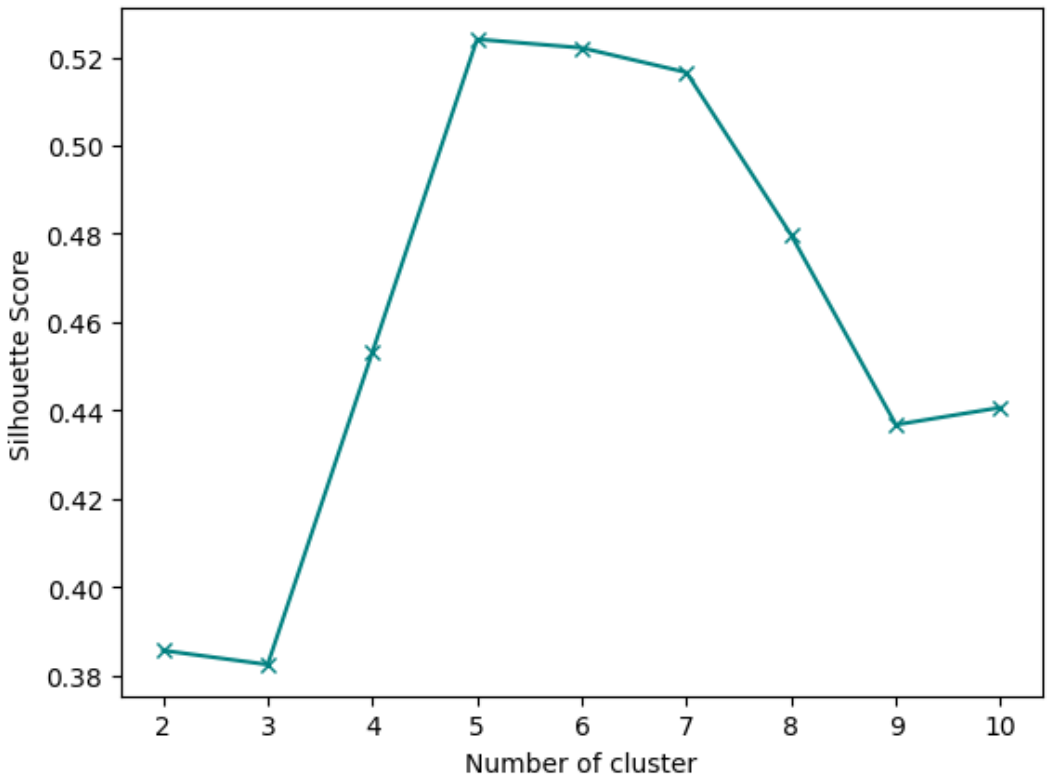


Figure 15 Silhouette Score Plot

- From the Elbow plot Figure 14 and Silhouette Score plot Figure 15, we have decided that the optimum number of cluster is 5.
- Once the number of clusters was decided, we then applied the K-Means algorithm with k number of clusters as 5.

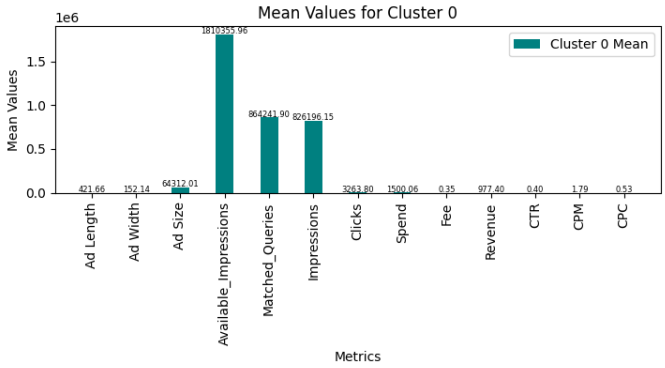
Cluster	Count of Rows
2	6524
0	6277
4	4674
1	4054
3	1537

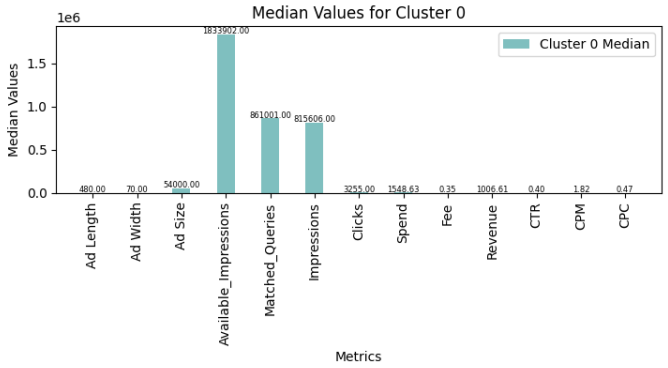
Table 7 K-Means Cluster Data

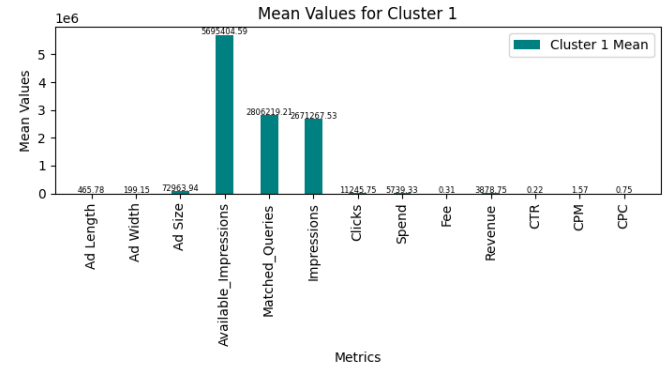
- Finally we grouped the data by clusters and calculated the mean and median values.

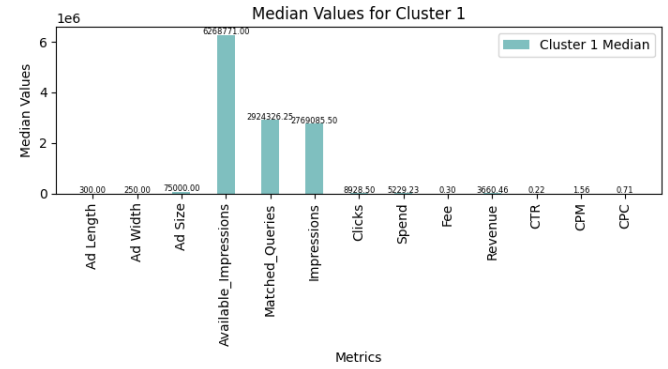
Columns	Cluster_0 Mean	Cluster_ 1 Mean	Cluster_ 2 Mean	Cluster_3 Mean	Cluster_4 Mean	Cluster_ 0 Median	Cluster_1 Median	Cluster_ 2 Median	Cluster_ 3 Median	Cluster_ 4 Median
Ad Length	421.66	465.78	143.28	141.45	683.99	480	300	120	120	720
Ad Width	152.14	199.15	572.1	572.45	303.66	70	250	600	600	300
Ad Size	64312.01	72963.94	73966.74	73686.4	100775.35	54000	75000	72000	72000	102000
Available Impressions	1810355.96	5695405	32093.56	806328.42	250623.18	1833902	6268771	13920.5	831024	213533.5
Matched Queries	864241.9	2806219	19624.06	566864.05	137267.4	861001	2924326.3	8263.5	583232	137245.5
Impressions	826196.15	2671268	13492.04	478148.52	116501.06	815606	2769085.5	3631.5	490310	115808.5
Clicks	3263.8	11245.75	1914.45	30572.44	14131.02	3255	8928.5	454	30919.38	14729
Spend	1500.06	5739.33	209.16	6546.37	1252.22	1548.63	5229.23	47.46	7172.6	1336.86
Fee	0.35	0.31	0.35	0.31	0.35	0.35	0.3	0.35	0.3	0.35
Revenue	977.4	3878.75	135.99	4471.78	815.5	1006.61	3660.46	30.85	5145.3	868.96
CTR	0.4	0.22	15.78	13.75	13.29	0.4	0.22	13.82	13.58	12.47
CPM	1.79	1.57	14.33	15.39	11.73	1.82	1.56	13.38	14.83	11.09
CPC	0.53	0.75	0.1	0.11	0.09	0.47	0.71	0.09	0.11	0.09

