Image courtesy: https://blog.reffascode.de/tag/machine-learning/

Machine Learning 1 Project

Business Report

May 05, 2024

Authored by: Kartik Trivedi

List of Contents

List of Figures

List of Tables

List of Equations

Data Dictionary

Problem 1

Column Name	Column Description	Data Type
Timestamp	The Timestamp of the particular Advertisement.	Object
InventoryType	The Inventory Type of the particular Advertisement. Format 1 to 7. This is a Categorical Variable.	Object
Ad - Length	The Length Dimension of the particular Advertisement.	Int 64
Ad- Width	The Width Dimension of the particular Advertisement.	Int 64
Ad Size	The Overall Size of the particular Advertisement. Length*Width.	Int 64
Ad Type	The type of the particular Advertisement. This is a Categorical Variable.	Object
Platform	The platform in which the particular Advertisement is displayed. Web, Video or App. This is a Categorical Variable.	Object
Device Type	The type of the device which supports the particular Advertisement. This is a Categorical Variable.	Object
Format	The Format in which the Advertisement is displayed. This is a Categorical Variable.	Object
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.	Int 64
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.	Int 64
Impressions	The impression counts of the particular Advertisement out of the total available impressions.	Int 64
Clicks	It is a marketing metric that counts the number of times users have clicked on the particular advertisement to reach an online property.	Int 64
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.	Float 64
Fee	The percentage of the Advertising Fees payable by Franchise Entities.	Float 64
Revenue	It is the income that has been earned from the particular advertisement.	Float 64
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.	Float 64
CPM	CPM stands for "cost per 1000 impressions." Formula used here is CPM = (Total Campaign Spend / Number of Impressions) * 1,000. Note	Float 64

	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.	Float 64

Problem 2

Name	Description	Data Type
State	State Code	Int 64
District	District Code	Int 64
State	State Name	Object
TRU1	Area Name	Object
No_HH	No of Household	Int 64
TOT_M	Total population Male	Int 64
TOT_F	Total population Female	Int 64
M_06	Population in the age group 0-6 Male	Int 64
F_06	Population in the age group 0-6 Female	Int 64
M_SC	Scheduled Castes population Male	Int 64
F_SC	Scheduled Castes population Female	Int 64
M_ST	Scheduled Tribes population Male	Int 64
F_ST	Scheduled Tribes population Female	Int 64
M_LIT	Literates' population Male	Int 64
F_LIT	Literates' population Female	Int 64
M_ILL	Illiterate Male	Int 64
F_ILL	Illiterate Female	Int 64
TOT_WORK_M	Total Worker Population Male	Int 64
TOT_WORK_F	Total Worker Population Female	Int 64
MAINWORK_M	Main Working Population Male	Int 64
MAINWORK_F	Main Working Population Female	Int 64
MAIN_CL_M	Main Cultivator Population Male	Int 64
MAIN_CL_F	Main Cultivator Population Female	Int 64
MAIN_AL_M	Main Agricultural Labourers Population Male	Int 64
MAIN_AL_F	Main Agricultural Labourers Population Female	Int 64
MAIN_HH_M	Main Household Industries Population Male	Int 64

MAIN_HH_F	Main Household Industries Population Female	Int 64
MAIN_OT_M	Main Other Workers Population Male	Int 64
MAIN_OT_F	Main Other Workers Population Female	Int 64
MARGWORK_M	Marginal Worker Population Male	Int 64
MARGWORK_F	Marginal Worker Population Female	Int 64
MARG_CL_M	Marginal Cultivator Population Male	Int 64
MARG_CL_F	Marginal Cultivator Population Female	Int 64
MARG_AL_M	Marginal Agriculture Labourers Population Male	Int 64
MARG_AL_F	Marginal Agriculture Labourers Population Female	Int 64
MARG_HH_M	Marginal Household Industries Population Male	Int 64
MARG_HH_F	Marginal Household Industries Population Female	Int 64
MARG_OT_M	Marginal Other Workers Population Male	Int 64
MARG_OT_F	Marginal Other Workers Population Female	Int 64
MARGWORK_3_6_M	Marginal Worker Population 3-6 Male	Int 64
MARGWORK_3_6_F	Marginal Worker Population 3-6 Female	Int 64
MARG_CL_3_6_M	Marginal Cultivator Population 3-6 Male	Int 64
MARG_CL_3_6_F	Marginal Cultivator Population 3-6 Female	Int 64
MARG_AL_3_6_M	Marginal Agriculture Labourers Population 3-6 Male	Int 64
MARG_AL_3_6_F	Marginal Agriculture Labourers Population 3-6 Female	Int 64
MARG_HH_3_6_M	Marginal Household Industries Population 3-6 Male	Int 64
MARG_HH_3_6_F	Marginal Household Industries Population 3-6 Female	Int 64
MARG_OT_3_6_M	Marginal Other Workers Population Person 3-6 Male	Int 64
MARG_OT_3_6_F	Marginal Other Workers Population Person 3-6 Female	Int 64
MARGWORK_0_3_M	Marginal Worker Population 0-3 Male	Int 64
MARGWORK_0_3_F	Marginal Worker Population 0-3 Female	Int 64
MARG_CL_0_3_M	Marginal Cultivator Population 0-3 Male	Int 64
MARG_CL_0_3_F	Marginal Cultivator Population 0-3 Female	Int 64
MARG_AL_0_3_M	Marginal Agriculture Labourers Population 0-3 Male	Int 64
MARG_AL_0_3_F	Marginal Agriculture Labourers Population 0-3 Female	Int 64
MARG_HH_0_3_M	Marginal Household Industries Population 0-3 Male	Int 64
MARG_HH_0_3_F	Marginal Household Industries Population 0-3 Female	Int 64
MARG_OT_0_3_M	Marginal Other Workers Population 0-3 Male	Int 64
MARG_OT_0_3_F	Marginal Other Workers Population 0-3 Female	Int 64
NON_WORK_M	Non-Working Population Male	Int 64
NON_WORK_F	Non-Working Population Female	Int 64

Executive Summary

Problem 1

Background Information

Ads 24x7, a digital marketing company has raised a funding of $ 10 Million in the seed round using which they are expanding into marketing analytics. For this they have hired a data analyst who has been tasked with a project aimed at categorizing ad types using the features contained in the data gathered by their marketing intelligence team.

Business Objective

Categorize advertisements into various groups based on ad types, target audiences, and marketing strategies. This insight will aid Ads 24x7 in optimizing their digital marketing campaigns by effectively allocating budgets and deploying customized ad content for specific audience segments.

Problem Statement

The aim of this analysis is to efficiently recognize various ad types and marketing strategies tailored to different target audiences, utilizing clustering techniques to generate clusters that can be used by Ads 24x7 in designing their future marketing campaigns.

Clusters

Cluster 0: Skyscraper being tall and sleek

Cluster 1: Long Rectangle being rectangle with dimensions 153x558

Cluster 2: Horizontal Rectangle being rectangle whose length is more than its width

Cluster 3: Wide Skyscraper being wider than skyscraper

Cluster 4: Wide Rectangle having slightly lower length and higher width than long rectangle

Conclusion

After analyzing the above data, Ads 24x7 needs to implement substantial changes to its business model. It not only falls short in financial terms, with ad spending nearly 50% higher than revenue, but also in terms of costs and meeting key sector metrics like CTR.

For improvement purposes the following recommendation are made

Recommendations

1. Based on the data provided by advertising platform, the CTR for them based on Matched_Queries where keywords searched matched keywords in the ads resulting in click to Available_Impressions is between the range of 50 and 70, while for Ads 24x7 the maximum CTR achieved is only 15, in fact for Horizontal Rectangle and Skyscraper which have most

impressions this value is below 0.4. Thus, it is highly recommended for Ads 24x7 that a detail Keyword analysis project is carried out to study and extract the keywords which can be used to improve the ad content.

2. Ads 24x7 needs to look at the ads spent, as per the principle of unit's economics per unit cost comes down as the number of units goes up, based on this principle for ad type which have high impressions their CPM should be low like in case of CPC we can see that Long Rectangle, Wide Rectangle, and Wide Skyscraper which have high number of clicks have low CPC. However, CPM for Horizontal Rectangle and Skyscrapers is very high despite them having high number of impressions.
3. Amongst the clusters Long Rectangle, Wide Rectangle, and Wide Skyscraper are performing much better than Horizontal Rectangle and Skyscraper in terms of CTR and CPC, for the time being it is highly recommended that Ads 24x7 should use these ad types more until the above mentioned studies are carried out changes are made, and though they have high CPM, we expect as the number of impressions for these clusters goes up the CPM would come down like in case of Horizontal Rectangle and Skyscraper which could help improve revenue while reducing spendings.

Problem 2

Background Information

Population census is held in India every ten years, a tradition dating back over a century. These censuses gather extensive data on various aspects of the population, which is then compiled and presented on a district-wise basis. However, due to the multitude of characteristics covered in the census, the compiled data contains numerous variables, making it challenging to extract useful insights.

Problem Statement

The objective of this analysis is to utilize PCA technique to determine the optimal number of principal components that capture the greatest variance in the abstract of population census data, specifically focusing on female-headed households excluding institutional households. This process aims to reduce the dimensionality of the data.

Inference

1. Based on the above heatmap different PC's explain highest variance for different features which could be understood better from the table below

PC1	PC2	PC3	PC4	PC5	PC6
No_HH	F_LIT	MAIN_AL_M	TOT_WORK_F	M_SC	MAIN_HH_M
TOT_M	MAINWORK_M	MARG_CL_F	MAINWORK_F	F_SC	MAIN_HH_F
TOT_F	MAIN_OT_M	MARG_AL_M	MAIN_CL_F	M_ST	MARG_HH_F
M_06	MAIN_OT_F	MARG_AL_F	MAIN_AL_F	F_ST	MARG_OT_3_6_F
F_06	MARG_CL_M	MARG_AL_3_6_F	MARG_HH_M	MAIN_CL_M	MARG_OT_0_3_F
M_LIT	MARG_AL_3_6_M	MARG_HH_3_6_M	MARG_OT_3_6_M	MARG_OT_F	
M_ILL	MARG_CL_0_3_M	MARG_HH_3_6_F	MARG_OT_0_3_M	NON_WORK_M	
F_ILL	MARG_CL_0_3_F	MARG_AL_0_3_M		NON_WORK_F	
TOT_WORK_M	MARG_HH_0_3_M	MARG_AL_0_3_F			
MARGWORK_M		MARG_HH_0_3_F			
MARGWORK_F					
MARG_OT_M					
MARGWORK_3_6_M					
MARGWORK_3_6_F					
MARG_CL_3_6_M					
MARG_CL_3_6_F					
MARGWORK_0_3_M					
MARGWORK_0_3_F					

Table 1: Features highest explaining PCs

2. PC1 captures most variance for 19 features and most of these features are related to overall male and female population demographics like total number of households, total male and female population, total male and female population between 0 and 6 years and so on. We can name this as PC_Population_demography as it explains key population questions like total population, literacy and illiteracy in population, sex ratio etc.
3. PC2 explains most variance for columns F_LIT, MAINWORK_M, MAIN_OT_M, MAIN_OT_F, MARG_CL_M, MARG_AL_3_6_M, MARG_CL_0_3_M, MARG_CL_0_3_F, MARG_HH_0_3_M, most of the features in this PC is related to working male population so we can name it as PC_Male_Workers.
4. PC3 has explain most variance for columns related to employment in farming sector providing information regarding main and marginal employment in farming sector, we can name it as PC_Farming_Workforce.
5. In PC4 most of the features are about female employment covering total working female, main working female etc so we can name it as PC_Female_Workers
6. PC5 captures variance for backward classes that is for SC and ST population so we can name it as PC_Backward_classes.
7. PC6 which captures the lowest variance in 6 PC's covers mostly columns related to workers in household industries so we can name it as PC_household_industries.

Problem 1

1.1 Background Information

Ads 24x7, a digital marketing company has raised a funding of $ 10 Million in the seed round using which they are expanding into marketing analytics. For this they have hired a data analyst who has been tasked with a project aimed at categorizing ad types using the features contained in the data gathered by their marketing intelligence team.

1.2 Business Context

Categorize advertisements into various groups based on ad types, target audiences, and marketing strategies. This insight will aid Ads 24x7 in optimizing their digital marketing campaigns by effectively allocating budgets and deploying customized ad content for specific audience segments.

1.3 Problem Statement

The aim of this analysis is to efficiently recognize various ad types and marketing strategies tailored to different target audiences, utilizing clustering techniques to generate clusters that can be used by Ads 24x7 in designing their future marketing campaigns.

1.4 METHODOLOGY

Import the libraries - Load the data - Check the structure of the data - Check the types of the data – Check for and treat (if needed) missing values - Check the statistical summary - Check for and treat (if needed) data irregularities – Univariate Analysis – Bivariate Analysis – Outlier Treatment – Data Scaling – Hierarchical Clustering – K-Means Clustering – Cluster Profiling – Data Analysis by Clusters – Conclusion

Key Points

1. **Data Collection**: Data was provided by marketing intelligence team which contains information related to various marketing campaigns run by them.
2. **Data Cleaning and Pre-processing:** Dataset was checked for duplicates, missing values, bad data and outliers. Missing values and outliers were found in the dataset, missing values and outliers were treated as per the procedure and data was scaled to make it fit for applying the clustering algorithm.
3. **Univariate Analysis:** Individual variables were analyzed using boxplot and histogram to understand distribution, central tendency and variability of variables.
4. **Bivariate Analysis:** All the variables were examined with the aim of gaining deeper insights about the various ad campaigns run by the company.

5. **Visualization Techniques:** In the report we have used histograms and boxplot for univariate analysis, in bivariate analysis, to understand correlation between numeric variables heatmap and pair plot are used, bar plot is used to understand relationship between categorical and numeric variables and cross tab is used to understand relationship between categorical variables.
6. **Tools and Software:** We have carried out the analysis using programming language python on Jupyter notebook. For this analysis Python libraries Numpy, Pandas, Matplotlib, Seaborn, Scipy, Scikit-learn and Math were used.
7. **Assumptions and Limitations:** For this analysis we took following assumptions
   a) Extracted hour data into another column from Timestamp column, naming the column as 'ads_hour' as we believe that hour at which ads were run might be useful in analysis and have deleted the Timestamp column as it was not relevant for the analysis.
   b) For 'ads_hour' column extracted from Timestamp column there are no bad data and so for our analysis we will consider that Timestamp column also does not has any bad data.

1.5 Data Overview

1. **Data Description:** Dataset has 23066 rows and 19 columns.

```
shape of the dataset
--------------------------------------------------------------------------------

(23066, 19)
```

Table 2: Dataset Shape

2. **Dataset Information:** Of the nineteen columns in the dataset 6 are object type, 7 are int 64 type and 6 are float type.

```
information of features
-----------------------------------------------------
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 23066 entries, 0 to 23065
Data columns (total 19 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Timestamp              23066 non-null  object
 1   InventoryType          23066 non-null  object
 2   Ad - Length            23066 non-null  int64
 3   Ad- Width              23066 non-null  int64
 4   Ad Size                23066 non-null  int64
 5   Ad Type                23066 non-null  object
 6   Platform               23066 non-null  object
 7   Device Type            23066 non-null  object
 8   Format                 23066 non-null  object
 9   Available_Impressions  23066 non-null  int64
 10  Matched_Queries        23066 non-null  int64
 11  Impressions            23066 non-null  int64
 12  Clicks                 23066 non-null  int64
 13  Spend                  23066 non-null  float64
 14  Fee                    23066 non-null  float64
 15  Revenue                23066 non-null  float64
 16  CTR                    18330 non-null  float64
 17  CPM                    18330 non-null  float64
 18  CPC                    18330 non-null  float64
dtypes: float64(6), int64(7), object(6)
memory usage: 3.3+ MB
```

Table 3: Dataset Information

3. **Missing Value Check:** There were missing values in 3 columns namely CTR, CPM and CPC, 4736 values were missing for each of these columns. These missing values were treated using formula available for each of these columns

$$\text{CTR} = \text{Total Measured Clicks} / \text{Total Measured Ad Impressions} \times 100$$

Equation 1: CTR calculation

$$\text{CPM} = (\text{Total Campaign Spend} / \text{Number of Impressions}) * 1{,}000$$

Equation 2: CPM calculation

$$\text{CPC} = \text{Total Cost (spend)} / \text{Number of Clicks.}$$

Equation 3: CPC calculation

- In the above equations Total Measured Ad Impressions/ Number of Impressions means Impressions.
- Total Measured Clicks/ Number of Clicks means Clicks
- Total Campaign Spend/ Total Cost (Spend) means Spend

```
missing values
--------------------------------------------------------------------------------

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
```

Table 4: Missing values information

4. **Duplicate Values:** Data was checked for duplicate values and no duplicates were found

```
checking for duplicates
--------------------------------------------------------------------------------
number of dupliacte rows: 0
```