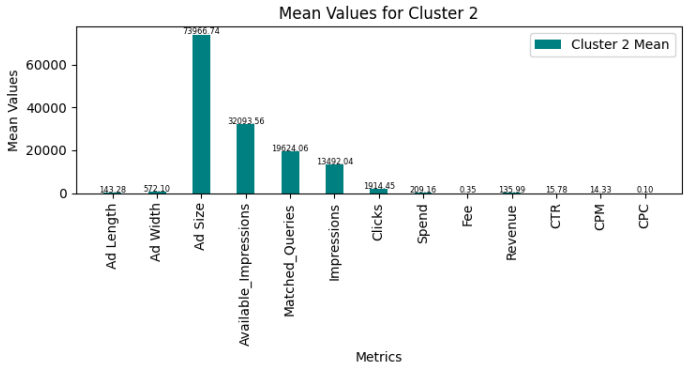
Table 8 Cluster profiling

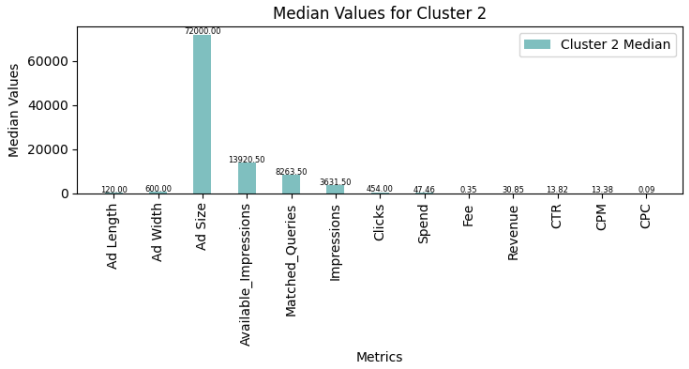












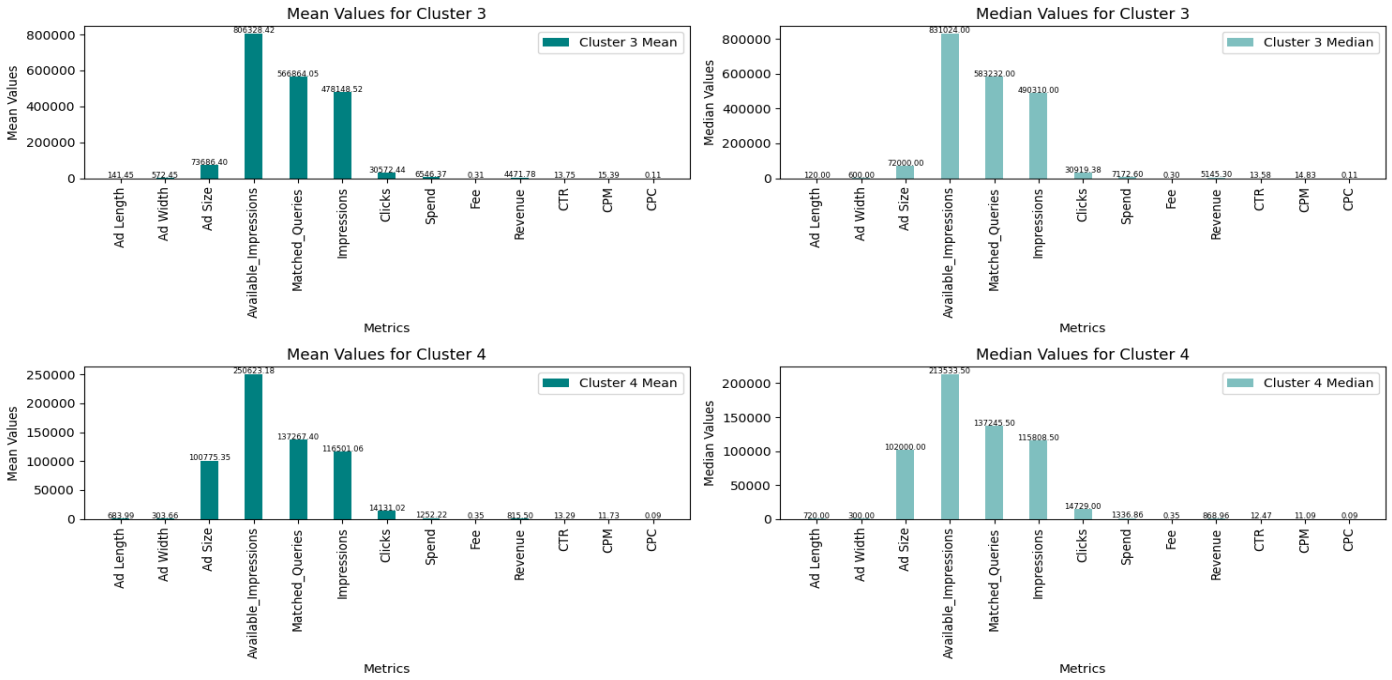


Figure 16 Bar Plots showing Mean & Median of Clusters

Problem 1 - Actionable Insights & Recommendations

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

Solution:

- There are 23066 rows, and 19 columns into the Dataset.
- There are no duplicate values in dataframe.
- There are 4636 Null values in CTR, CPM, and CPC Columns.
- I have treated missing values in CPC, CTR, and CPM columns using the given formula
- It seems that there are Outliers into the Dataset
- We treated outliers using upper and lower bound values
- I have applied z-score method on the dataframe for scaling
- I have plotted Dendrogram and got optimum number of clusters is 5
- Plotted elbow plot and got optimum value is 5
- As per Silhouette score, we concluded that the optimal number of clusters should be 5.
- I have created 5 clusters for the Dataset

Recommendations:

- When Click on Ads gets increases then Revenue is also increases.
- When amount of money spent on specific ad variations within a specific campaign or ad set increases then Revenue is also increases.
- When impression count of the particular Advertisement increases then Revenue is also increases

Problem 2

Context

PCA:

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

- Note: The 24 variables given in the Rubric is just for performing EDA. You will have to consider the entire dataset, including all the variables for performing PCA.

Objective

The objective is to conduct exploratory data analysis (EDA) on the Primary Census Abstract dataset for female-headed households in India, excluding institutional households. Utilize Sklearn to perform detailed analysis on the dataset comprising numerous variables. Identify and extract optimum principal components that explain the maximum variance within the data. The objective is to gain insights into the underlying patterns and structures of the dataset, facilitating a deeper understanding of demographic characteristics and societal trends captured by the Indian Census of 2011.

Problem 2 - Define the problem and perform Exploratory Data Analysis

- Problem Definition - Check shape, Data types, statistical summary - Perform an EDA on the data to extract useful insights Note: 1. Pick 5 variables out of the given 24 variables below for EDA: No_HH, TOT_M, TOT_F, M_06, F_06, M_SC, F_SC, M_ST, F_ST, M_LIT, F_LIT, M_ILL, F_ILL, TOT_WORK_M, TOT_WORK_F, MAINWORK_M, MAINWORK_F, MAIN_CL_M, MAIN_CL_F, MAIN_AL_M, MAIN_AL_F, MAIN_HH_M, MAIN_HH_F, MAIN_OT_M, MAIN_OT_F 2. Example questions to answer from EDA - (i) Which state has highest gender ratio and which has the lowest? (ii) Which district has the highest & lowest gender ratio?

Solution:

Analyze the Primary Census Abstract dataset for female-headed households in India, excluding institutional households, using Sklearn. Explore demographic trends and socio-economic factors through exploratory data analysis (EDA) to identify optimal principal components explaining maximum variance. This aims to provide insights for informed decision-making and policy formulation regarding household composition and resource allocation.

- We used shape attribute of pandas library [df2.shape] to determine the number of rows and columns of the dataframe,

Shape	Description
640, 61	We have 640 rows and 61 columns in the dataset

Table 9 Shape of Population Dataset

- To check the types of data, we used info() method of pandas library to list the basic information of the data such as rows count, columns count and datatype of the columns.

Column	Non-Null Count	Dtype
State Code	640 non-null	int64
Dist.Code	640 non-null	int64
State	640 non-null	object
Area Name	640 non-null	object
No_HH	640 non-null	int64
TOT_M	640 non-null	int64
TOT_F	640 non-null	int64
M_06	640 non-null	int64
F_06	640 non-null	int64
M_SC	640 non-null	int64
F_SC	640 non-null	int64
M_ST	640 non-null	int64
F_ST	640 non-null	int64
M_LIT	640 non-null	int64
F_LIT	640 non-null	int64
M_ILL	640 non-null	int64
F_ILL	640 non-null	int64
TOT_WORK_M	640 non-null	int64
TOT_WORK_F	640 non-null	int64
MAINWORK_M	640 non-null	int64
MAINWORK_F	640 non-null	int64
MAIN_CL_M	640 non-null	int64
MAIN_CL_F	640 non-null	int64
MAIN_AL_M	640 non-null	int64
MAIN_AL_F	640 non-null	int64
MAIN_HH_M	640 non-null	int64
MAIN_HH_F	640 non-null	int64
MAIN_OT_M	640 non-null	int64
MAIN_OT_F	640 non-null	int64
MARGWORK_M	640 non-null	int64

MARGWORK_F	640 non-null	int64
MARG_CL_M	640 non-null	int64
MARG_CL_F	640 non-null	int64
MARG_AL_M	640 non-null	int64
MARG_AL_F	640 non-null	int64
MARG_HH_M	640 non-null	int64
MARG_HH_F	640 non-null	int64
MARG_OT_M	640 non-null	int64
MARG_OT_F	640 non-null	int64
MARGWORK_3_6_M	640 non-null	int64
MARGWORK_3_6_F	640 non-null	int64
MARG_CL_3_6_M	640 non-null	int64
MARG_CL_3_6_F	640 non-null	int64
MARG_AL_3_6_M	640 non-null	int64
MARG_AL_3_6_F	640 non-null	int64
MARG_HH_3_6_M	640 non-null	int64
MARG_HH_3_6_F	640 non-null	int64
MARG_OT_3_6_M	640 non-null	int64
MARG_OT_3_6_F	640 non-null	int64
MARGWORK_0_3_M	640 non-null	int64
MARGWORK_0_3_F	640 non-null	int64
MARG_CL_0_3_M	640 non-null	int64
MARG_CL_0_3_F	640 non-null	int64
MARG_AL_0_3_M	640 non-null	int64
MARG_AL_0_3_F	640 non-null	int64
MARG_HH_0_3_M	640 non-null	int64
MARG_HH_0_3_F	640 non-null	int64
MARG_OT_0_3_M	640 non-null	int64
MARG_OT_0_3_F	640 non-null	int64
NON_WORK_M	640 non-null	int64
NON_WORK_F	640 non-null	int64

Table 10 Basic Information of Population Dataset

- From the above table, We have **2** object and **59** int data types in the dataset. Also we can see that there are no null rows in the CTR, CPM and CPC columns.
- To check if there are any duplicate columns, we used df1.duplicated().sum(). We can see that there are no duplicate rows in the dataset.
- Once the above is done, we found the statistical summary of the dataset using describe method()

Index	Count	Mean	STD	Min	25%	50%	75%	Max
State Code	640	17.1140625	9.4264863	1	9	18	24	35
Dist.Code	640	320.5	184.896367	1	160.75	320.5	480.25	640
No_HH	640	51222.8719	48135.4055	350	19484	35837	68892	310450
TOT_M	640	79940.5766	73384.5111	391	30228	58339	107918.5	485417
TOT_F	640	122372.084	113600.717	698	46517.75	87724.5	164251.75	750392
M_06	640	12309.0984	11500.9069	56	4733.75	9159	16520.25	96223
F_06	640	11942.3	11326.2946	56	4672.25	8663	15902.25	95129

M_SC	640	13820.9469	14426.3731	0	3466.25	9591.5	19429.75	103307
F_SC	640	20778.3922	21727.8877	0	5603.25	13709	29180	156429
M_ST	640	6191.80781	9912.66895	0	293.75	2333.5	7658	96785
F_ST	640	10155.6406	15875.7015	0	429.5	3834.5	12480.25	130119
M_LIT	640	57967.9797	55910.2825	286	21298	42693.5	77989.5	403261
F_LIT	640	66359.5656	75037.8602	371	20932	43796.5	84799.75	571140
M_ILL	640	21972.5969	19825.6053	105	8590	15767.5	29512.5	105961
F_ILL	640	56012.5188	47116.6938	327	22367	42386	78471	254160
TOT_WORK_M	640	37992.4078	36419.5375	100	13753.5	27936.5	50226.75	269422
TOT_WORK_F	640	41295.7609	37192.3609	357	16097.75	30588.5	53234.25	257848
MAINWORK_M	640	30204.4469	31480.9157	65	9787	21250.5	40119	247911
MAINWORK_F	640	28198.8469	29998.2627	240	9502.25	18484	35063.25	226166
MAIN_CL_M	640	5424.34219	4739.16197	0	2023.5	4160.5	7695	29113
MAIN_CL_F	640	5486.04219	5326.36273	0	1920.25	3908.5	7286.25	36193
MAIN_AL_M	640	5849.10938	6399.50797	0	1070.25	3936.5	8067.25	40843
MAIN_AL_F	640	8925.99531	12864.2876	0	1408.75	3933.5	10617.5	87945
MAIN_HH_M	640	883.89375	1278.64234	0	187.5	498.5	1099.25	16429
MAIN_HH_F	640	1380.77344	3179.41445	0	248.75	540.5	1435.75	45979
MAIN_OT_M	640	18047.1016	26068.4809	36	3997.5	9598	21249.5	240855
MAIN_OT_F	640	12406.0359	18972.2024	153	3142.5	6380.5	14368.25	209355
MARGWORK_M	640	7787.96094	7410.79169	35	2937.5	5627	9800.25	47553
MARGWORK_F	640	13096.9141	10996.4745	117	5424.5	10175	18879.25	66915
MARG_CL_M	640	1040.7375	1311.54685	0	311.75	606.5	1281	13201
MARG_CL_F	640	2307.68281	3564.6261	0	630.25	1226	2659.25	44324
MARG_AL_M	640	3304.32656	3781.55571	0	873.5	2062	4300.75	23719
MARG_AL_F	640	6463.28125	6773.8763	0	1402.5	4020.5	9089.25	45301
MARG_HH_M	640	316.742188	462.661891	0	71.75	166	356.5	4298
MARG_HH_F	640	786.626563	1198.71821	0	171.75	429	962.5	15448
MARG_OT_M	640	3126.15469	3609.39182	7	935.5	2036	3985.25	24728
MARG_OT_F	640	3539.32344	4115.19131	19	1071.75	2349.5	4400.5	36377
MARGWORK_3_6_M	640	41948.1688	39045.3169	291	16208.25	30315	57218.75	300937
MARGWORK_3_6_F	640	81076.3234	82970.4062	341	26619.5	56793	107924	676450
MARG_CL_3_6_M	640	6394.9875	6019.80664	27	2372	4630	8167	39106
MARG_CL_3_6_F	640	10339.8641	8467.47343	85	4351.5	8295	15102	50065
MARG_AL_3_6_M	640	789.848438	905.639279	0	235.5	480.5	986	7426
MARG_AL_3_6_F	640	1749.58438	2496.54151	0	497.25	985.5	2059	27171
MARG_HH_3_6_M	640	2743.63594	3059.58639	0	718.75	1714.5	3702.25	19343
MARG_HH_3_6_F	640	5169.85	5335.64096	0	1113.75	3294	7502.25	36253
MARG_OT_3_6_M	640	245.3625	358.728567	0	58	129.5	276	3535
MARG_OT_3_6_F	640	585.884375	900.025817	0	127.75	320.5	719.25	12094
MARGWORK_0_3_M	640	2616.14063	3036.96438	7	755	1681.5	3320.25	20648
MARGWORK_0_3_F	640	2834.54531	3327.83693	14	833.5	1834.5	3610.5	25844
MARG_CL_0_3_M	640	1392.97344	1489.70705	4	489.5	949	1714	9875
MARG_CL_0_3_F	640	2757.05	2788.77668	30	957.25	1928	3599.75	21611
MARG_AL_0_3_M	640	250.889063	453.336594	0	47	114.5	270.75	5775
MARG_AL_0_3_F	640	558.098438	1117.64275	0	109	247.5	568.75	17153
MARG_HH_0_3_M	640	560.690625	762.578991	0	136.5	308	642	6116
MARG_HH_0_3_F	640	1293.43125	1585.37794	0	298	717	1710.75	13714
MARG_OT_0_3_M	640	71.3796875	107.897627	0	14	35	79	895
MARG_OT_0_3_F	640	200.742188	309.740854	0	43	113	240	3354
NON_WORK_M	640	510.014063	610.603187	0	161	326	604.5	6456
NON_WORK_F	640	704.778125	910.209225	5	220.5	464.5	853.5	10533

Table 11 Statistical Summary of the Population Data

- Once the basic check was done, we then proceeded with EDA. We have so many attributes in the data so, we chose TOT_M, TOT_F, M_LIT, F_LIT, M_ILL, F_ILL attributes which enables us to perform better analysis.