Table 5: Data Duplicates

5. **Statistical Summary:**

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

Table 6: Statistical Summary

6. Frequency Distribution of Categorical Columns:

```
frequency distribution of categorical columns
------------------------------------------------------------------------------
value counts for Timestamp
------------------------------------------------------------------------------
Timestamp
2020-11-13-22    13
2020-11-20-9     13
2020-11-14-23    13
2020-10-18-1     13
2020-9-23-15     13
                 ..
2020-9-2-5       10
2020-9-4-19      10
2020-9-2-11      10
2020-9-3-11       9
2020-9-1-16       2
Name: count, Length: 2018, dtype: int64
```

```
value counts for InventoryType
--------------------------------------------------------------------------------
InventoryType
Format4    7165
Format5    4249
Format1    3814
Format3    3540
Format6    1850
Format2    1789
Format7     659
Name: count, dtype: int64


value counts for Ad Type
--------------------------------------------------------------------------------
Ad Type
Inter224    1658
Inter217    1655
Inter223    1654
Inter219    1650
Inter221    1650
Inter222    1649
Inter229    1648
Inter227    1647
Inter218    1645
inter230    1644
Inter220    1644
Inter225    1643
Inter226    1640
Inter228    1639
Name: count, dtype: int64
```

```
value counts for Platform
--------------------------------------------------------------------------------
Platform
Video    9873
Web      8251
App      4942
Name: count, dtype: int64


value counts for Device Type
--------------------------------------------------------------------------------
Device Type
Mobile     14806
Desktop     8260
Name: count, dtype: int64


value counts for Format
--------------------------------------------------------------------------------
Format
Video      11552
Display    11514
Name: count, dtype: int64
```

Table 7: Frequency Distribution of categorical columns

Key observations

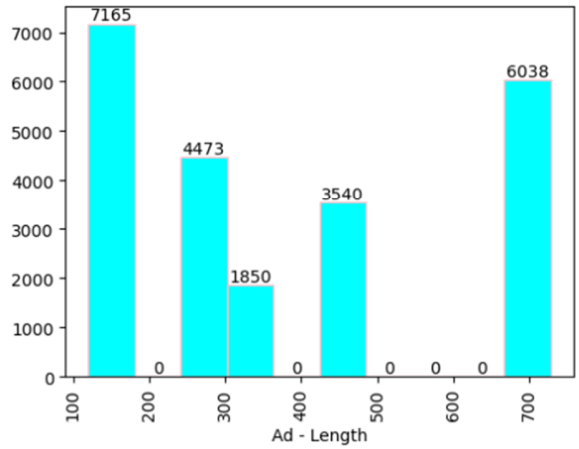
1. Dataset has 23066 rows and 19 columns.
2. According to the data dictionary, we anticipated the data frame to comprise 13 numeric columns, 6 categorical/object columns, and 1 datetime datatype column. However, as per the data info dataset has 13 numeric columns and 6 categorical/object column. This discrepancy is due to the fact that timestamp column which was expected to have datetime datatype has object datatype.
3. Based on the statistical summary of the data we can observe that there is significantly high variance for the numeric columns and this variance is present both between the features as well as within the features. Due to this we will have to scale the data before we can apply clustering models.
4. There are missing values for columns CTR, CPM and CPC which we have imputed during pre-processing phase.
5. There are no duplicates in the data.
6. For the columns with datatype as object there is no bad data except for column Timestamp which cannot be checked completely due to the fact that there are over 2000 unique values in this column. We checked for bad data in that column during pre-processing phase when we created a new column by extracting the hour from Timestamp column.

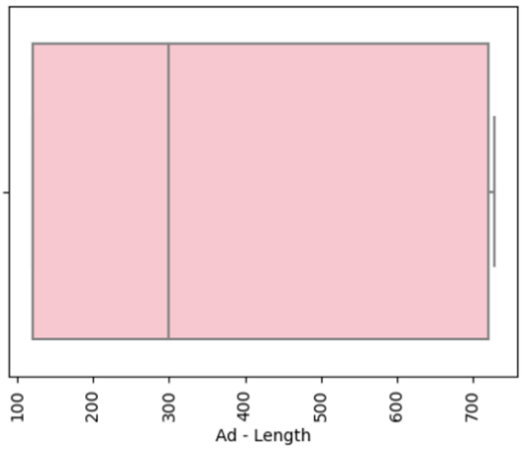
1.6 Exploratory Data Analysis

1.6.1 Univariate Analysis

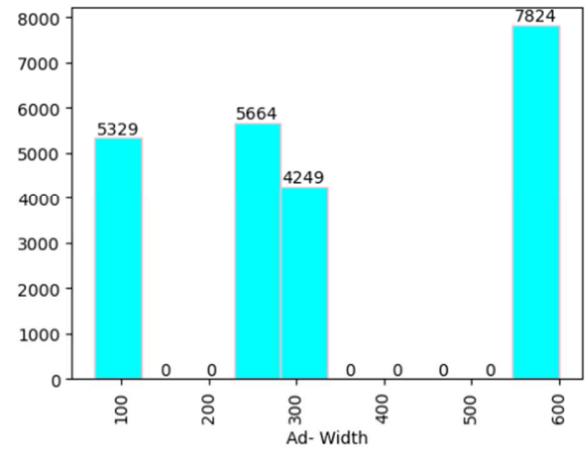
For numeric columns

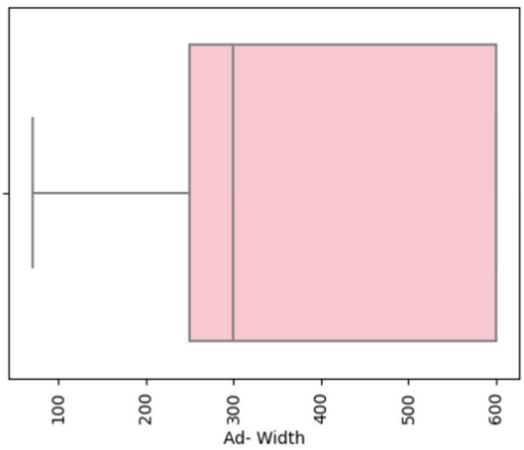
```
Distribution of Ad - Length
--------------------------------------------------------------------------------
```



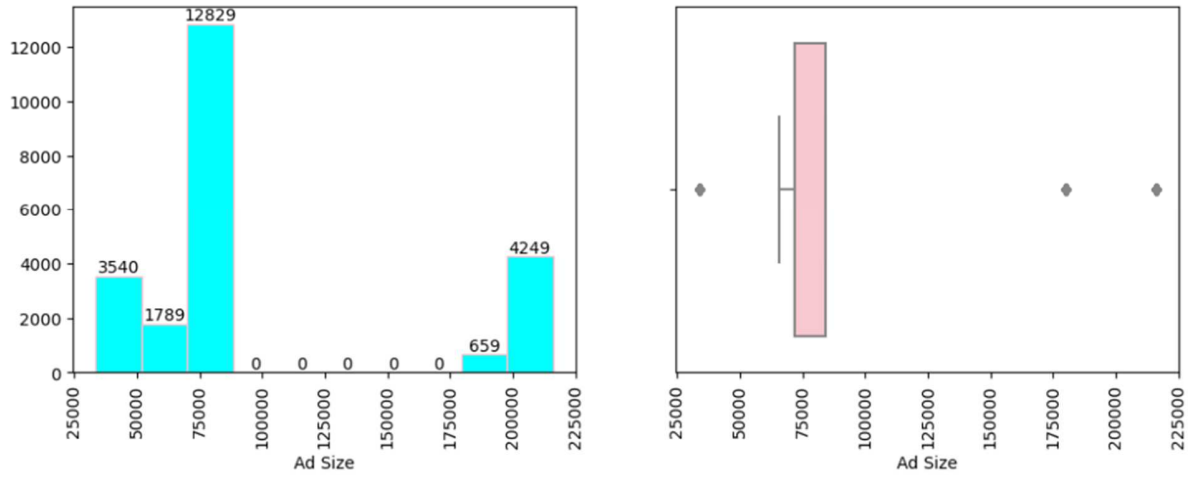



```
Distribution of Ad- Width
--------------------------------------------------------------------------------
```

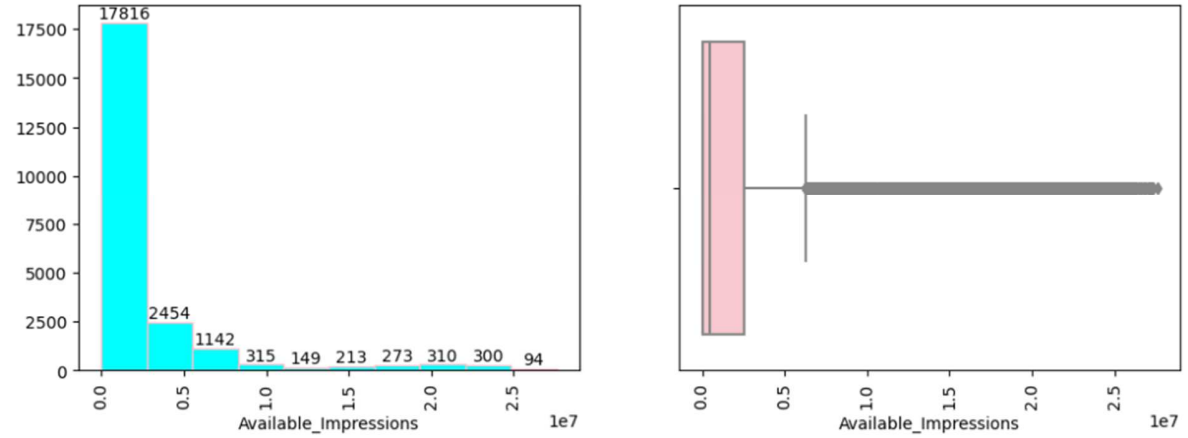




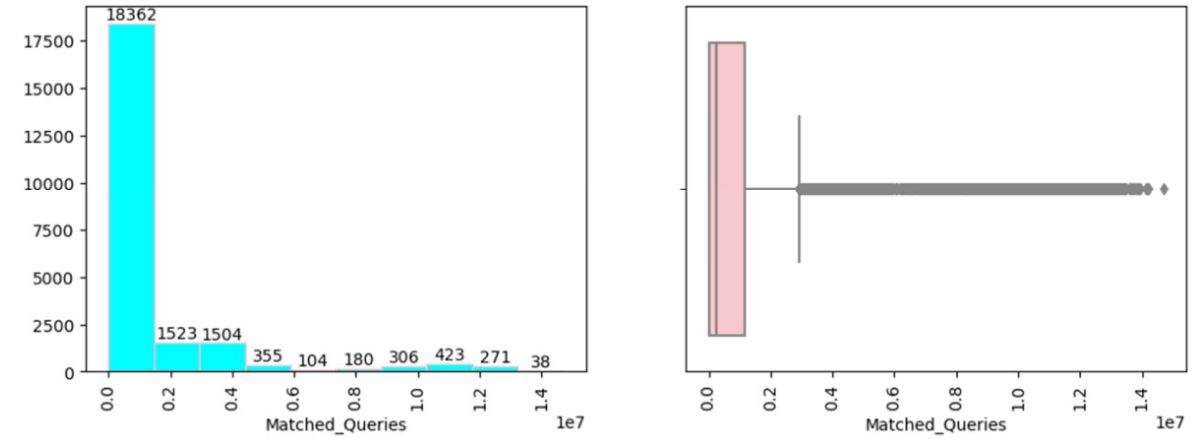
Distribution of Ad Size



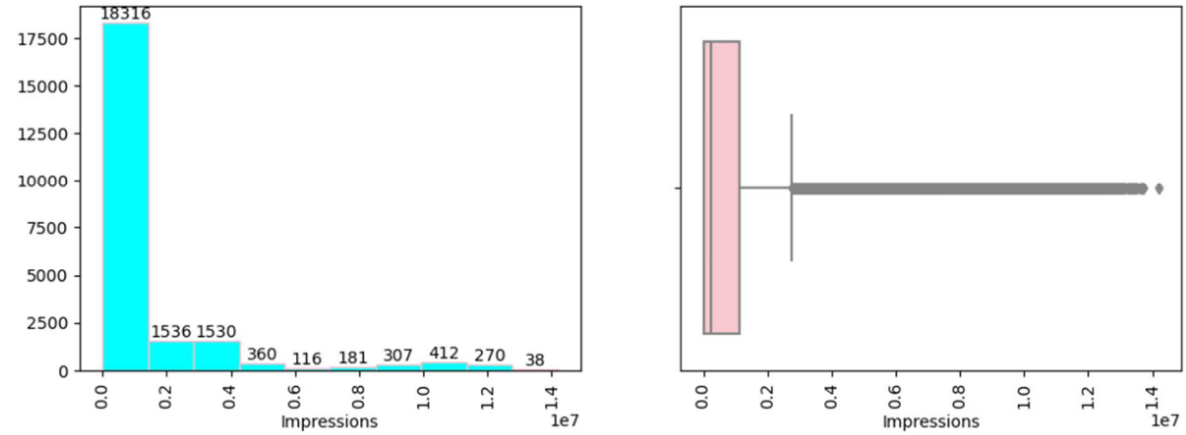
Distribution of Available_Impressions



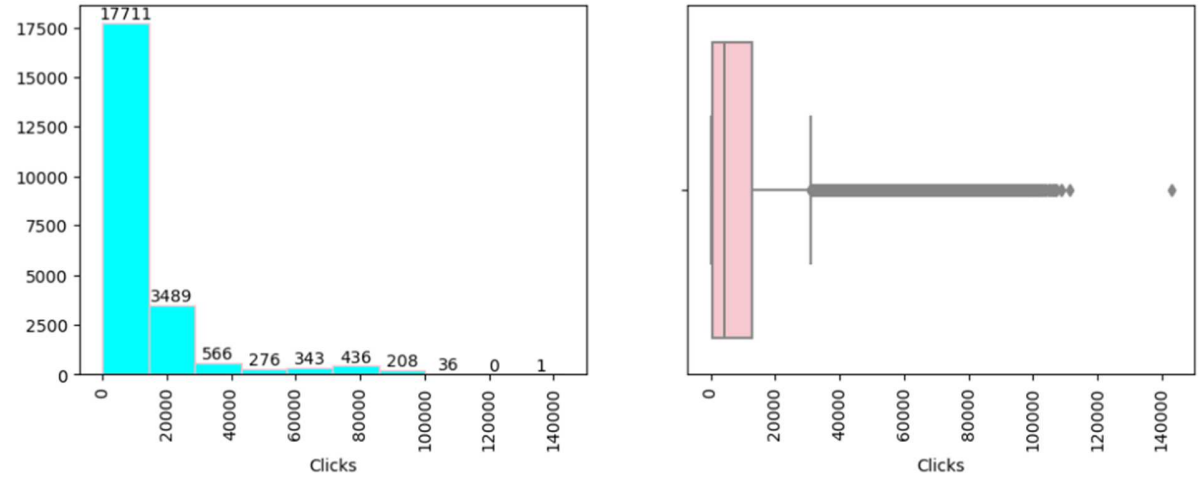
Distribution of Matched_Queries



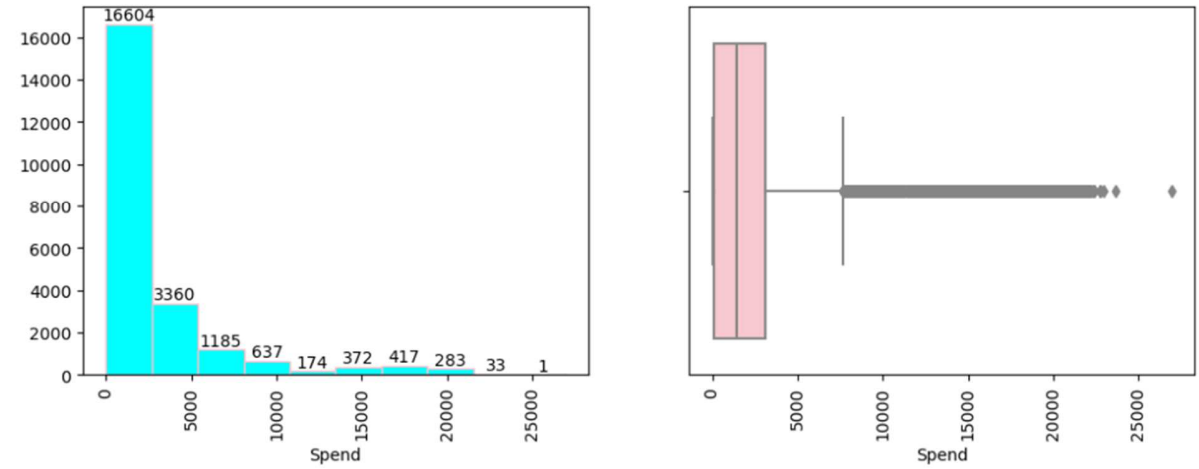
Distribution of Impressions



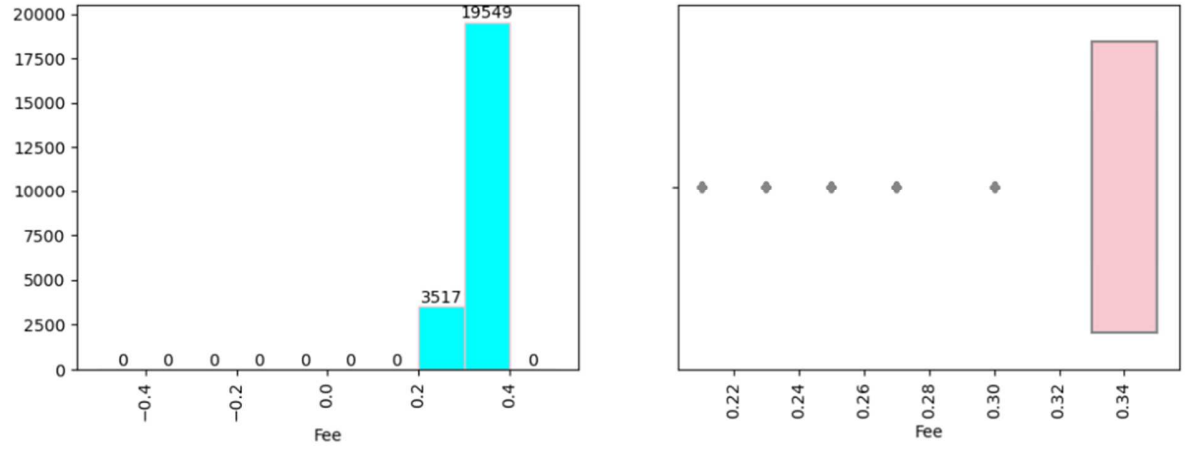
Distribution of Clicks



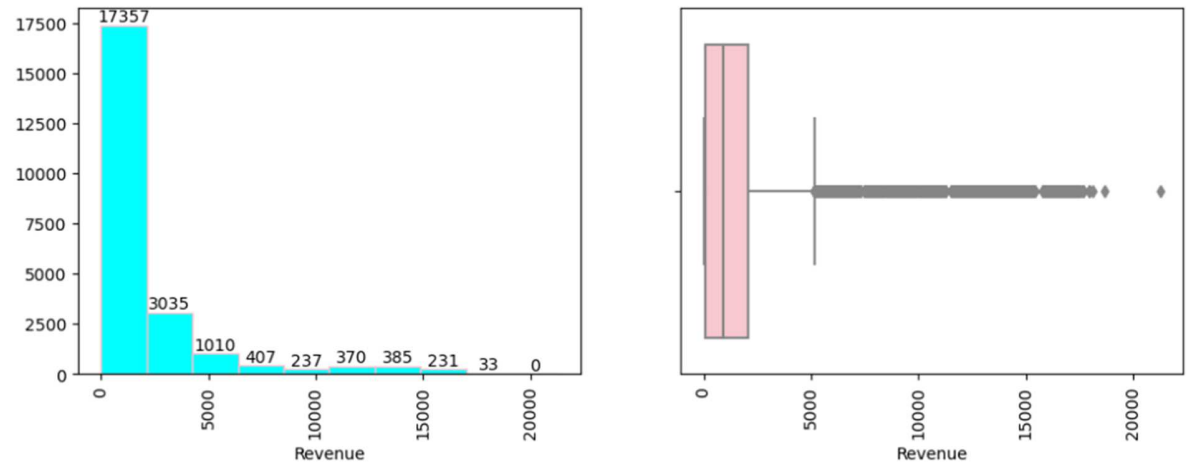
Distribution of Spend



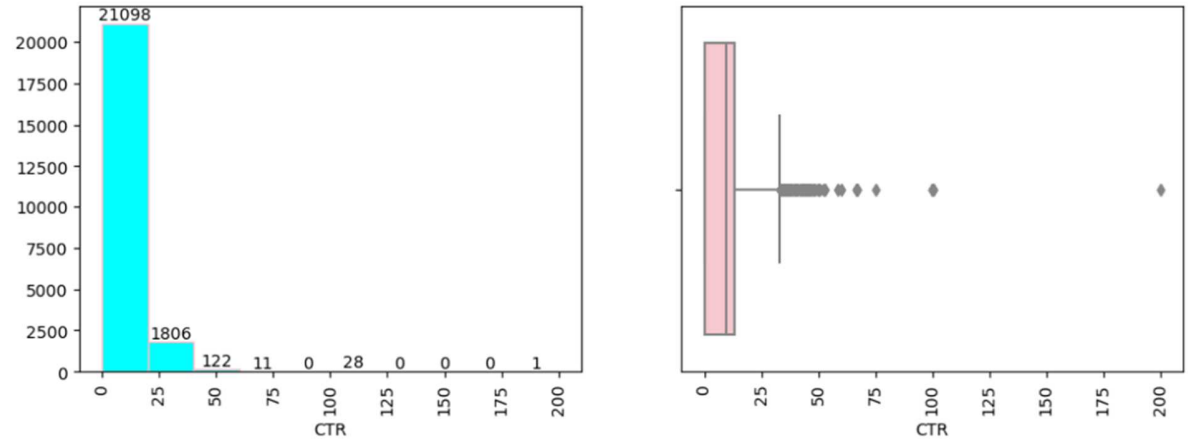
Distribution of Fee



Distribution of Revenue



Distribution of CTR



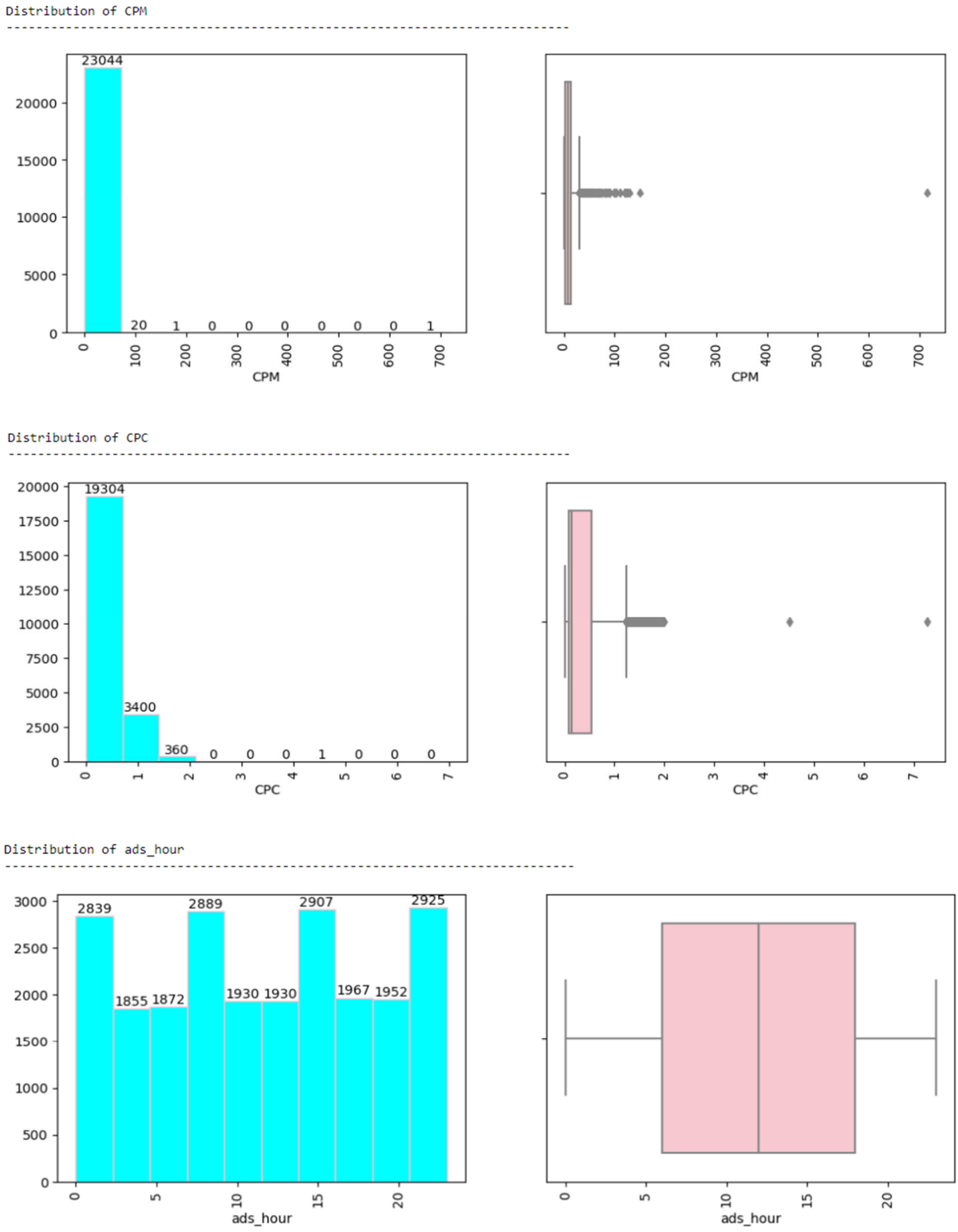


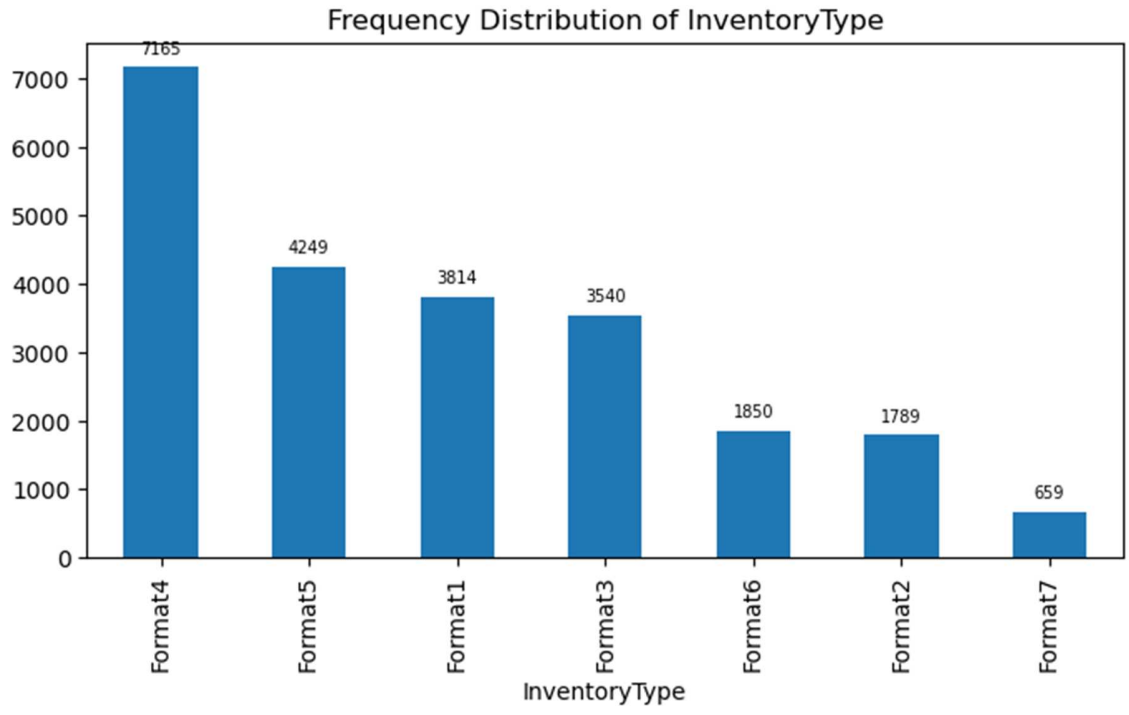
Figure 1: Univariate Analysis numeric columns

Key Observations

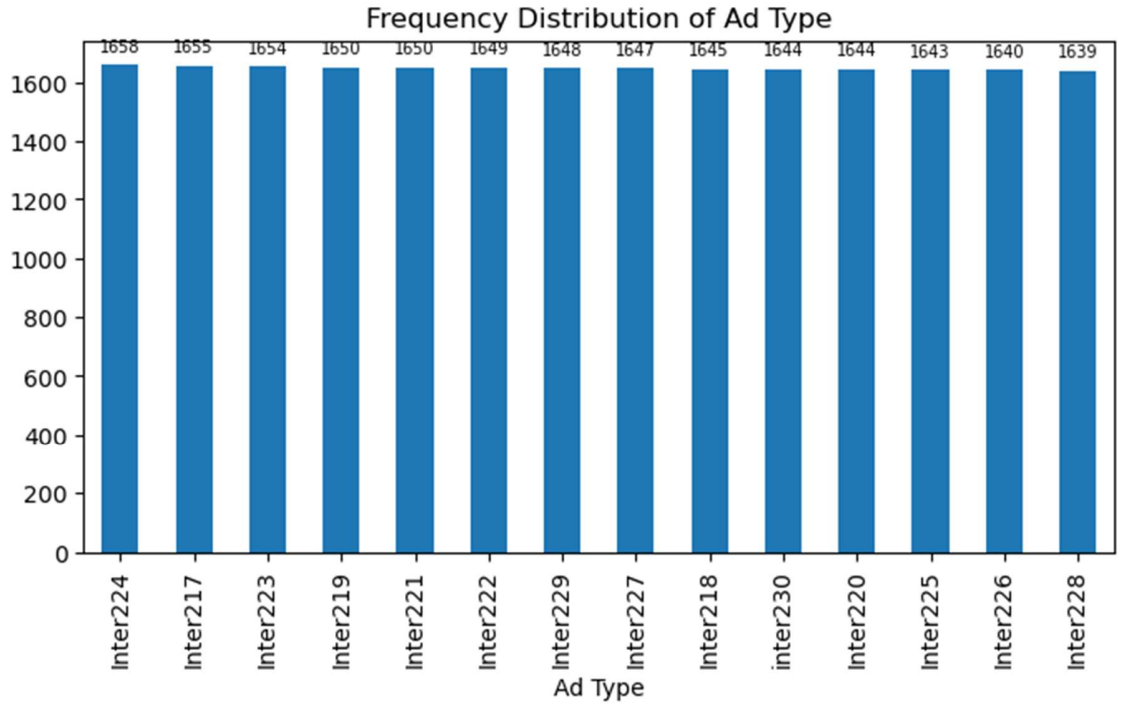
1. Outliers are present in 11 out of the 14 numeric columns. Given that clustering is sensitive to outliers, particularly for the k-means method, as it relies on centroids to form groups, and these centroids can be substantially influenced by outliers, we will address these outliers.
2. There is skewness in the data with except for columns ad-length and ads_hour all the columns are skewed. Ad-length and Fee are left skewed while the others are right skewed.
3. While for both ad length and ad width, the median values are almost similar. However, for ad length, the box covers almost the entire range but in case ad width data is left skewed this deviation might be due to the device type on which ads were played.
4. From the statistical summary and univariate analysis, we have observed that the ratio of 'Matched_Queries' that is number of ads run where the keywords used in these ads matched the keywords entered in the search engine and this resulting in a click on the ad against 'Available_Impressions' that is total available advertising slots available on network is very almost 1:2 that is for every 2 'Available Impressions' is resulting in a click or an instance of 'Matched_Queries'. However, this ratio in case of Ads 24x7 is 1:10 that for every 10 impressions Ads 24x7 gets only 1 click.

For categorical columns

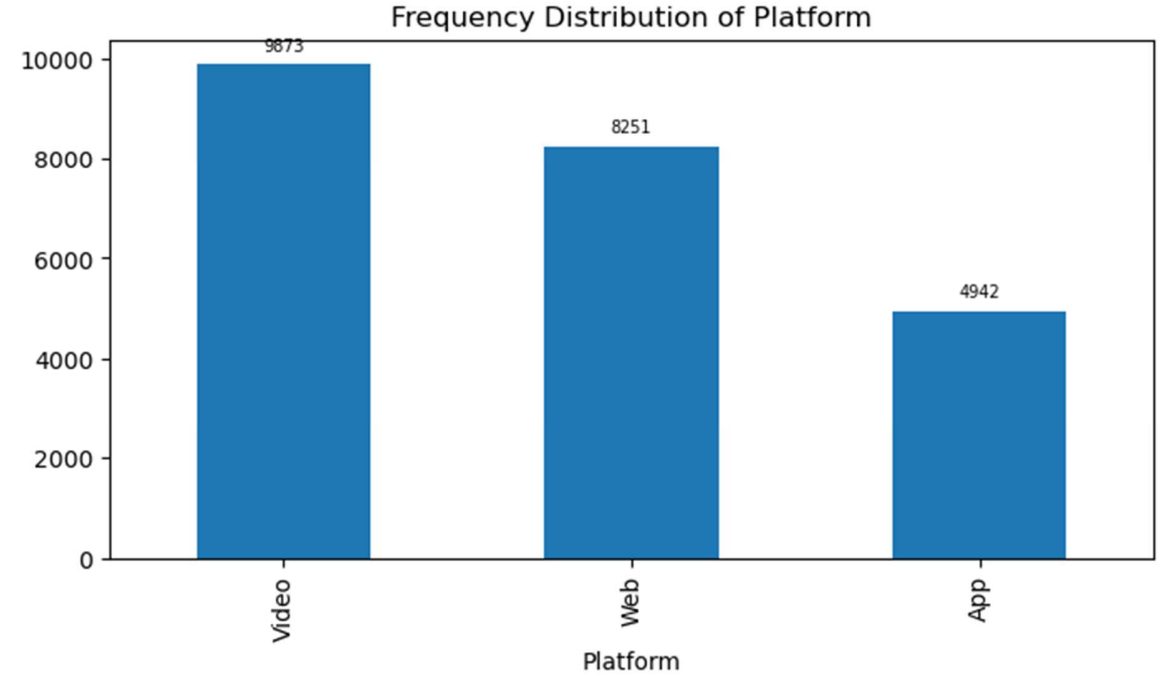
```
Details of InventoryType
------------------------------------------------------------------
```