Column	Count	Unique	Top	Freq	Mean	STD	Min	25%	50%	75%	Max
State	640	35	Uttar Pradesh	71	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Area Name	640	635	Raigarh	2	NaN	NaN	NaN	NaN	NaN	NaN	NaN
TOT_M	640	NaN	NaN	NaN	79940.57656	73384.51111	391	30228	58339	107918.5	485417
TOT_F	640	NaN	NaN	NaN	122372.0844	113600.7173	698	46517.75	87724.5	164251.75	750392
M_LIT	640	NaN	NaN	NaN	57967.97969	55910.28247	286	21298	42693.5	77989.5	403261
F_LIT	640	NaN	NaN	NaN	66359.56563	75037.86021	371	20932	43796.5	84799.75	571140
M_ILL	640	NaN	NaN	NaN	21972.59688	19825.60527	105	8590	15767.5	29512.5	105961
F_ILL	640	NaN	NaN	NaN	56012.51875	47116.69377	327	22367	42386	78471	254160

Table 12 Statistical Summary of the EDA Population Data

- The average male & female population is approximately 79940 and 122372 respectively
- The average literate male & female population is approximately 57967 and 66359 respectively
- The average illiterate male & female population is approximately 21972 and 56012 respectively.
- We plotted the Box Plots & Histogram of Numerical Columns and Count Plot of Categorical Columns

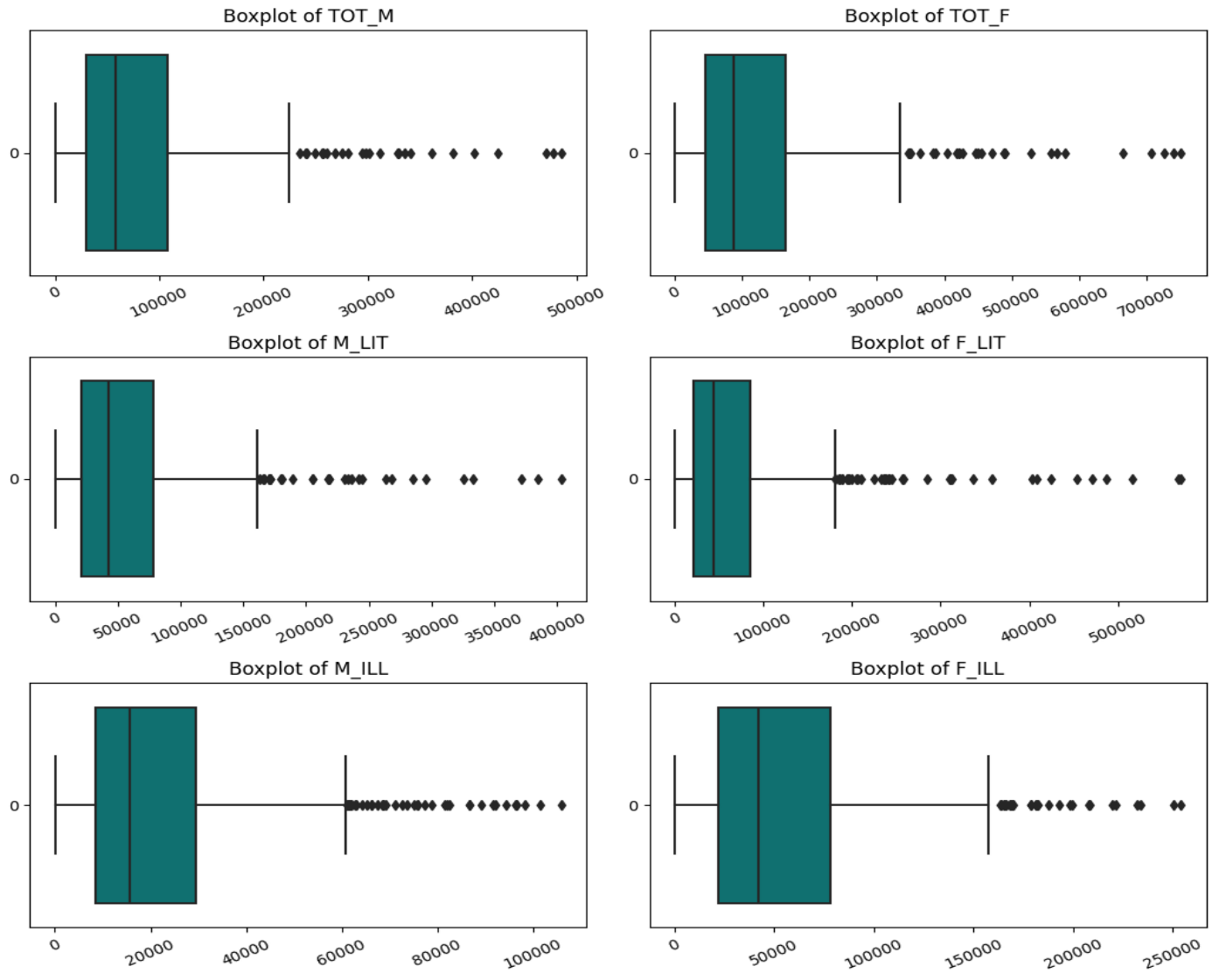


Figure 17 Box Plot EDA Population Data

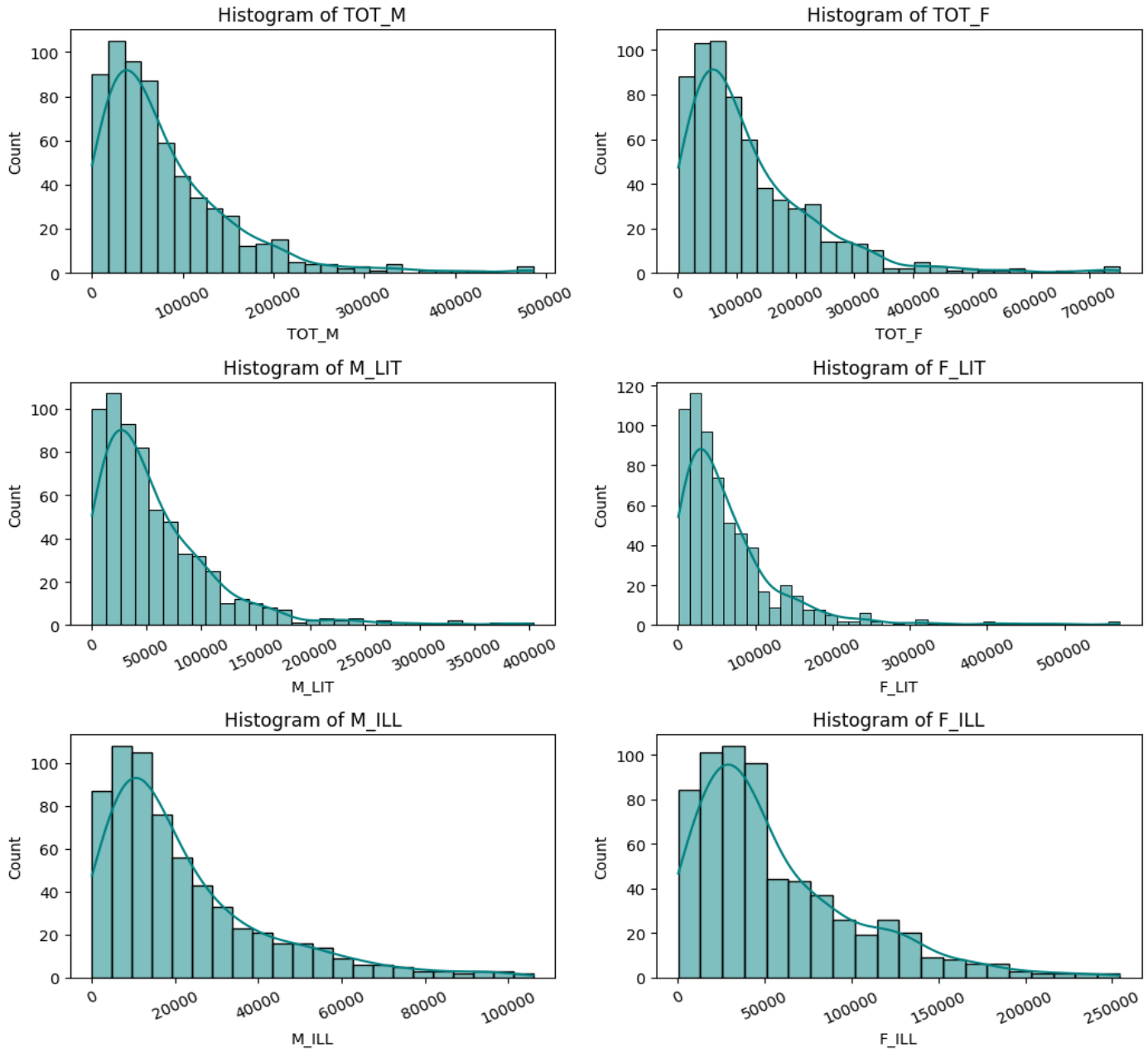


Figure 18 Histogram of EDA Population Data

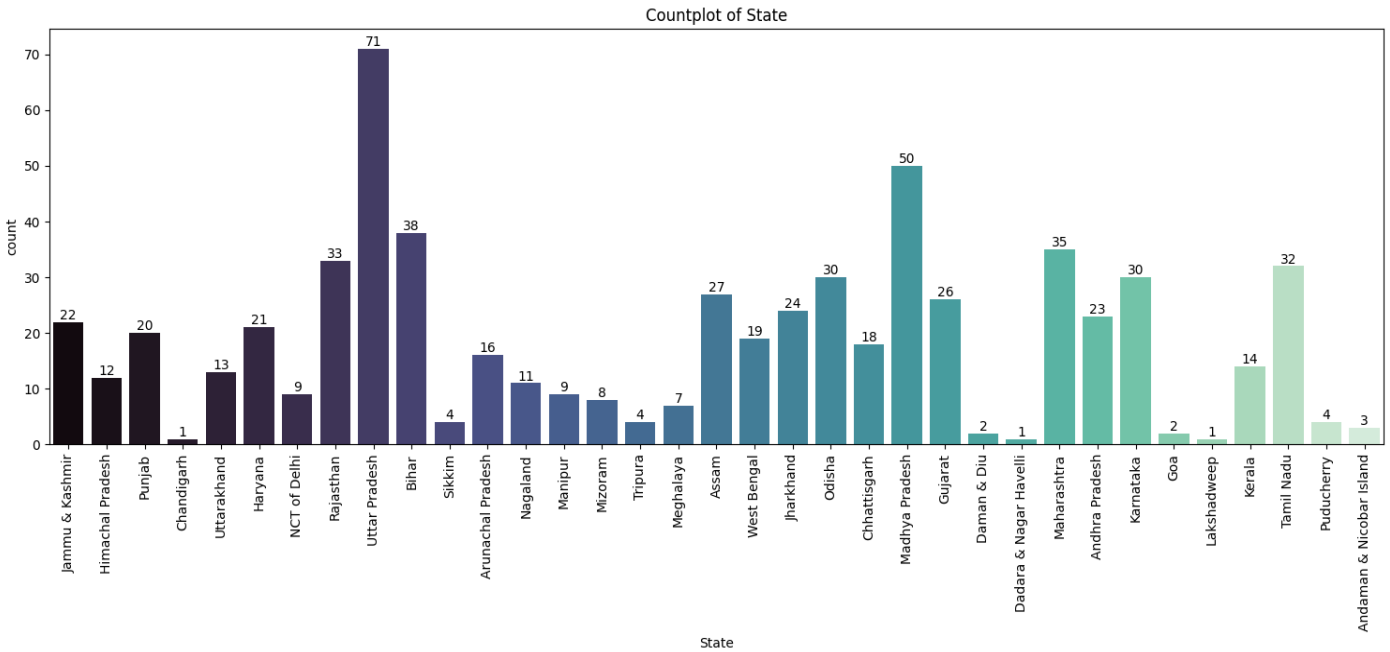


Figure 19 Countplot of States

From the Univariate analysis,

- We can see outliers in the numerical columns from the boxplot
- The selected numerical variables are all Right Skewed
- From the countplot, Uttar Pradesh has high frequency of 71 and Chandigarh, Dadar & Lakshadweep has low frequency of 1
- We then plotted a pairgrid showing the relationship between all the numerical variables.

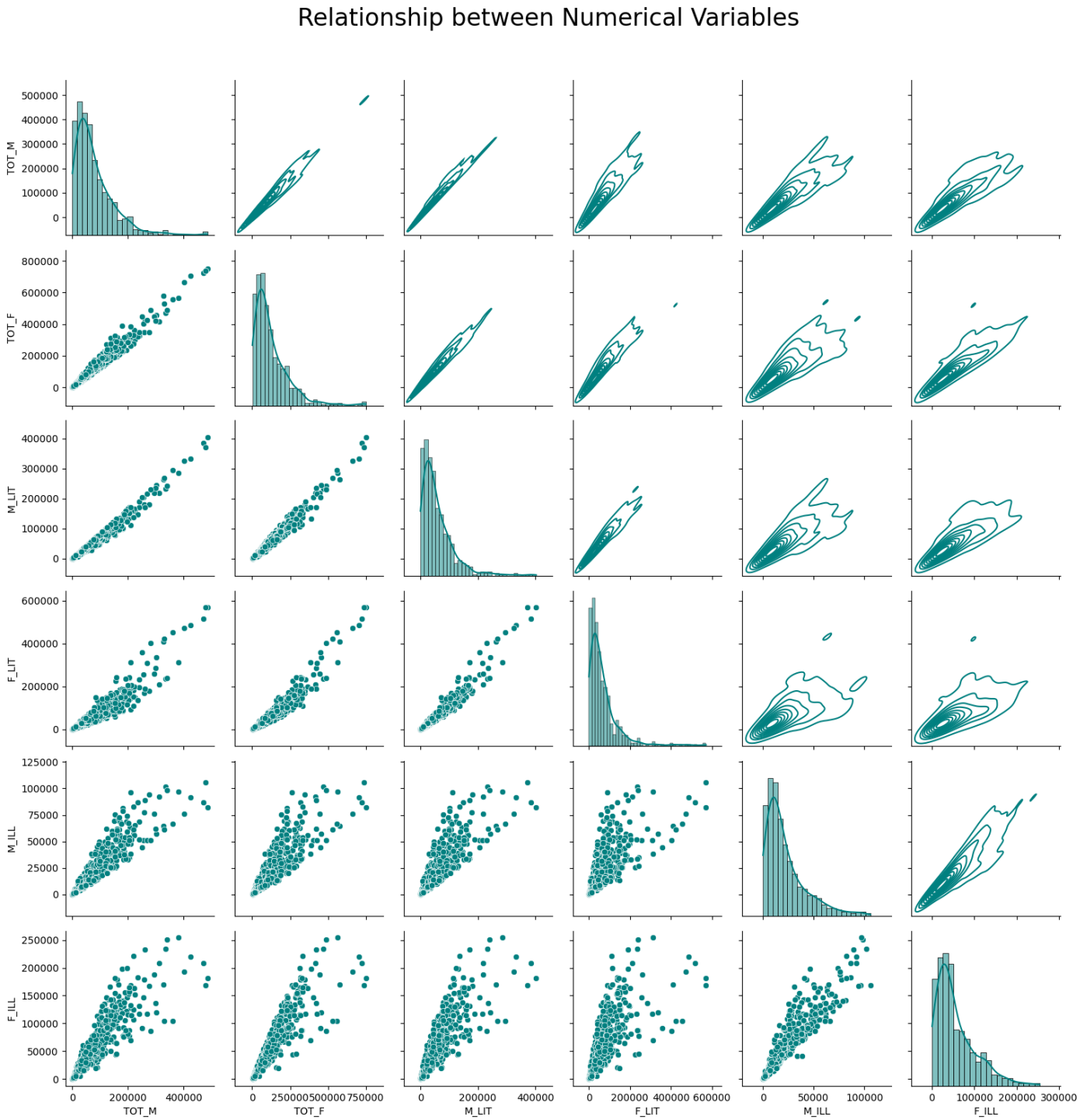


Figure 20 Relationship Between Numerical Variables

- From the above figure Figure 20, We can see that all the variables we chose for performing EDA are positively correlated.