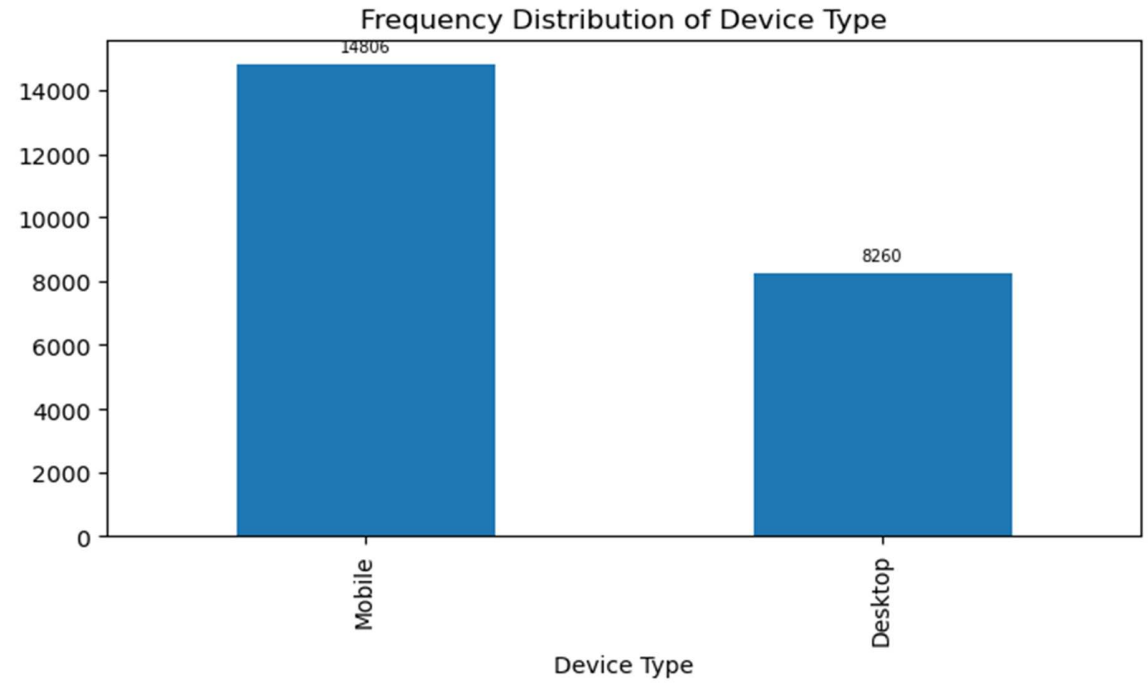
```
Details of Ad Type
-----------------------------------------------------------------
```




```
Details of Platform
-----------------------------------------------------------------
```



Details of Device Type



Details of Format

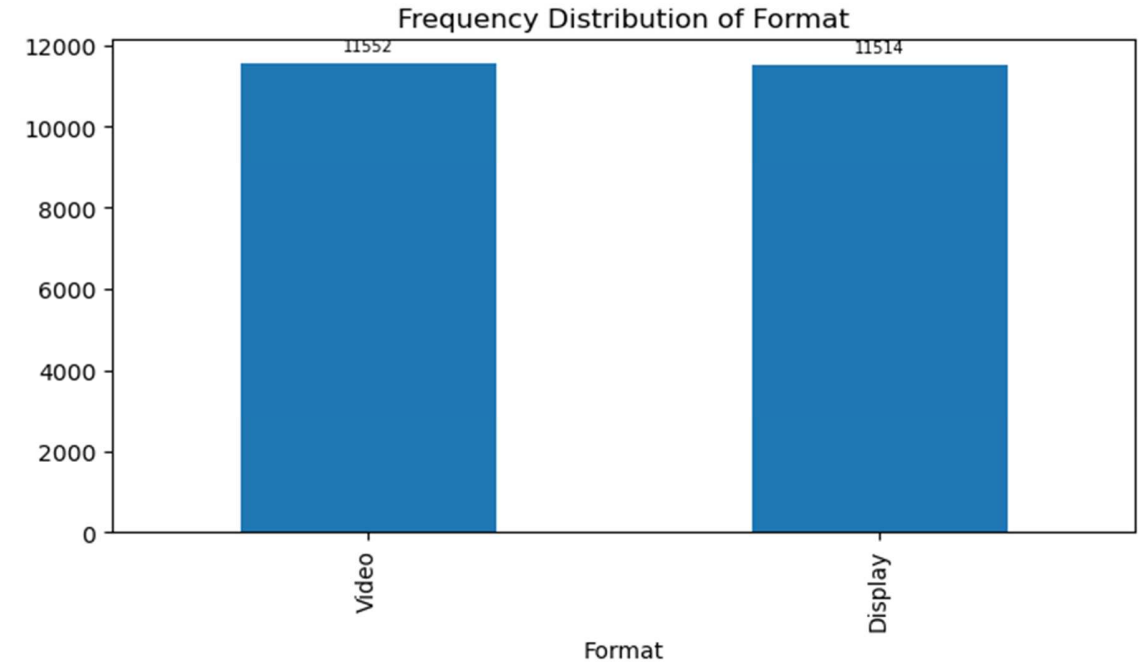


Figure 2: Univariate Analysis categorical columns

Key Observations

1. Data distribution for 'Ad Type' and 'Format' is uniform.
2. While ads run Mobiles are significantly more than those run on Desktops, however, Video and Web ads are displayed in far greater numbers than those on Apps.
3. Format 4 is the most preferred 'Inventory Type' while Format 7 is the least preferred.

1.6.2 Bivariate Analysis

Relation between numeric columns

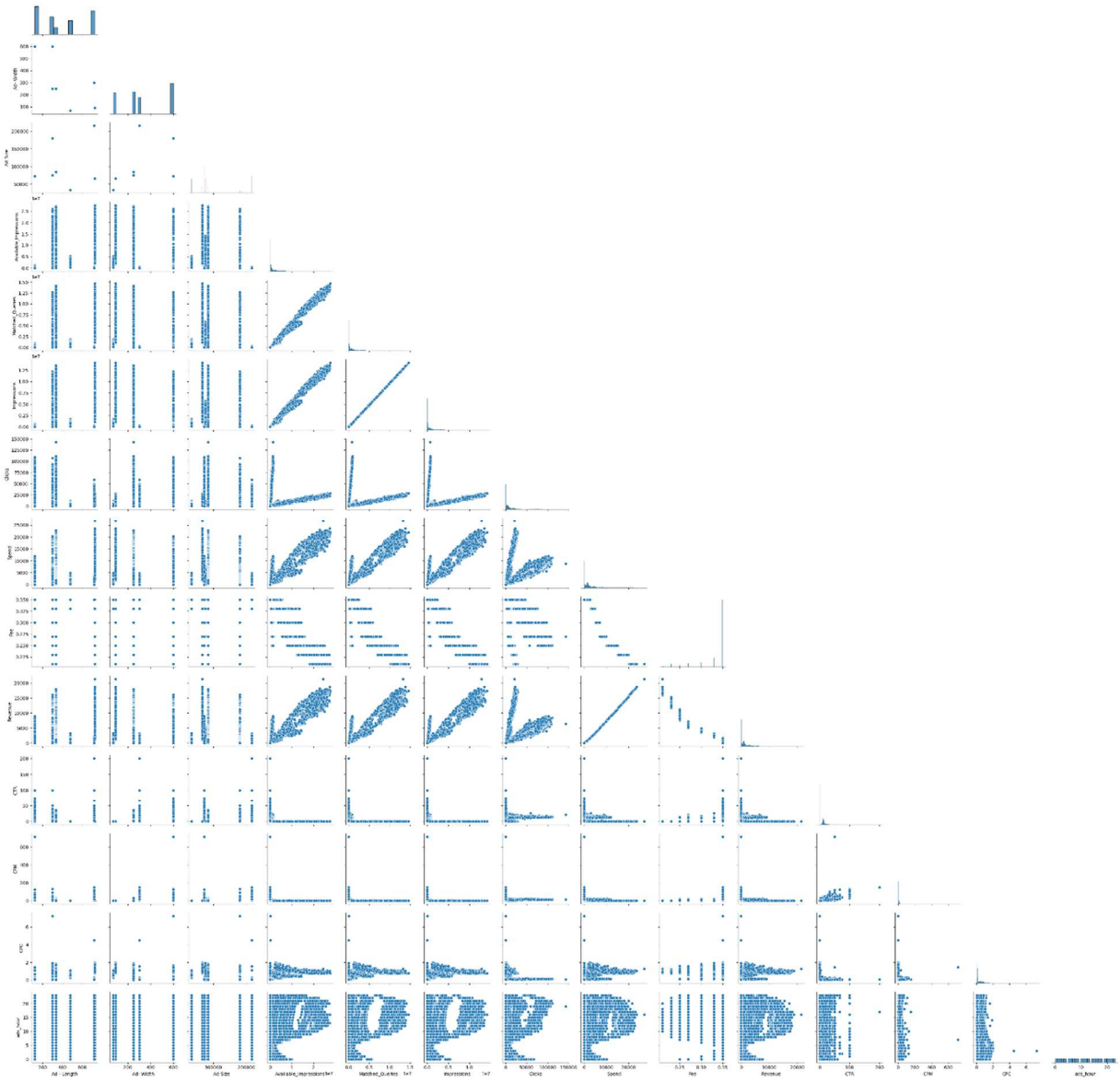


Figure 3: Pair plot

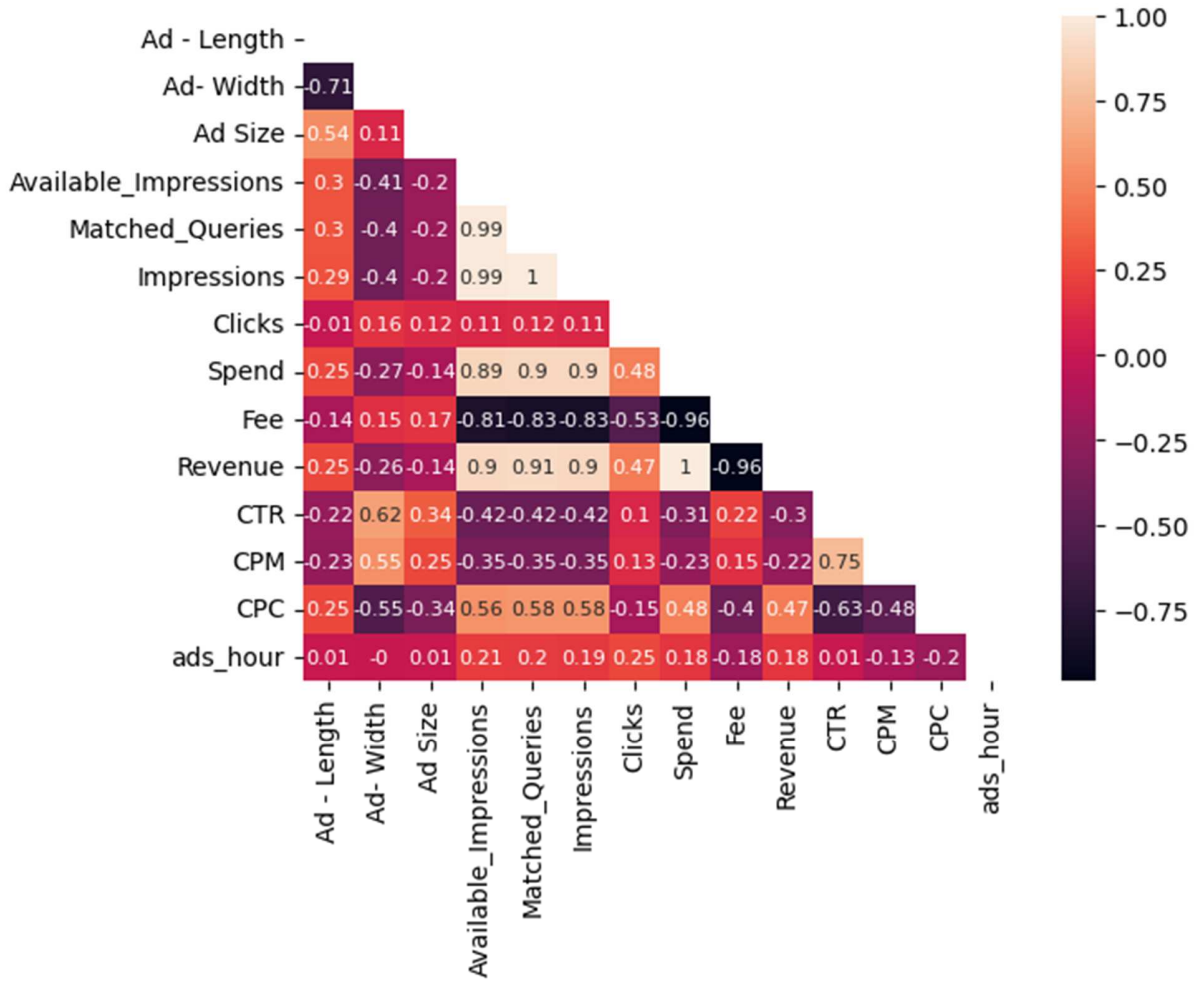


Figure 4: Heatmap

Key Observations

1. There is strong positive as well as negative correlation in the data.
2. Ad Length and Ad Width have strong negative correlation and this might be due to the fact that ads are run on Mobile and Desktop where for mobiles ads might be longer while in desktop the ads will be wider due to device screen dimensions.
3. Ad Width has a strong positive correlation with CTR and CPM, strong negative correlation with CPC. There is some relation that wider ads receive more clicks and due to this CPC goes down.
4. Columns Matched_Queries, Impressions, Spend, Fee and Revenue are showing very strong correlation amongst each other.