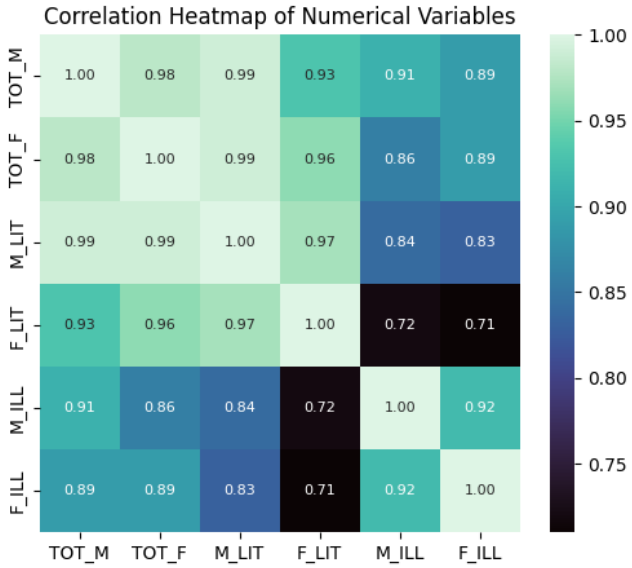


Figure 21 Correlation Heatmap of Numerical Variables

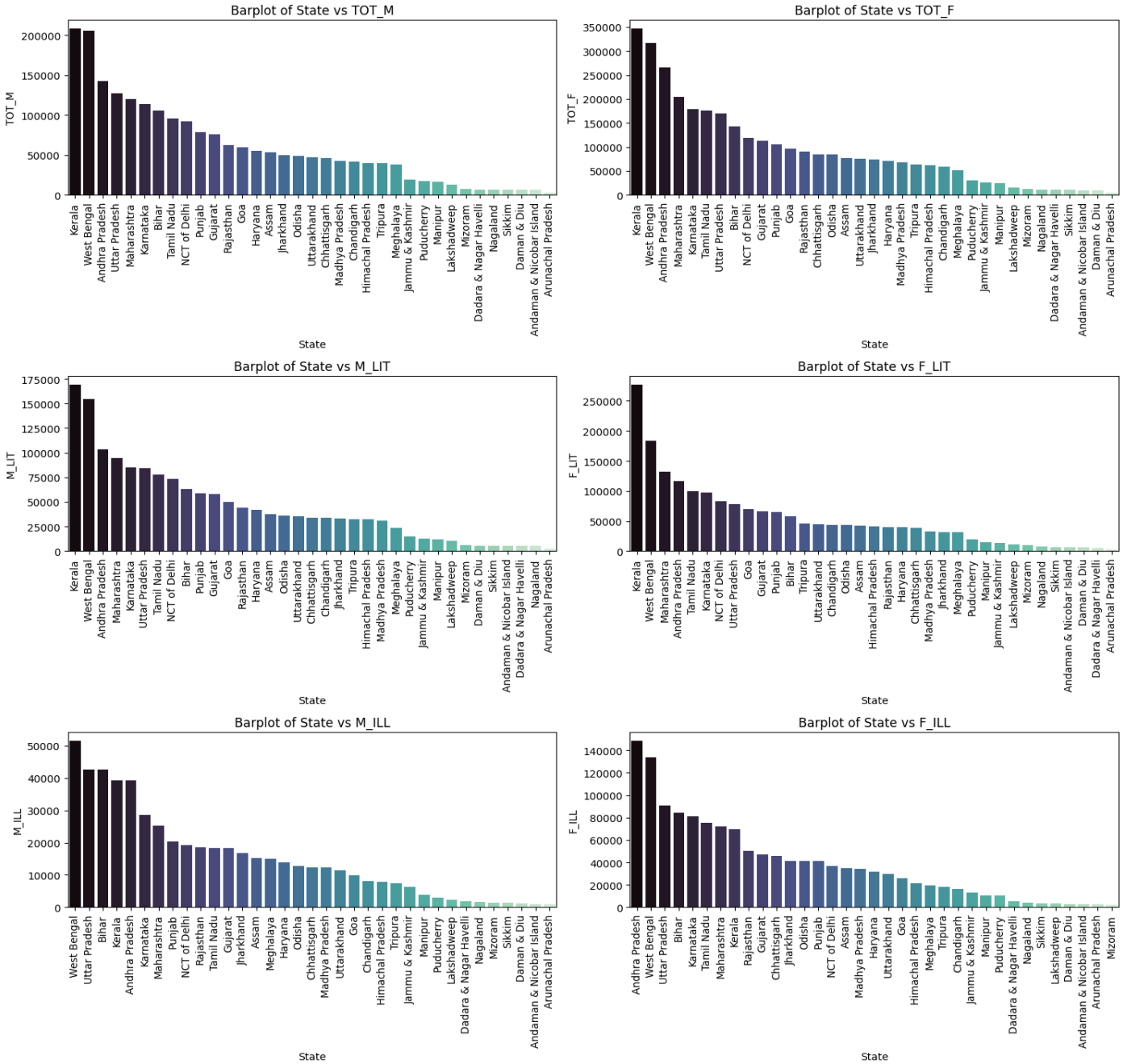


Figure 22 Bar Plot of State vs Numerical Variables

From the Bivariate analysis,

- The selected numerical variables are positively correlated
- Kerala has the highest Male, Female, Literate Male and Literate Female Population
- Arunachal has the lowest Male, Female, Literate Male and Literate Female Population
- West Bengal has the highest Illiterate Male population whereas Arunachal has the lowest illiterate male population
- Andhra Pradesh has the highest Illiterate Female population whereas Mizoram has the lowest illiterate female population
- Please note that the above high and low values entirely depends on the male & female population of the state. High values doesn't necessarily mean high as the corresponding state has high population. Percentage value from the State's population decides the high or low values.

Problem 2 - Data Preprocessing

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

- From Table 11 Statistical Summary of the Population DataWe can see that there are no null values. We also confirmed that there are no duplicates in the dataset.
- We could not find any data irregularities as the dataset contains 2 Categorical Columns and the remaining columns are numerical.
- Before scaling, we dropped 'State Code', 'Dist.Code', 'State', 'Area Name' from our data as these columns do not provide any value for our analysis

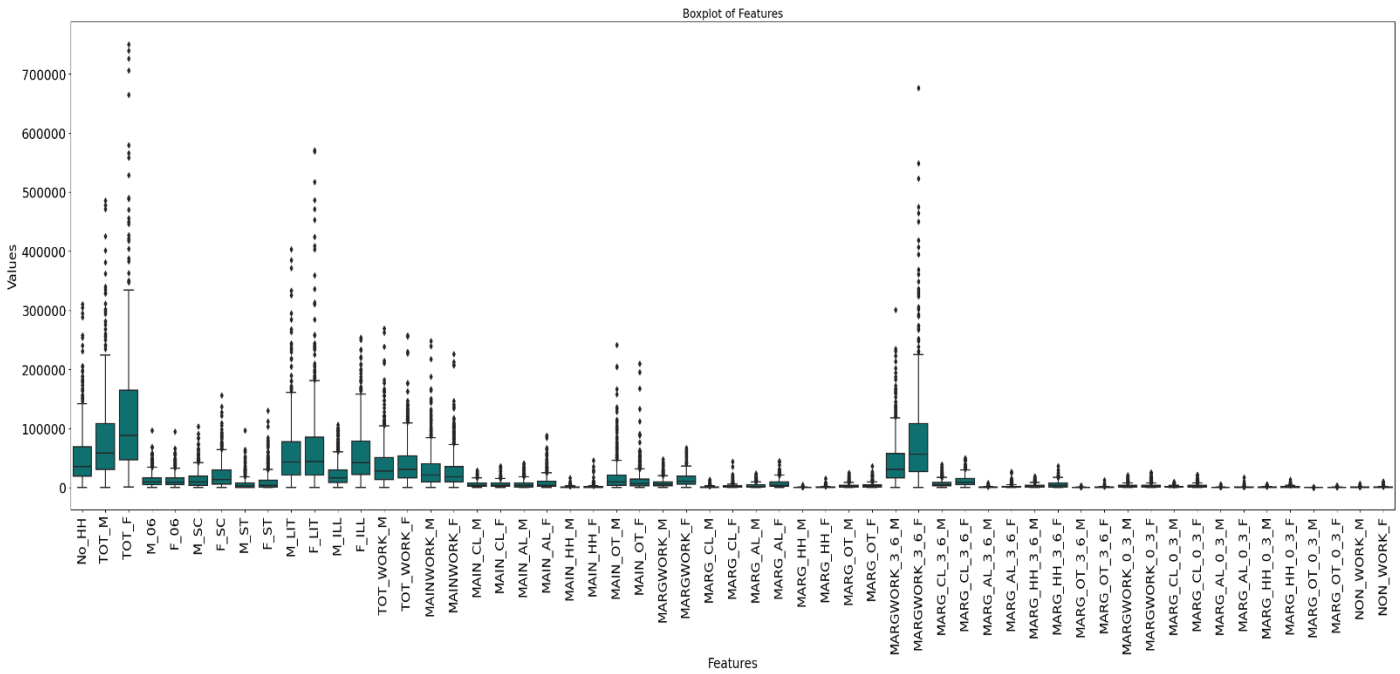


Figure 23 Boxplot of Features Before Scaling

- From the Figure 23 we could see that there are outliers in all the features.
- We are not treating outliers here as these values are true values and should be kept in data.
- We used StandardScaler from scikit learn library to scale our data and we check the effect of scaling on the outliers.

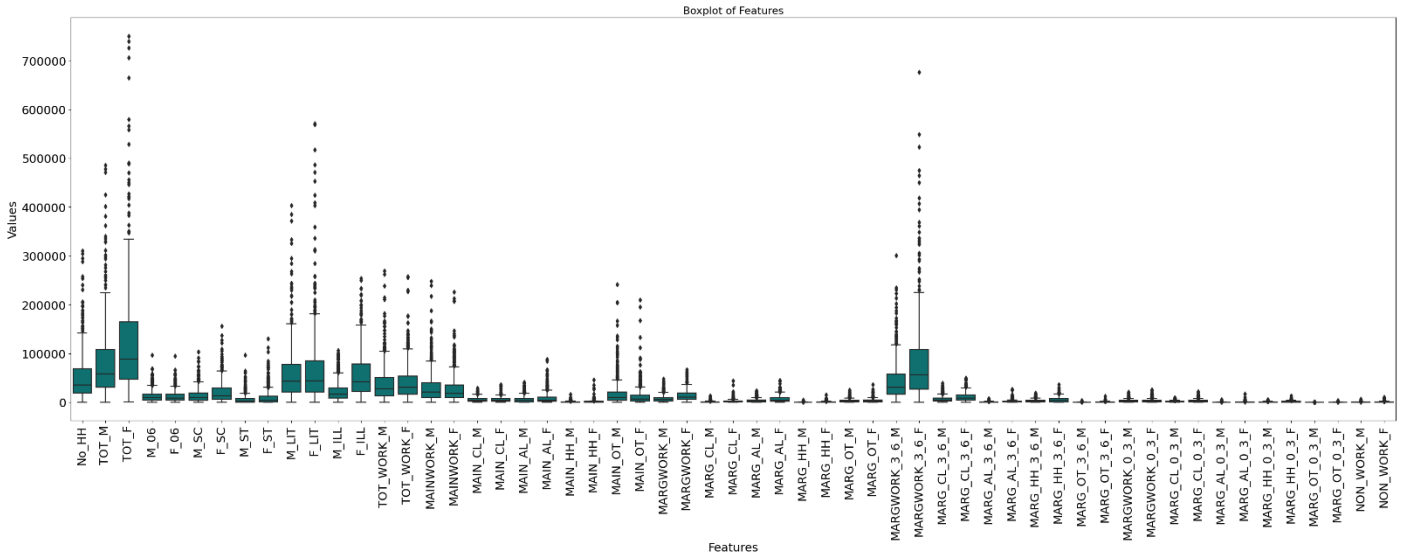


Figure 24 Boxplot of Population Data After Scaling

- From Figure 23 and Figure 24, we could interpret that there is no effect of scaling on the outliers of the features.