Relation between numeric and categorical columns

Bivariate analysis for InventoryType

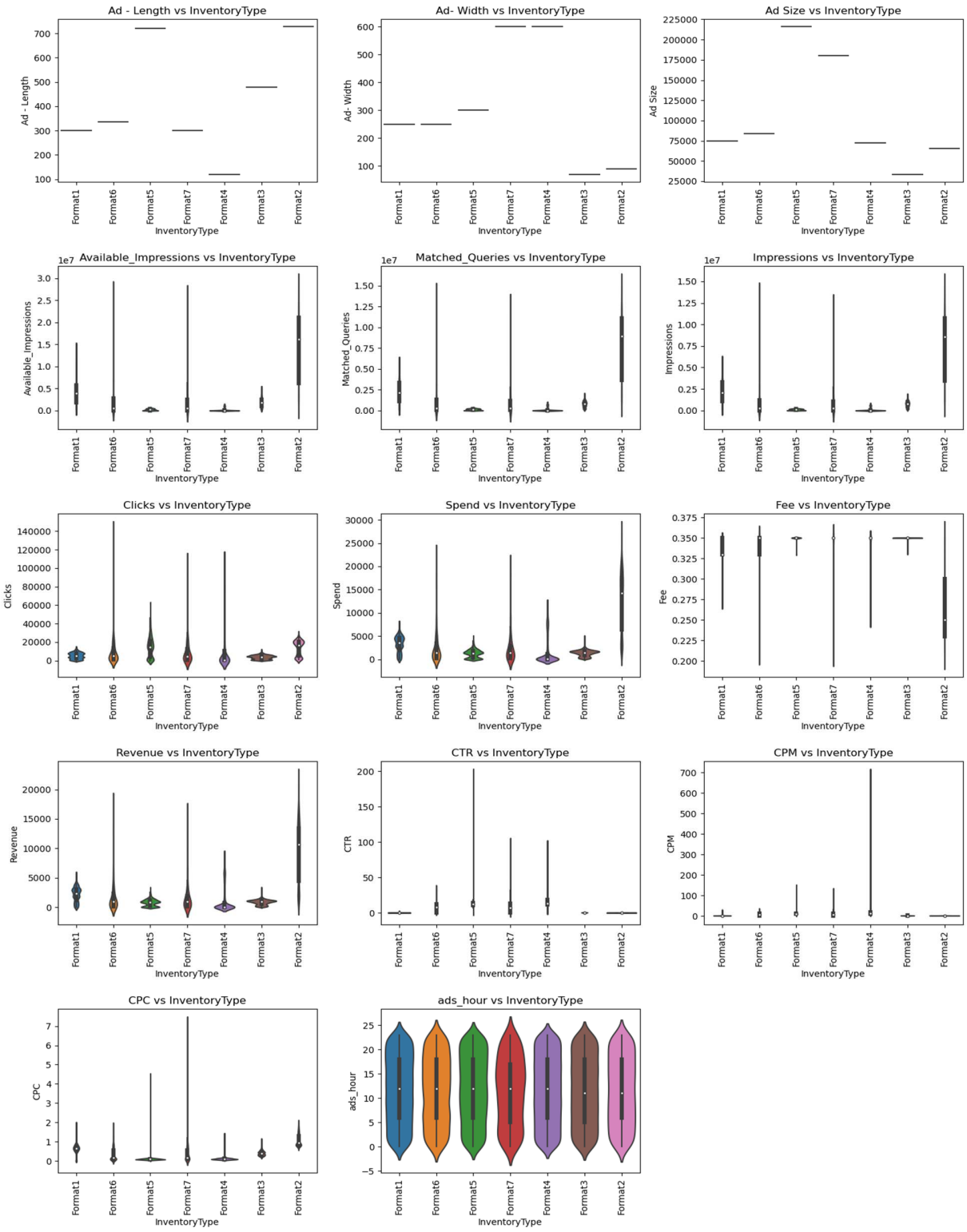


Figure 5: Bivariate analysis InventoryType

Bivariate analysis for Ad Type

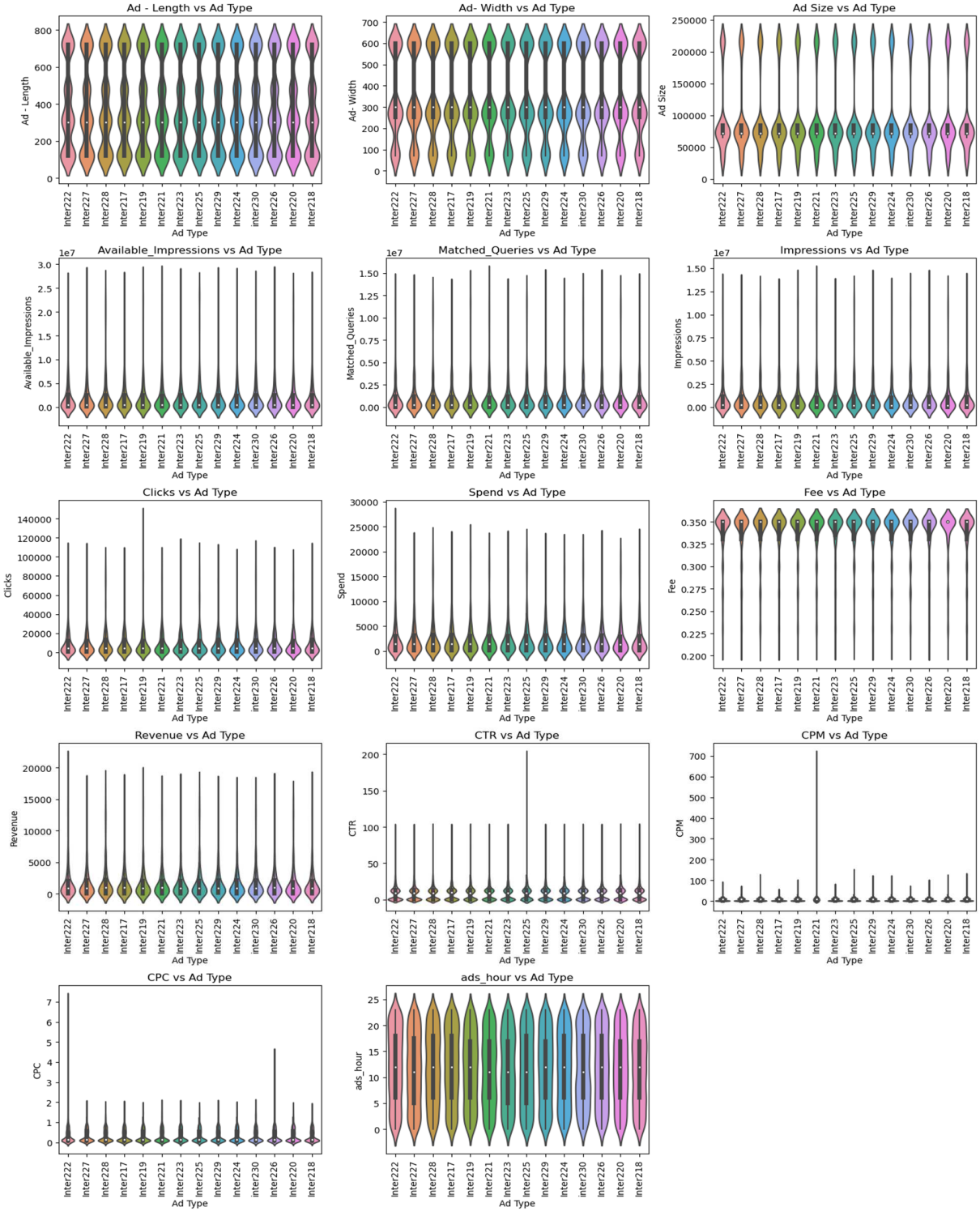


Figure 6: Bivariate analysis for Ad Type

Bivariate analysis for Platform

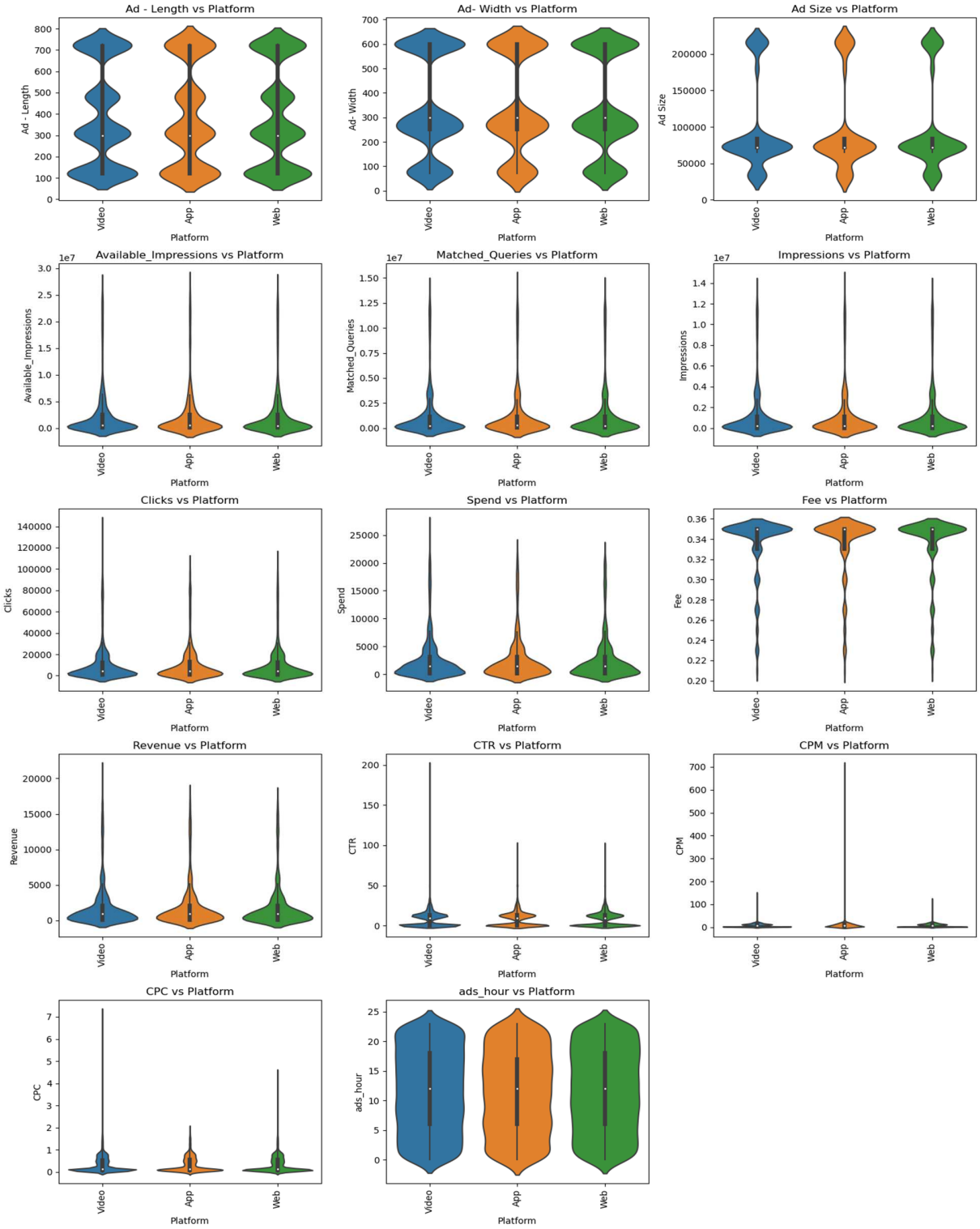


Figure 7: Bivariate analysis for Platform

Bivariate analysis for Device Type

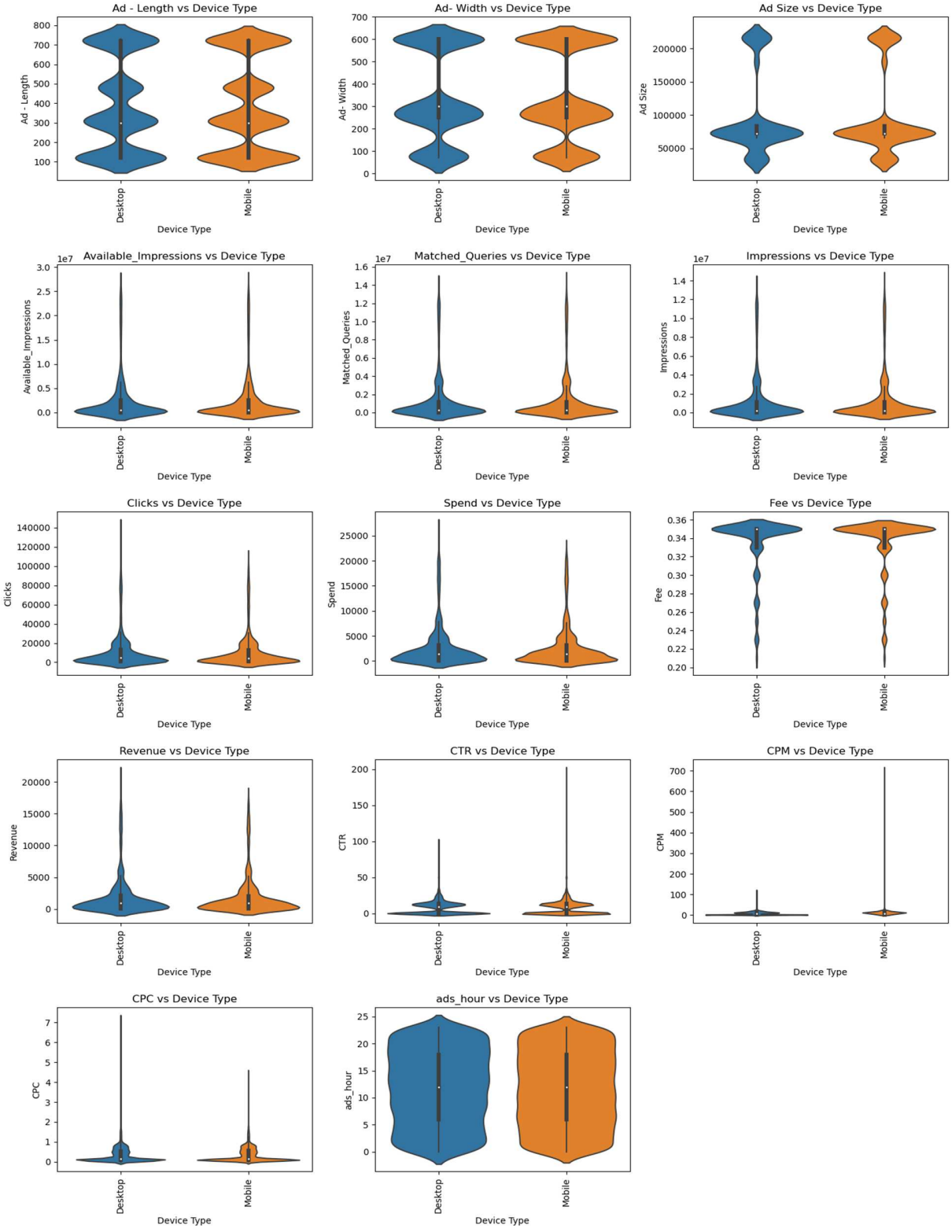


Figure 8: Bivariate analysis for Device Type

Bivariate analysis for Format

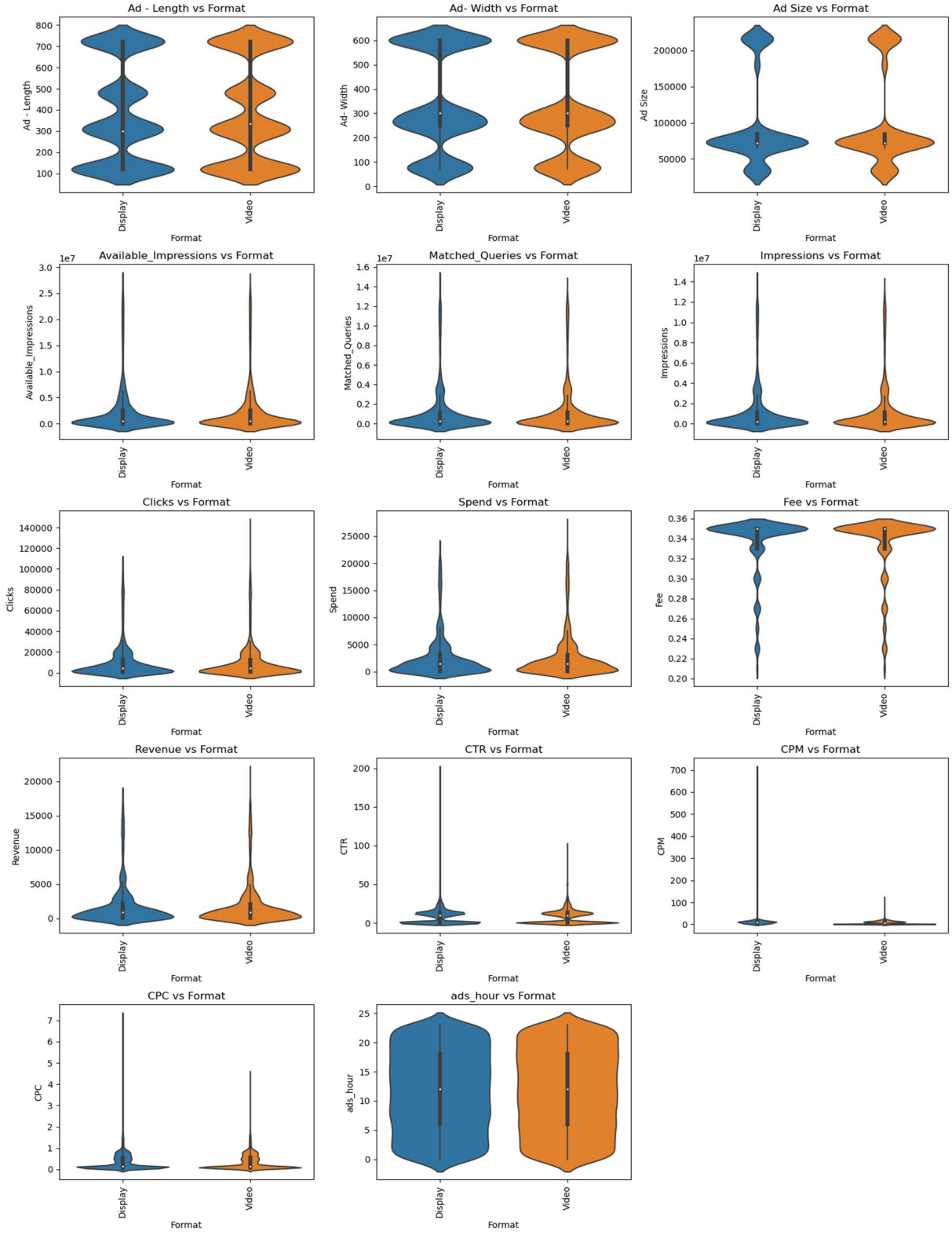


Figure 9: Bivariate analysis for Format

Key Observations

1. Each format in Inventory Type has a unique dimension as well as available impressions and these might have been designed keeping in mind the available impressions on the network.
2. Ad-Length for videos is longer than display ads.

Relationship between categorical columns

Cross-tabulation between InventoryType and Ad Type

Ad Type	Inter217	Inter218	Inter219	Inter220	Inter221	Inter222	Inter223	Inter224	Inter225	Inter226	Inter227	Inter228	Inter229	inter230
InventoryType														
Format1	273	276	276	265	270	268	274	278	268	269	278	269	279	271
Format2	126	127	127	127	127	127	127	127	127	127	132	129	129	130
Format3	253	253	253	253	253	253	253	253	253	253	253	253	252	252
Format4	511	510	514	513	512	514	513	513	511	511	510	511	509	513
Format5	308	304	301	305	309	304	307	306	300	299	298	302	304	302
Format6	136	127	130	135	134	135	132	133	135	134	131	130	129	129
Format7	48	48	49	46	45	48	48	48	49	47	45	45	46	47

Cross-tabulation between InventoryType and Platform

Platform	App	Video	Web
InventoryType			
Format1	824	1631	1359
Format2	386	769	634
Format3	759	1516	1265
Format4	1532	3072	2561
Format5	911	1813	1525
Format6	392	788	670
Format7	138	284	237

Cross-tabulation between InventoryType and Device Type

Device Type	Desktop	Mobile
InventoryType		
Format1	1374	2440
Format2	636	1153
Format3	1264	2276
Format4	2561	4604
Format5	1523	2726
Format6	663	1187
Format7	239	420

Cross-tabulation between InventoryType and Format

Format	Display	Video
InventoryType		
Format1	1926	1888
Format2	896	893
Format3	1746	1794
Format4	3629	3536
Format5	2081	2168
Format6	905	945
Format7	331	328

Cross-tabulation between Ad Type and Platform

Platform	App	Video	Web
Ad Type			
Inter217	0	0	1655
Inter218	1645	0	0
Inter219	0	1650	0
Inter220	0	0	1644
Inter221	1650	0	0
Inter222	0	1649	0
Inter223	0	0	1654
Inter224	0	0	1658
Inter225	0	1643	0
Inter226	0	0	1640
Inter227	1647	0	0
Inter228	0	1639	0
Inter229	0	1648	0
inter230	0	1644	0

Cross-tabulation between Ad Type and Device Type

Device Type	Desktop	Mobile
Ad Type		
Inter217	1655	0
Inter218	0	1645
Inter219	1650	0
Inter220	0	1644
Inter221	0	1650
Inter222	1649	0
Inter223	0	1654
Inter224	1658	0
Inter225	0	1643
Inter226	0	1640
Inter227	0	1647
Inter228	0	1639
Inter229	1648	0
inter230	0	1644

Cross-tabulation between Ad Type and Format

Format	Display	Video
Ad Type		
Inter217	826	829
Inter218	836	809
Inter219	831	819
Inter220	793	851
Inter221	871	779
Inter222	808	841
Inter223	794	860
Inter224	837	821
Inter225	813	830
Inter226	818	822
Inter227	842	805
Inter228	836	803
Inter229	812	836
inter230	797	847

Cross-tabulation between Platform and Device Type

Device Type	Desktop	Mobile
Platform		
App	0	4942
Video	4947	4926
Web	3313	4938

Cross-tabulation between Platform and Format

Platform	Display	Video
App	2549	2393
Video	4897	4976
Web	4068	4183

Cross-tabulation between Device Type and Format

Device Type	Display	Video
Desktop	4114	4146
Mobile	7400	7406

Table8: Cross tabs of categorical columns

Key Observations

1. There are different ad types specifically for different platforms and devices.
2. Ads on app is only done for mobiles.