Problem 2 - PCA

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

- Before proceeding with PCA, we have to perform few tests to make sure that our dataset is suitable for performing PCA.

Bartletts Test of Sphericity:

- Bartlett's test of sphericity tests the hypothesis that the variables are uncorrelated in the population.
- H0: All variables in the data are uncorrelated
- Ha: At least one pair of variables in the data are correlated
- If the null hypothesis cannot be rejected, then PCA is not advisable.
- If the p-value is small, then we can reject the null hypothesis and agree that there is atleast one pair of variables in the data wihich are correlated hence PCA is recommended.

KMO Test:

- The Kaiser-Meyer-Olkin (KMO) - measure of sampling adequacy (MSA) is an index used to examine how appropriate PCA is.
- Generally, if MSA is less than 0.5, PCA is not recommended, since no reduction is expected. On the other hand, MSA > 0.7 is expected to provide a considerable reduction is the dimension and extraction of meaningful components.

- We passed the scaled data to the **calculate_bartlett_sphericity** and **calculate_kmo** methods from factor_analyzer library to calculate significance of correlations and adequacy of sample size and we obtained 0 for Bartletts Test of Sphericity and 0.803 for the KMO test.
- We can reject the H0 for Bartletts Test of Sphericity as the value is 0 and from the KMO test value we have adequate sample size. Hence PCA is recommended.

- We used the PCA method and passed number of columns in the scaled data as the n_components.
- We created a covariance matrix showing the relationship between the variables and plotted them on a heatmap for better understanding.

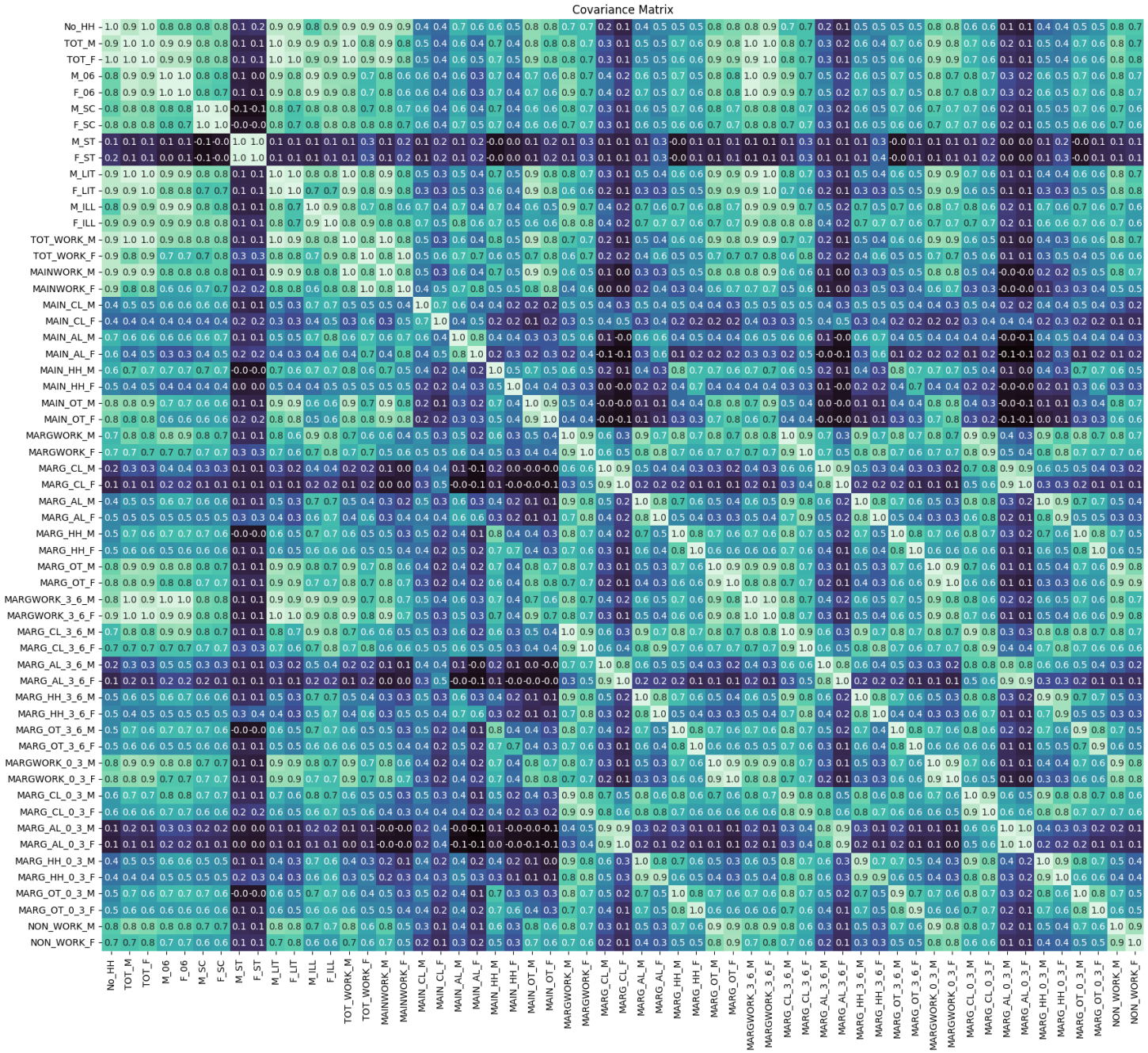


Figure 25 Covariance Matrix

- The below image shows us the eigen vectors.

```
array([[ 0.16,  0.17,  0.17, ...,  0.13,  0.15,  0.13],
       [-0.13, -0.09, -0.1 , ...,  0.05, -0.07, -0.07],
       [-0.  ,  0.06,  0.04, ..., -0.08,  0.11,  0.1 ],
       ...,
       [ 0.  ,  0.21,  0.25, ..., -0.07,  0.  , -0.07],
       [ 0.  ,  0.29, -0.21, ...,  0.04, -0.03,  0.01],
       [-0.  ,  0.19,  0.03, ..., -0.03, -0.14, -0.02]])
```

Figure 26 Eigen Vectors

- The below image shows us the eigen values.

```
array([3.18135647e+01, 7.86942415e+00, 4.15340812e+00, 3.66879058e+00,
       2.20652588e+00, 1.93827502e+00, 1.17617374e+00, 7.51159086e-01,
       6.17053743e-01, 5.28300887e-01, 4.29831189e-01, 3.53440201e-01,
       2.96163013e-01, 2.81275560e-01, 1.92158325e-01, 1.36267920e-01,
       1.13389199e-01, 1.06303946e-01, 9.72885376e-02, 8.01062194e-02,
       5.76089954e-02, 4.43955966e-02, 3.78910846e-02, 2.96360194e-02,
       2.70797618e-02, 2.34458139e-02, 1.45111511e-02, 1.09852268e-02,
       9.31507853e-03, 8.13540203e-03, 7.89250253e-03, 5.02601514e-03,
       2.59771182e-03, 1.06789820e-03, 7.13559124e-04, 2.47799812e-31,
       2.47799812e-31, 2.47799812e-31, 2.47799812e-31, 2.47799812e-31,
       2.47799812e-31, 2.47799812e-31, 2.47799812e-31, 2.47799812e-31,
       2.47799812e-31, 2.47799812e-31, 2.47799812e-31, 2.47799812e-31,
       2.47799812e-31, 2.47799812e-31, 2.47799812e-31, 2.47799812e-31,
       2.47799812e-31, 2.47799812e-31, 2.47799812e-31, 2.47799812e-31,
       2.47799812e-31])
```

Figure 27 Eigen Values

- We plotted a scree plot showing the percentage of information contained in each principal component.

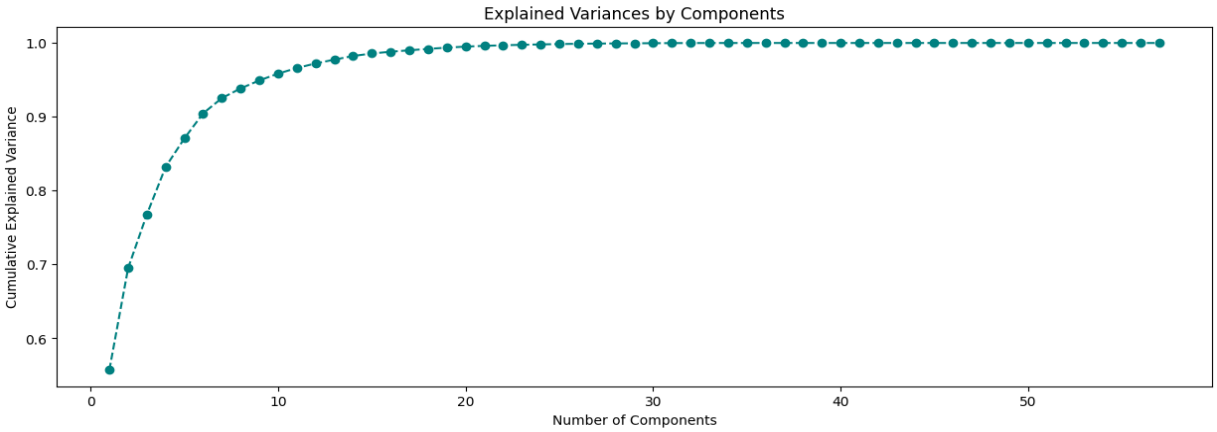


Figure 28 Scree Plot Showing Explained Variance by Components

- Also we plotted a waterfall chart showing similar information as shown in Figure 28.

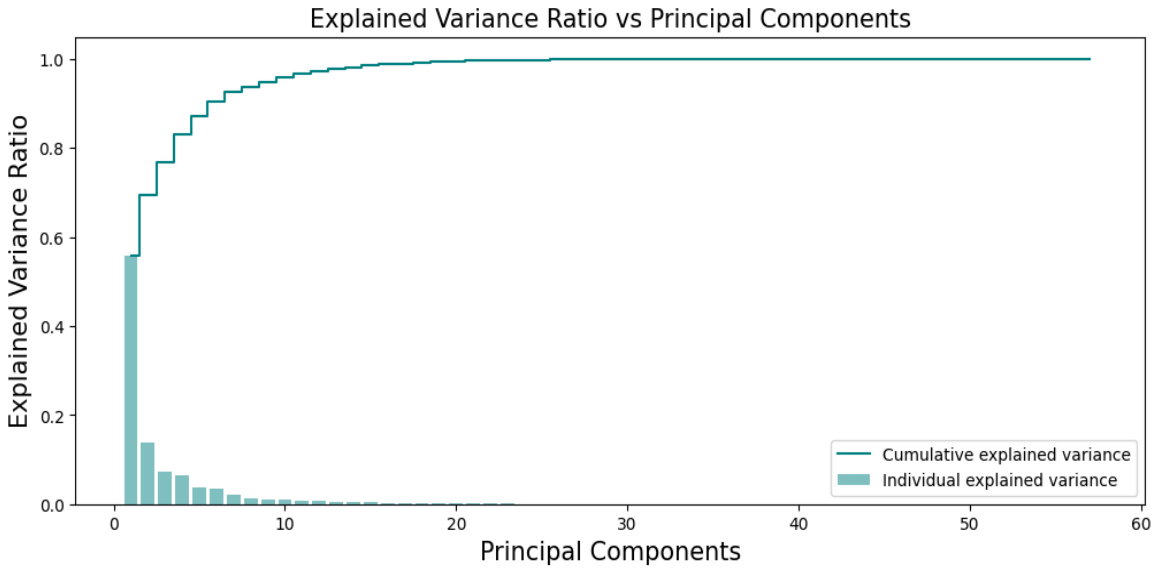


Figure 29 Waterfall chart Showing Explained Variance by Components

- The Number of PCs that explain at least 90% variance is **6**.

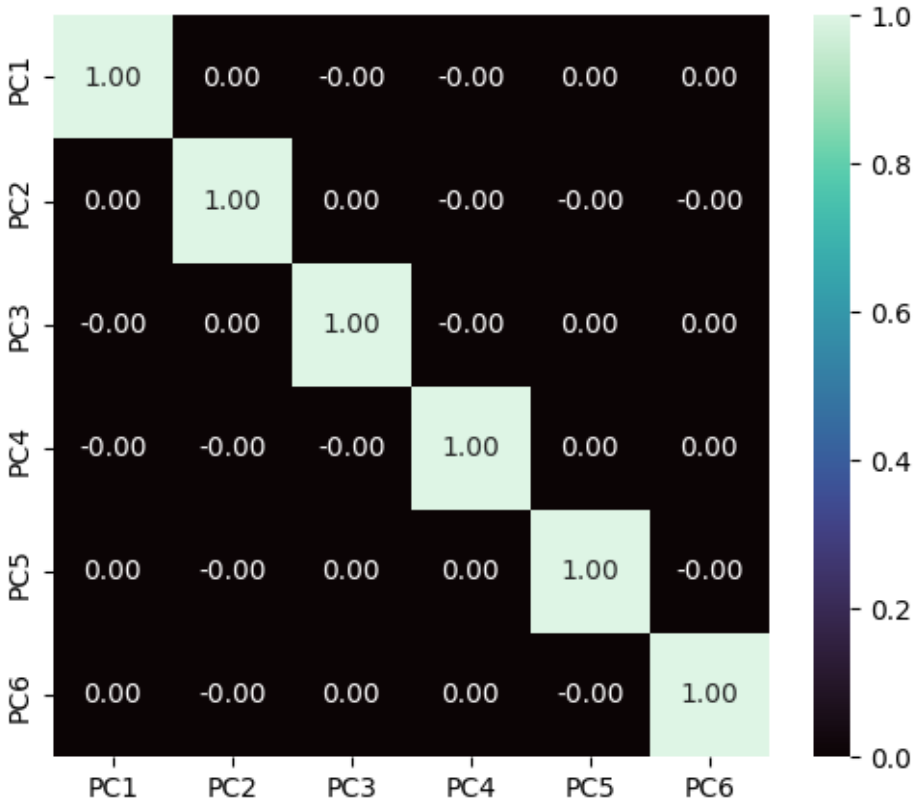


Figure 30 Correlation Heatmap of Principal Components

- The above heatmap confirms us that the Principal Components derived are valid.
- The linear equation for first principal component is,
  0.16 * No_HH (+) 0.17 * TOT_M (+) 0.17 * TOT_F (+) 0.16 * M_06 (+) 0.16 * F_06 (+) 0.15 * M_SC (+) 0.15 * F_SC (+) 0.03 * M_ST (+) 0.03 * F_ST (+) 0.16 * M_LIT (+) 0.15 * F_LIT (+) 0.16 * M_ILL (+) 0.17 * F_ILL (+) 0.16 * TOT_WORK_M (+) 0.15 * TOT_WORK_F (+) 0.15 * MAINWORK_M (+) 0.12 * MAINWORK_F (+) 0.1 * MAIN_CL_M (+) 0.07 * MAIN_CL_F (+) 0.11 * MAIN_AL_M (+) 0.07 * MAIN_AL_F (+) 0.13 * MAIN_HH_M (+) 0.08 * MAIN_HH_F (+) 0.12 * MAIN_OT_M (+) 0.11 * MAIN_OT_F (+) 0.16 * MARGWORK_M (+) 0.16 * MARGWORK_F (+) 0.08 * MARG_CL_M (+) 0.05 * MARG_CL_F (+) 0.13 * MARG_AL_M (+) 0.11 * MARG_AL_F (+) 0.14 * MARG_HH_M (+) 0.13 * MARG_HH_F (+) 0.16 * MARG_OT_M (+) 0.15 * MARG_OT_F (+) 0.16 * MARGWORK_3_6_M (+) 0.16 * MARGWORK_3_6_F (+) 0.17 * MARG_CL_3_6_M (+) 0.16 * MARG_CL_3_6_F (+) 0.09 * MARG_AL_3_6_M (+) 0.05 * MARG_AL_3_6_F (+) 0.13 * MARG_HH_3_6_M (+) 0.11 * MARG_HH_3_6_F (+) 0.14 * MARG_OT_3_6_M (+) 0.12 * MARG_OT_3_6_F (+) 0.15 * MARGWORK_0_3_M (+) 0.15 * MARGWORK_0_3_F (+) 0.15 * MARG_CL_0_3_M (+) 0.14 * MARG_CL_0_3_F (+) 0.05 * MARG_AL_0_3_M (+) 0.04 * MARG_AL_0_3_F (+) 0.12 * MARG_HH_0_3_M (+) 0.12 * MARG_HH_0_3_F (+) 0.14 * MARG_OT_0_3_M (+) 0.13 * MARG_OT_0_3_F (+) 0.15 * NON_WORK_M (+) 0.13 * NON_WORK_F (+)