1.7 Outlier Treatment

For outlier treatment of numeric columns, we will be using Winsorization method where we will replace the extreme values lying beyond 1.5 times the Inter Quartile Range (IQR) from 75 percentile mark and below 25 percentile mark with 95 percentile value and 5 percentile value respectively.

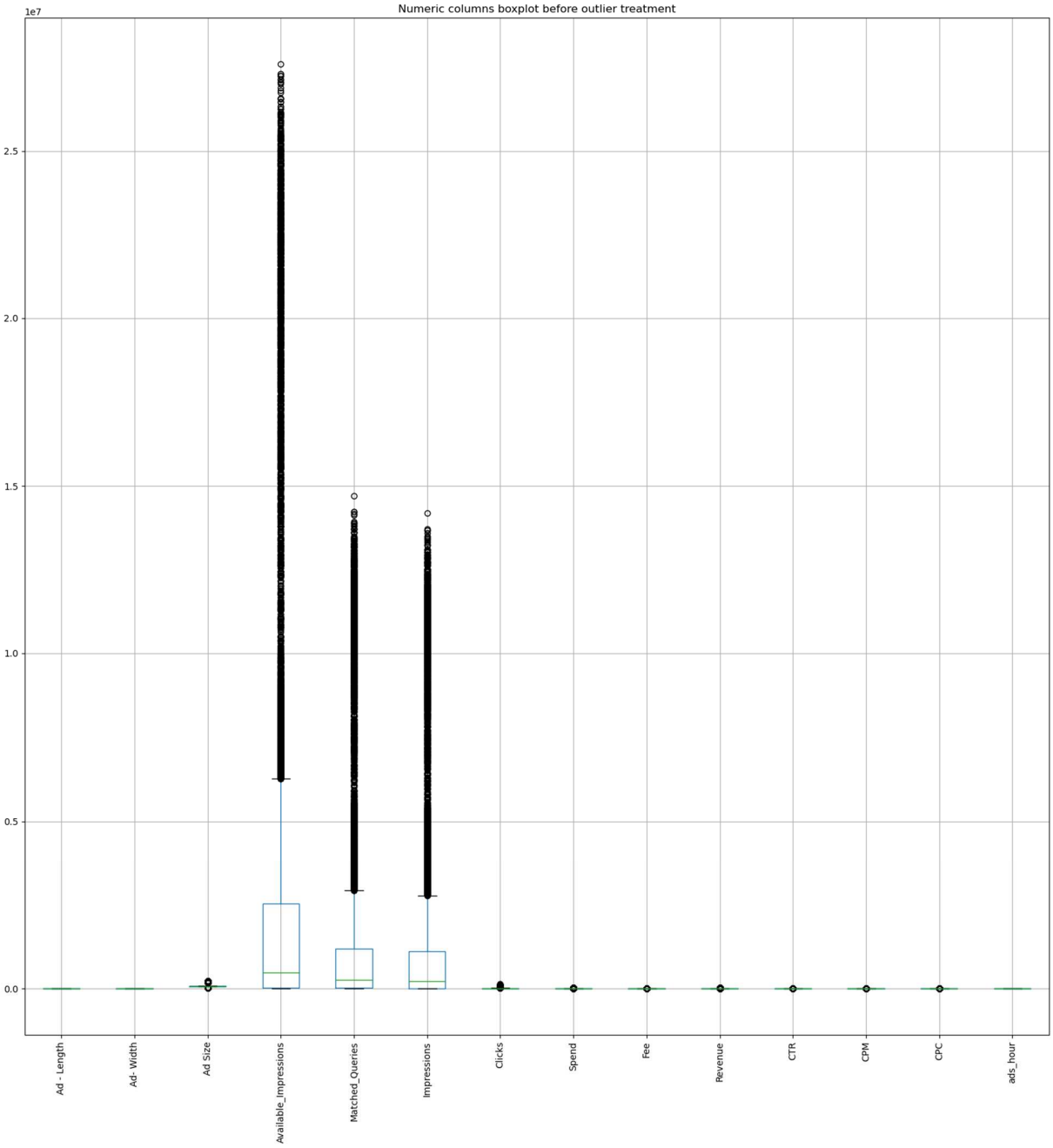


Figure 10: Boxplot numeric columns

As per the original data the range for features like ‘Available_impressions’ go well over 25000000.

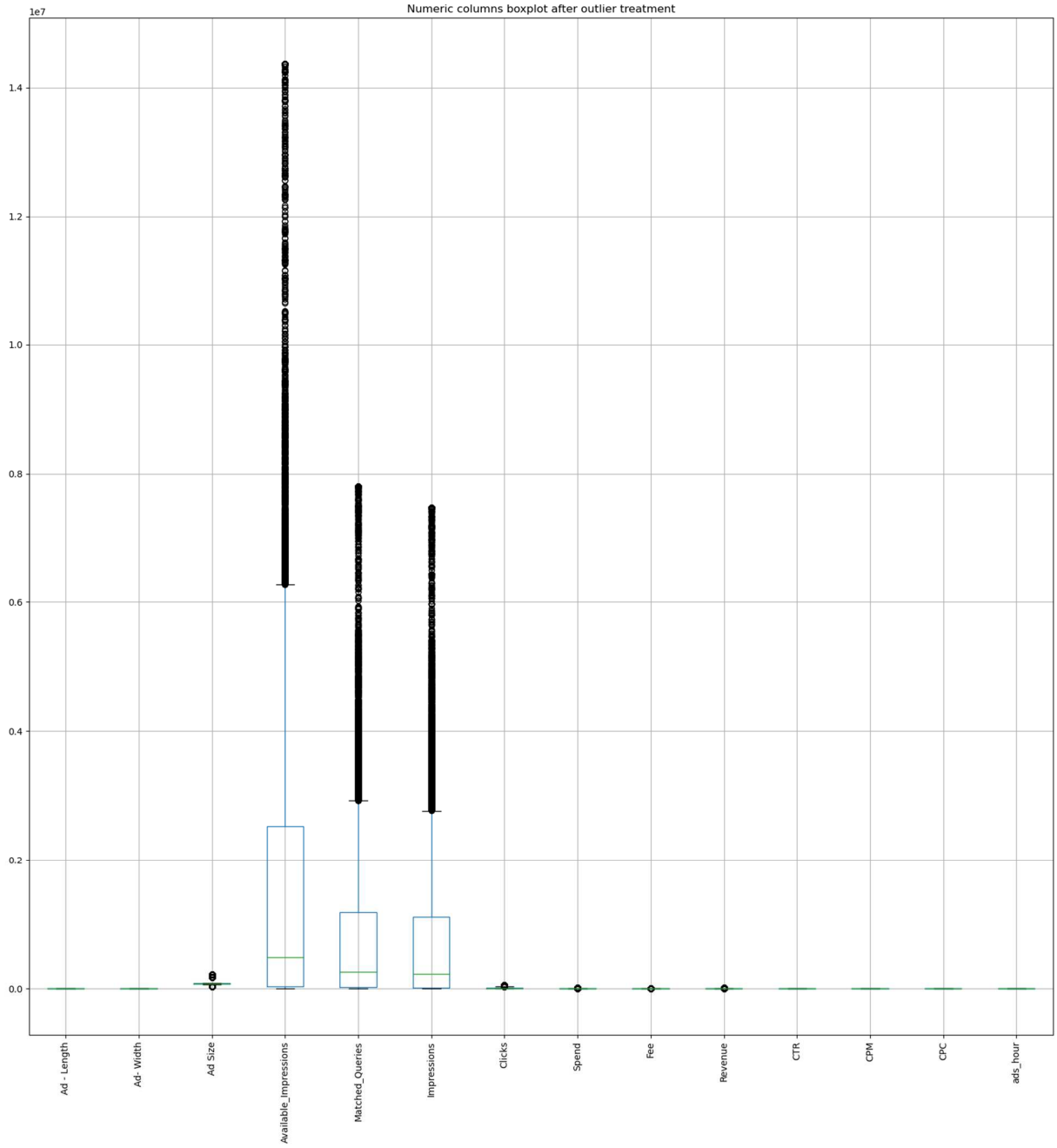


Figure 11: Boxplot outlier treated numeric columns

The outlier treatment has helped bring down the range for features like for ‘Available_impressions it has come down to below 15000000, similarly for ‘Matched_Queries’ and ‘Impressions’ it has come down from just under 15000000 to below 8000000.

1.8 Data Scaling

For scaling we will use Z score scaling since data in features are of different scale and Z score method will help reduce this difference bring data on almost similar scales where mean for all the features is almost 0 and standard deviation is 1, ensuring that all features contribute equally to clustering.

Also, since clustering algorithms use distance-based method to measure similarity and dissimilarity between data points, scaling reduces the magnitude of variability between the data point helping in faster convergences, reducing the time taken by the algorithm in clustering drastically.

Statistical Summary of scaled data

	count	mean	std	min	25%	50%	75%	max
Ad - Length	23066.0	1.281478e-16	1.000022	-1.134891	-1.134891	-0.364496	1.433093	1.467332
Ad- Width	23066.0	-1.182903e-16	1.000022	-1.319110	-0.432797	-0.186599	1.290590	1.290590
Ad Size	23066.0	2.464381e-17	1.000022	-1.024985	-0.400970	-0.400970	-0.205965	1.939086
Available_Impressions	23066.0	0.000000e+00	1.000022	-0.593128	-0.583891	-0.458606	0.110324	3.404928
Matched_Queries	23066.0	1.971505e-17	1.000022	-0.586173	-0.576910	-0.454345	0.017206	3.402121
Impressions	23066.0	-3.943010e-17	1.000022	-0.581070	-0.576915	-0.461761	0.008361	3.379223
Clicks	23066.0	3.943010e-17	1.000022	-0.737121	-0.682799	-0.393262	0.258973	3.210313
Spend	23066.0	0.000000e+00	1.000022	-0.754487	-0.728988	-0.322959	0.191044	3.154074
Fee	23066.0	0.000000e+00	1.000022	-2.973434	-0.209252	0.481794	0.481794	0.481794
Revenue	23066.0	-3.943010e-17	1.000022	-0.712603	-0.690262	-0.334496	0.141374	3.239009
CTR	23066.0	8.871773e-17	1.000022	-1.015889	-1.005333	0.182298	0.713161	2.055065
CPM	23066.0	2.217943e-16	1.000022	-1.067315	-0.980968	0.050671	0.778227	1.921144
CPC	23066.0	2.168656e-16	1.000022	-0.908581	-0.795531	-0.624259	0.780447	2.089664
ads_hour	23066.0	1.035040e-16	1.000022	-1.563029	-0.825671	0.059159	0.943988	1.533875

Table 9: Statistical Summary of scaled data

From the statistical summary we can clearly infer that for all the features mean is very small, nearly zero and standard deviation is almost 1 and range of data is between -3 to +3.5.

1.9 Clustering

Hierarchical Clustering

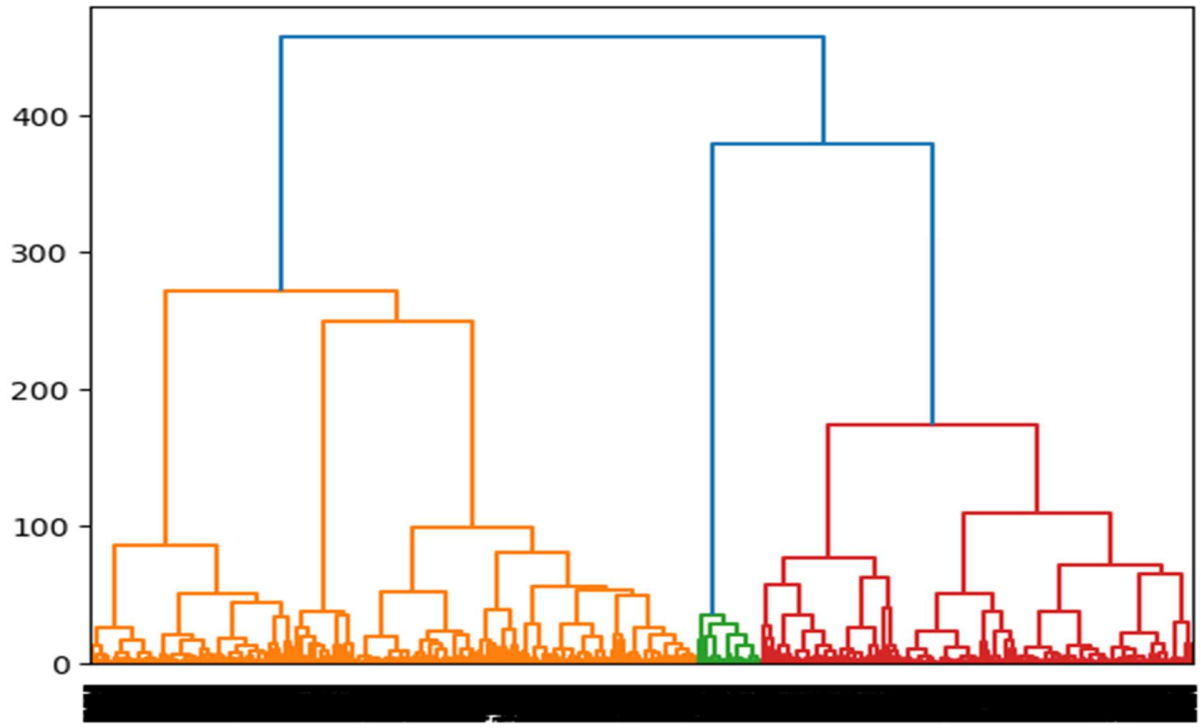


Figure 12: Dendrogram

The above dendrogram is very cluttered due to the hierarchical clustering process, however, it clearly highlights three clusters by 3 different colours. From it we can conclude that as per hierarchical clustering the optimum number of clusters are 3. Below we will try to understand better by reducing this cluttering.

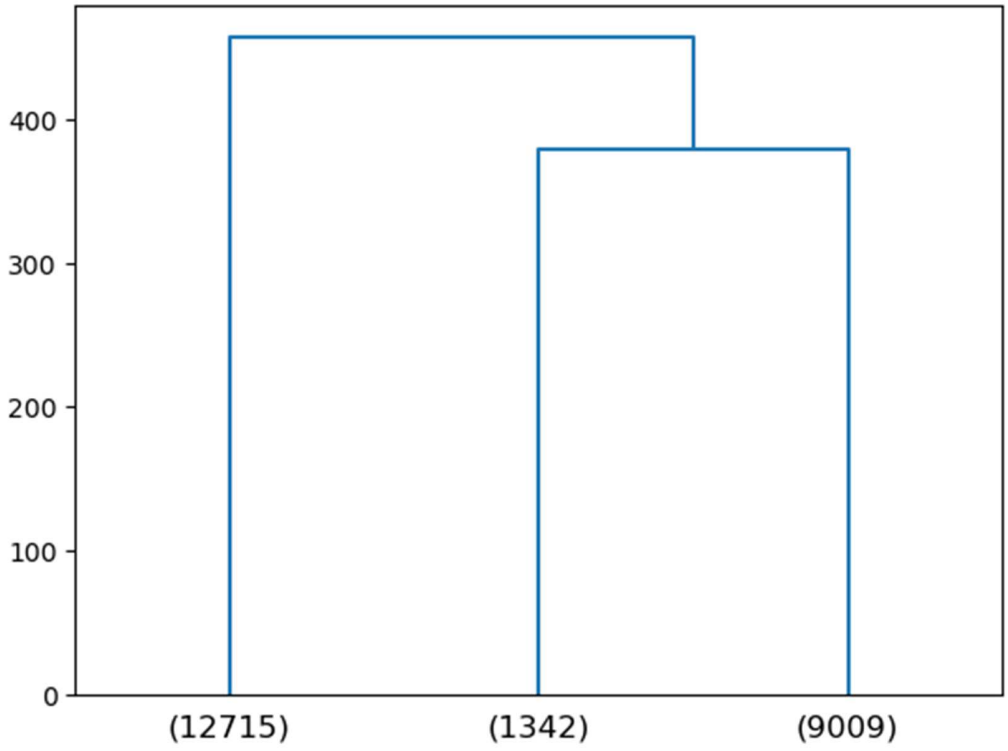


Figure 13: Dendrogram

Here we have cut the dendrogram into taking only 3 clusters as we knew that as per hierarchical clustering the optimum number of clusters is 3 where each horizontal line going from the x-axis represents a cluster and x-ticks represents the number of instances recorded for each cluster.

As per the above dendrogram, cluster 1 accounts for 12715 instances covering over 50% of the observations followed by cluster 3 with 9009 instances and cluster 2 account for only 1342 instances. To confirm that we have got correct number of clusters using hierarchical clustering, we will do K-means clustering and try to confirm the optimum number of clusters using Silhouette scores.

K-Means Clustering

In K-Means we have to pre-define the number of clusters, since, it is impossible for us to know the correct optimum value so we take a range of numbers as clusters and using the K-elbow method find the optimum number of clusters

K-elbow Method

In the K-elbow method, we explored the range of 1 to 10 for the number of clusters and plotted the corresponding K-mean values. In the plot below where an elbow-shaped pattern is formed for certain K

values. The K value associated with this elbow point is considered the optimal number of clusters.

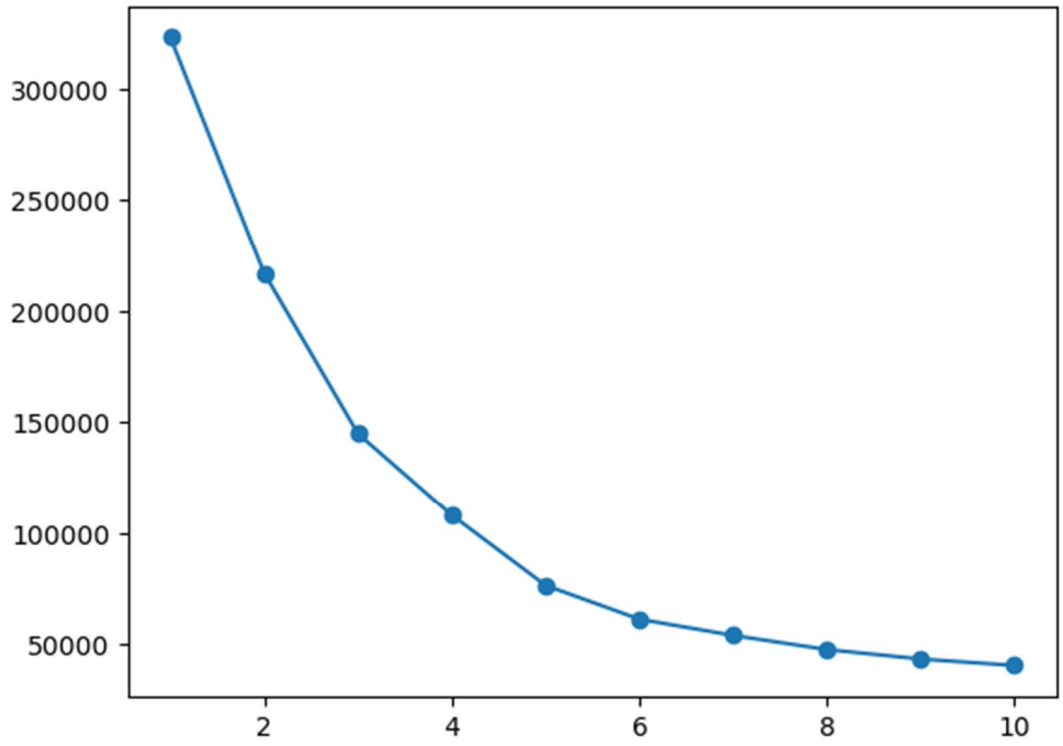


Figure 14: K-Means elbow plot

The elbow chart above does not have a clear breakout in the elbow after K = 3 as there is there is consistent drop in the within sum of squares up to K = 6 before we can see any signs of the plot getting flattening out. So, to find the optimum number of clusters we will calculate silhouette score for all values of K taken above.

Silhouette Score

In Silhouette score technique we calculate Silhouette score (Sil score) for all the values of K taken above where Sil score values could range between -1 and +1. Here, -1 would mean that entire clustering is inappropriately done, 0 would mean that it is difficult to perfectly separate the data point and +1 would means that data points are segregated appropriately, clusters could be clearly distinguished. As per the Sil score technique the K value for which Sil score is highest is the most optimum number of clusters.

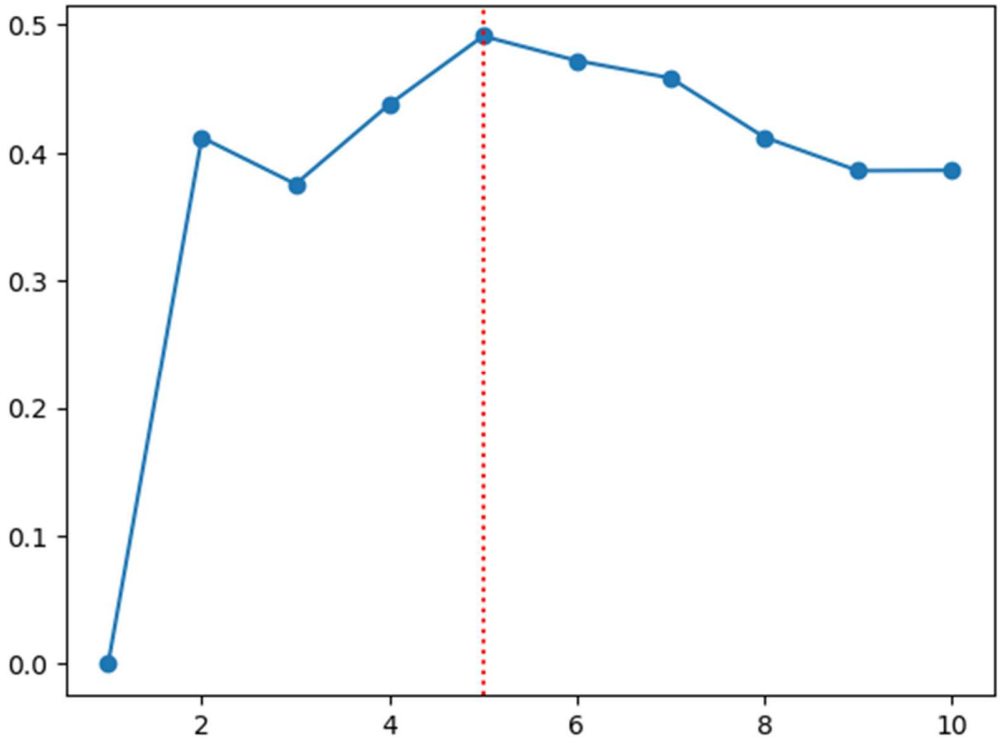


Figure 15: Silhouette Score plot

In the above plot red dotted line is used to mark the K value for which Silhouette score is highest at 0.48, K = 5 is considered the optimum number of clusters for this data, as far as K = 3 is concerned which we got from hierarchical clustering it has the second lowest value at 0.37 so we reject hierarchical clustering finding of 3 clusters and will continue with K = 5.

Cluster Profiling

Based on the K-means clustering for K = 5, each cluster has following number of instances in the data.

```
Clus_kmeans5
0    1529
1    6977
2    8820
3    4294
4    1446
Name: count, dtype: int64
```

Table 10: Clusters value counts

Clus_kmeans5	0	1	2	3	4
Ad - Length	6.754506e+02	152.780278	3.976939e+02	715.346996	142.531120
Ad- Width	1.199804e+02	558.212699	1.801995e+02	302.945971	570.954357
Ad Size	7.028947e+04	77627.633653	6.190121e+04	215485.794131	75759.336100
Available_Impressions	1.339262e+07	48493.492081	2.956154e+06	246421.787087	840171.217842
Matched_Queries	7.320533e+06	29658.611115	1.545106e+06	134714.179145	588656.074689
Impressions	7.015764e+06	22108.002508	1.497152e+06	114349.264264	496622.592669
Clicks	1.748777e+04	3047.977641	4.462684e+03	14162.820913	48599.696404
Spend	1.209195e+04	326.782750	2.382393e+03	1223.867068	7205.564668
Fee	2.537083e-01	0.349693	3.412324e-01	0.349544	0.285705
Revenue	9.039052e+03	213.485554	1.590548e+03	797.055839	5182.665433
CTR	1.952174e-01	15.044317	3.598913e-01	13.032602	13.767551
CPM	1.697421e+00	13.838584	1.748668e+00	11.485962	15.010082
CPC	8.489766e-01	0.102967	5.468139e-01	0.090070	0.109943
ads_hour	1.435971e+01	10.936936	1.109649e+01	11.821379	14.276625
freq	1.529000e+03	6977.000000	8.820000e+03	4294.000000	1446.000000

Table 11: Mean value of numeric columns based on clusters

The above tables show average values for all numeric columns based on clusters.

Key Observations

1. All the clusters have different length and width. Ad size for cluster 3 is largest and for cluster 2 it is smallest.
2. Cluster 0 and 2 have most available impressions due to which they have highest impression count, since, cluster 1 had least number of Available_impressions, it got least number of impressions.
3. In the key metrics of CTR and CPC, Cluster 1 has the best CTR of 15 while Cluster 0 and 2 have very bad CTR values. Since, CTR and CPC are highly correlated so CPC also follows the similar trend.
4. Cluster 0 and 2 have lowest CPM values, this might be due to high number of impressions while cluster 4 has highest CPM value despite having good number of impressions.

Cluster Naming

Based on the cluster's properties, especially their size dimensions we can name them as following:

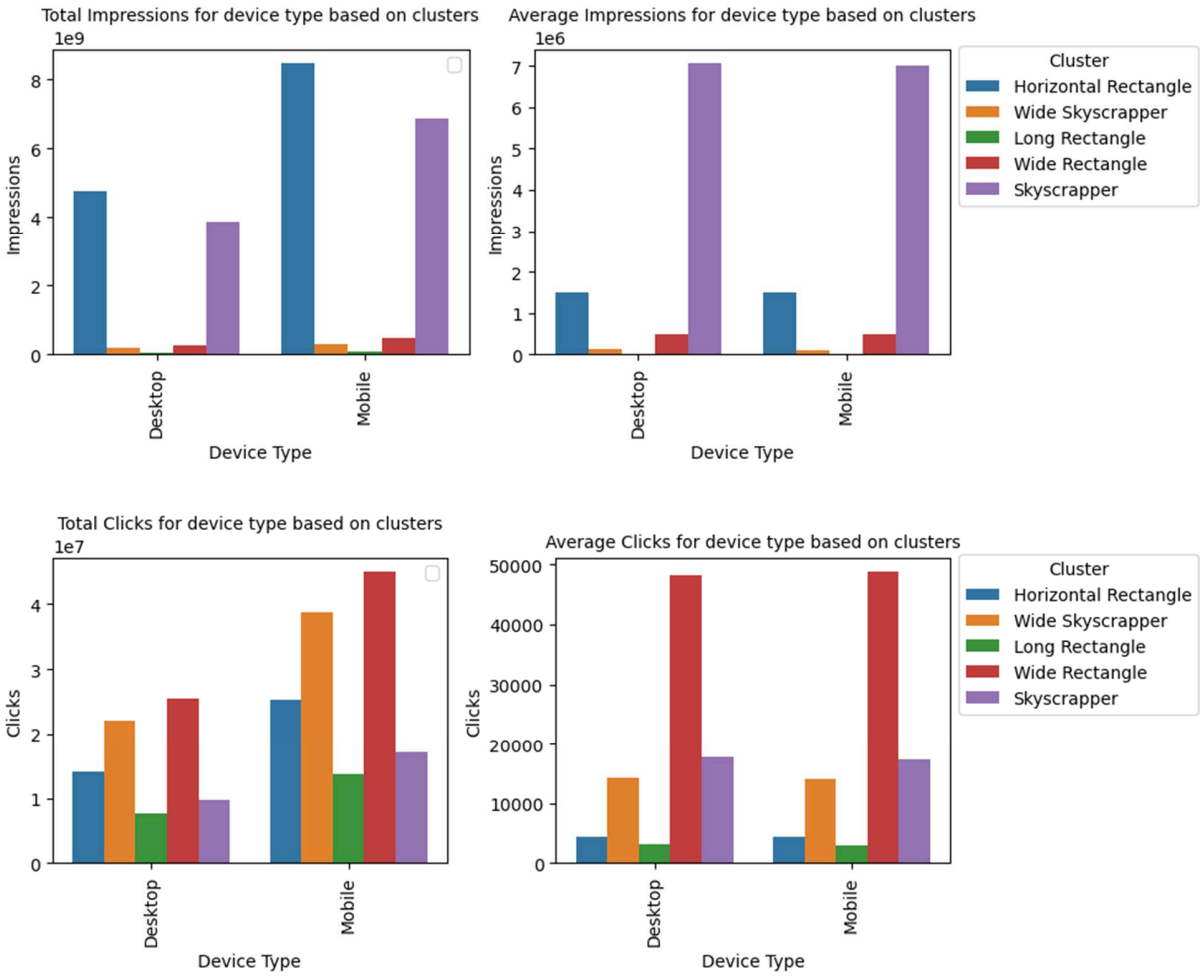
- Cluster 0: Skyscraper being tall and sleek
- Cluster 1: Long Rectangle being rectangle with dimensions 153x558
- Cluster 2: Horizontal Rectangle being rectangle whose length is more than its width

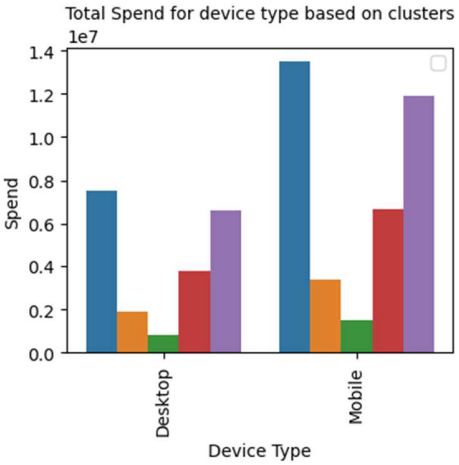
- Cluster 3: Wide Skyscraper being wider than skyscraper
- Cluster 4: Wide Rectangle having slightly lower length and higher width than long rectangle

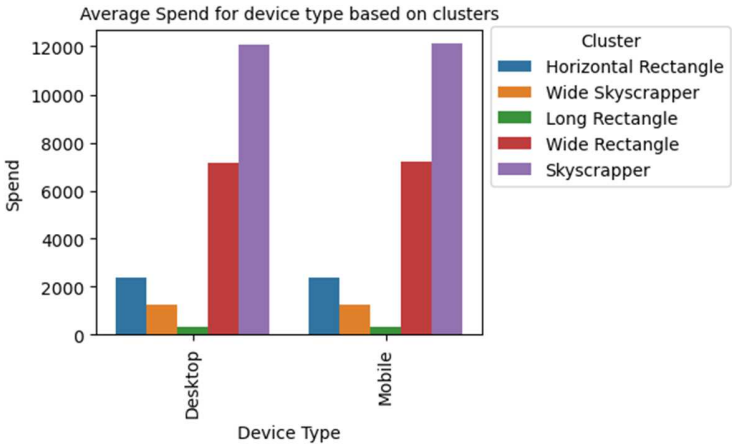
We replaced the numeric values of the clusters with the above names given for further analysis.

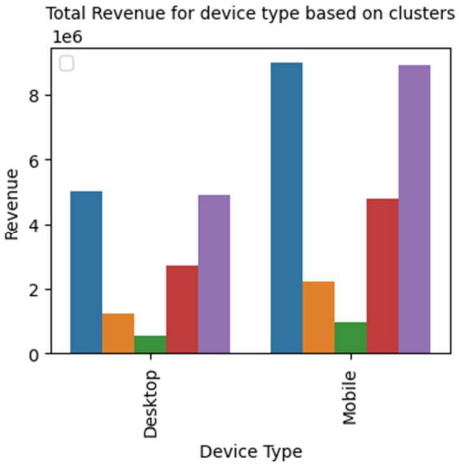
1.10 Data Analysis based on Clusters

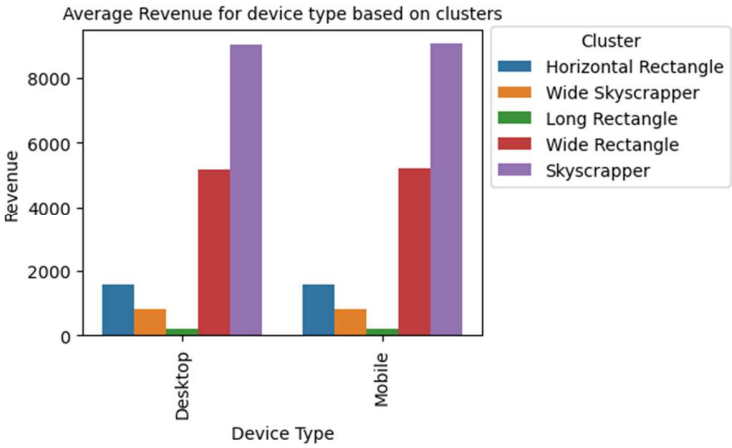
We visually analyze the different numeric fields like Impressions, Clicks, Spend, Revenue etc. against different devices on which ads were showcased for different clusters.

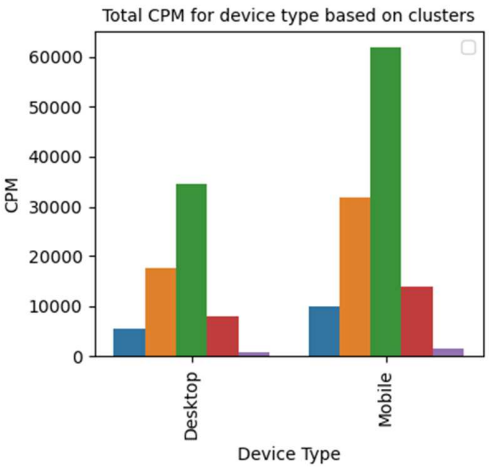


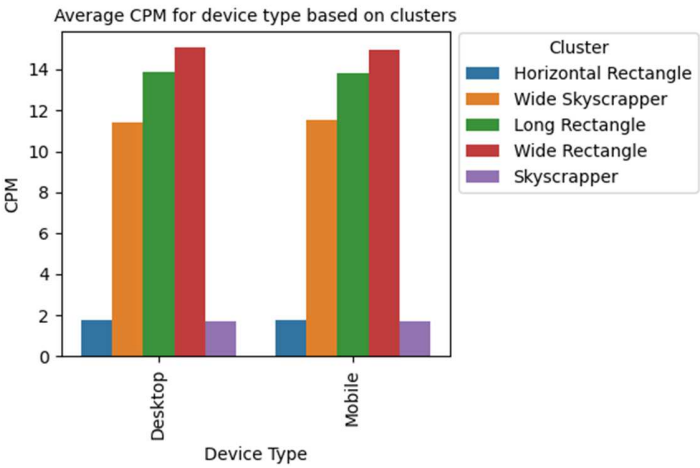

Total Spend for device type based on clusters
1e7
Spend
Device Type
Desktop
Mobile


Average Spend for device type based on clusters
Spend
Device Type
Desktop
Mobile
Cluster
Horizontal Rectangle
Wide Skyscrapper
Long Rectangle
Wide Rectangle
Skyscrapper


Total Revenue for device type based on clusters
1e6
Revenue
Device Type
Desktop
Mobile


Average Revenue for device type based on clusters
Revenue
Device Type
Desktop
Mobile
Cluster
Horizontal Rectangle
Wide Skyscrapper
Long Rectangle
Wide Rectangle
Skyscrapper


Total CPM for device type based on clusters
CPM
Device Type
Desktop
Mobile


Average CPM for device type based on clusters
CPM
Device Type
Desktop
Mobile
Cluster
Horizontal Rectangle
Wide Skyscrapper
Long Rectangle
Wide Rectangle
Skyscrapper

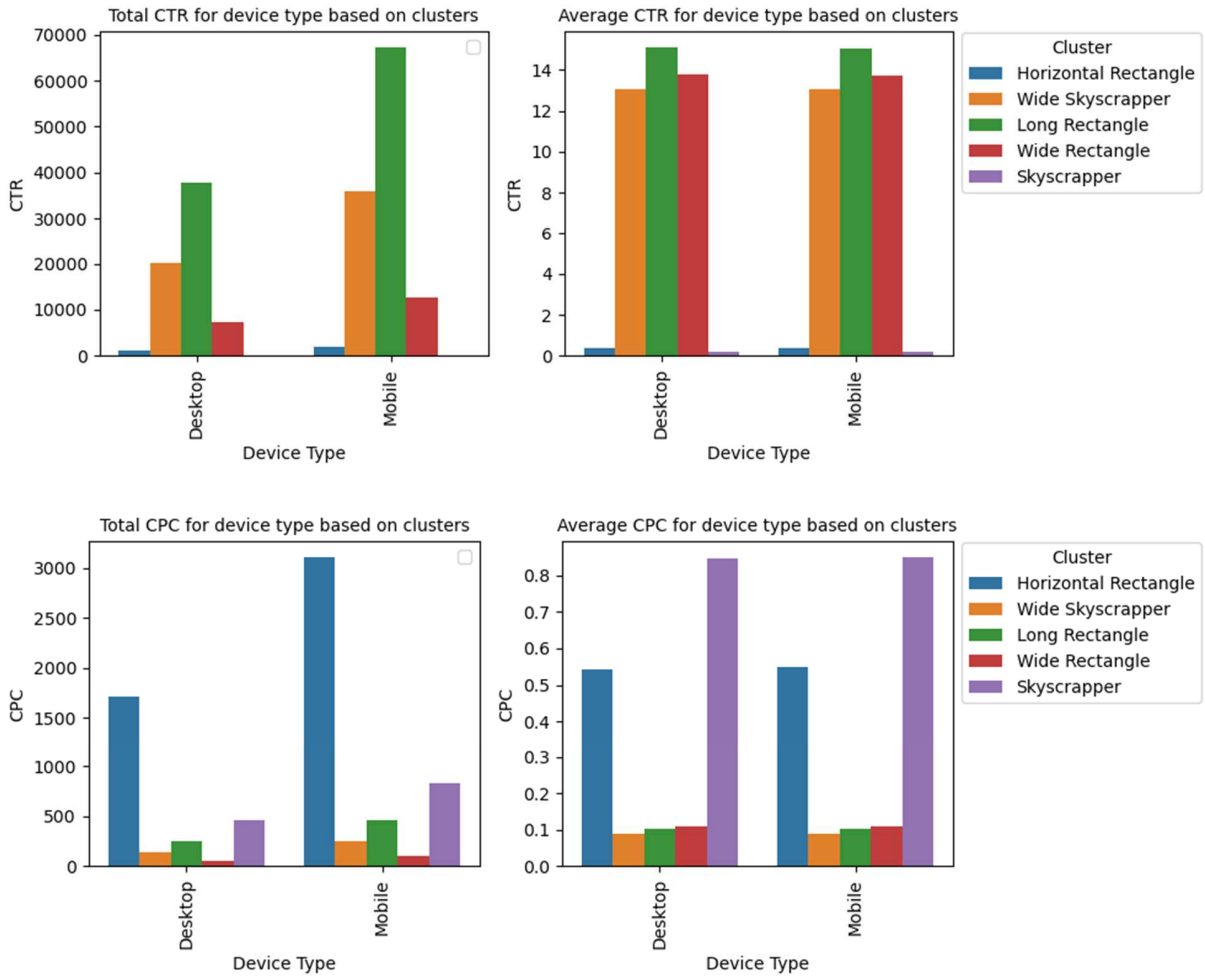


Figure 16: Visual analysis based on clusters

Key Observations

1. First thing that we can take from the above plot on an overall basis is that trends for both device type is very similar both in terms of totals and average.
2. While total impressions count for Horizontal Rectangle clusters in highest followed by Skyscraper for both mobile and desktop, in average terms impressions count for Skyscraper is substantially larger than all the others which means for Skyscraper these impressions are much more concentrated.
3. While in impressions terms there is a huge difference in count for Horizontal Rectangle and Skyscraper with other three cluster types in terms of clicks in totality Wide Rectangle and Wide Skyscraper are performing much better in fact even in average terms Wide Rectangle types are providing highest returns in terms of clicks.
4. Since most impressions are of Horizontal Rectangle and Skyscrapers, ads spent are also in line with this as these clusters have the highest spent in totality and also on average terms most is spent on skyscrapers which means unit economics for Skyscrapers is not very efficient.

5. Revenue contribution from different clusters of ads is similar in proportion terms to ads spent with Horizontal Rectangle and Skyscrapers contributing the most in totality and skyscrapers providing the highest revenue on average terms.
6. The revenue generated by various ad clusters is approximately 30% lower than the expenditure incurred on these ads, which is not an ideal business scenario.
7. Since impression count for Horizontal Rectangle and Skyscrapers are most thus, they have lowest cost per 1000 impressions (CPM) and other 3 cluster have significantly higher CPM with Long Rectangle having the highest in total terms while Wide Rectangle having the highest in average terms.
8. Click through ratio (CTR) for the two clusters with most impressions is minimal which means that though Horizontal Rectangle and Skyscrapers types of ads are used a lot but they are not receiving clicks on the contrary Long Rectangle ads are performing very well both in totality and average terms, Wide Rectangle and Wide Skyscraper types are also performing well in average CTR terms.
9. Given the low click-through rate (CTR) for Horizontal Rectangle and Skyscraper ads, their cost per click (CPC) is high. Conversely, Long Rectangle, Wide Rectangle, and Wide Skyscraper ads, which boast high CTR, exhibit low CPC.

1.11 Conclusion

After analyzing the above data, Ads 24x7 needs to implement substantial changes to its business model. It not only falls short in financial terms, with ad spending nearly 50% higher than revenue, but also in terms of costs and meeting key sector metrics like CTR.

For improvement purposes the following recommendation are made

Recommendations

1. Based on the data provided by advertising platform, the CTR for them based on Matched_Queries where keywords searched matched keywords in the ads resulting in click to Available_Impressions is between the range of 50 and 70, while for Ads 24x7 the maximum CTR achieved is only 15, in fact for Horizontal Rectangle and Skyscraper which have most impressions this value is below 0.4. Thus, it is highly recommended for Ads 24x7 that a detail Keyword analysis project is carried out to study and extract the keywords which can be used to improve the ad content.
2. Ads 24x7 needs to look at the ads spent, as per the principle of unit economics per unit cost comes down as the number of units goes up, based on this principle for ad type which have high impressions their CPM should be low, however, clusters like Wide Rectangle, Wide Skyscraper and Long Rectangle have very high CPM. On the other hand, Horizontal Rectangle and Skyscraper have very poor CTR and CPC. The cost part and budgetary allocation should be relooked at to get better returns.
3. Amongst the clusters Long Rectangle, Wide Rectangle, and Wide Skyscraper are performing much better than Horizontal Rectangle and Skyscraper in terms of CTR and CPC, for the time

being it is highly recommended that Ads 24x7 should use these ad types more until the above mentioned studies are carried out and changes are made, and though they have high CPM, we expect as the number of impressions for these clusters goes up the CPM would come down like in case of Horizontal Rectangle and Skyscraper which could help improve revenue while reducing spendings.

Problem 2

2.1 Background Information

Population census is held in India every ten years, a tradition dating back over a century. These censuses gather extensive data on various aspects of the population, which is then compiled and presented on a district-wise basis. However, due to the multitude of characteristics covered in the census, the compiled data contains numerous variables, making it challenging to extract useful insights.

2.2 Problem Statement

The objective of this analysis is to utilize PCA technique to determine the optimal number of principal components that capture the greatest variance in the abstract of population census data, specifically focusing on female-headed households excluding institutional households. This process aims to reduce the dimensionality of the data.

2.3 METHODOLOGY

Import the libraries - Load the data - Check the structure of the data - Check the types of the data – Check for and treat (if needed) missing values - Check the statistical summary - Check for and treat (if needed) data irregularities – Univariate Analysis – Bivariate Analysis – Data Scaling – PCA – Inference

Key Points

1. **Data Collection**: Data was part of the Population Census 2011 by the Government of India and though the source of data is external, since it belongs to the Government of India it is considered highly reliable.
2. **Data Cleaning and Pre-processing:** Dataset was checked for duplicates, missing values, bad data and outliers. No irregularity was found in the data, however, State Code and Dist.Code columns we irrelevant for analysis so they were dropped during pre-processing and data was scaled to make it fit for applying the PCA algorithm.
3. **Univariate Analysis:** Only 5 numeric columns namely No_HH, TOT_M, TOT_F, TOT_WORK_M and TOT_WORK_F along with State and Area Name were taken and individual variables were analyzed using boxplot and histogram to understand distribution, central tendency and variability of variables.
4. **Bivariate Analysis:** All the above-mentioned variables were examined with the aim of gaining deeper insights about population.
5. **Visualization Techniques:** In the report we have used histograms and boxplot for univariate analysis, in bivariate analysis, to understand correlation between numeric variables heatmap and pair plot are used, bar plot and tables are used to understand relationship between categorical and numeric variable.

6. **Tools and Software:** We have carried out the analysis using programming language python on Jupyter notebook. For this analysis Python libraries Numpy, Pandas, Matplotlib, Seaborn, Scikit-learn, factor_analyzer and PIL were used.
7. **Assumptions and Limitations:** For this analysis we took following assumptions
   a) As per Census 2011 there were 640 districts in India and this data set has 640 rows where each row has a unique Area Name each representing a district, since manually checking these values will not be viable, we assume that there is no irregularity within Area.
   b) The key objective of a census is to record the demographic data compiled on state and district wise basis to understand the changes, similarities and dissimilarities and for this reason we have not treated the outliers as outlier treatment flattens out the data which will have direct impact changes and dissimilarity part of data.

2.4 Data Overview

1. **Data Description:** Dataset has 640 rows and 61 columns

```
shape of the dataset
--------------------------------------------------------------------------------

(640, 61)
```

Table 12: Dataset shape

2. **Data Information:** In the 61 columns 2 are object type and 59 are numeric int64 type.

```
information of features
-------------------------------------------------------------------------------
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 640 entries, 0 to 639
Data columns (total 61 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   State Code      640 non-null    int64
 1   Dist.Code       640 non-null    int64
 2   State           640 non-null    object
 3   Area Name       640 non-null    object
 4   No_HH           640 non-null    int64
 5   TOT_M           640 non-null    int64
 6   TOT_F           640 non-null    int64
 7   M_06            640 non-null    int64
 8   F_06            640 non-null    int64
 9   M_SC            640 non-null    int64
 10  F_SC            640 non-null    int64
 11  M_ST            640 non-null    int64
 12  F_ST            640 non-null    int64
 13  M_LIT           640 non-null    int64
 14  F_LIT           640 non-null    int64
 15  M_ILL           640 non-null    int64
 16  F_ILL           640 non-null    int64
 17  TOT_WORK_M      640 non-null    int64
 18  TOT_WORK_F      640 non-null    int64
 19  MAINWORK_M      640 non-null    int64
 20  MAINWORK_F      640 non-null    int64
```

```
21  MAIN_CL_M        640 non-null    int64
22  MAIN_CL_F        640 non-null    int64
23  MAIN_AL_M        640 non-null    int64
24  MAIN_AL_F        640 non-null    int64
25  MAIN_HH_M        640 non-null    int64
26  MAIN_HH_F        640 non-null    int64
27  MAIN_OT_M        640 non-null    int64
28  MAIN_OT_F        640 non-null    int64
29  MARGWORK_M       640 non-null    int64
30  MARGWORK_F       640 non-null    int64
31  MARG_CL_M        640 non-null    int64
32  MARG_CL_F        640 non-null    int64
33  MARG_AL_M        640 non-null    int64
34  MARG_AL_F        640 non-null    int64
35  MARG_HH_M        640 non-null    int64
36  MARG_HH_F        640 non-null    int64
37  MARG_OT_M        640 non-null    int64
38  MARG_OT_F        640 non-null    int64
39  MARGWORK_3_6_M   640 non-null    int64
40  MARGWORK_3_6_F   640 non-null    int64
41  MARG_CL_3_6_M    640 non-null    int64
42  MARG_CL_3_6_F    640 non-null    int64
43  MARG_AL_3_6_M    640 non-null    int64
44  MARG_AL_3_6_F    640 non-null    int64
45  MARG_HH_3_6_M    640 non-null    int64
46  MARG_HH_3_6_F    640 non-null    int64
47  MARG_OT_3_6_M    640 non-null    int64
48  MARG_OT_3_6_F    640 non-null    int64
49  MARGWORK_0_3_M   640 non-null    int64

 50  MARGWORK_0_3_F   640 non-null    int64
 51  MARG_CL_0_3_M    640 non-null    int64
 52  MARG_CL_0_3_F    640 non-null    int64
 53  MARG_AL_0_3_M    640 non-null    int64
 54  MARG_AL_0_3_F    640 non-null    int64
 55  MARG_HH_0_3_M    640 non-null    int64
 56  MARG_HH_0_3_F    640 non-null    int64
 57  MARG_OT_0_3_M    640 non-null    int64
 58  MARG_OT_0_3_F    640 non-null    int64
 59  NON_WORK_M       640 non-null    int64
 60  NON_WORK_F       640 non-null    int64
dtypes: int64(59), object(2)
memory usage: 305.1+ KB

None
```

Table 13: Dataset information

3. **Missing Data:** There are no missing values in the dataset.

1. No missing values for column State Code
2. No missing values for column Dist.Code
3. No missing values for column State
4. No missing values for column Area Name
5. No missing values for column No_HH
6. No missing values for column TOT_M
7. No missing values for column TOT_F
8. No missing values for column M_06
9. No missing values for column F_06
10. No missing values for column M_SC
11. No missing values for column F_SC
12. No missing values for column M_ST
13. No missing values for column F_ST
14. No missing values for column M_LIT
15. No missing values for column F_LIT
16. No missing values for column M_ILL
17. No missing values for column F_ILL
18. No missing values for column TOT_WORK_M
19. No missing values for column TOT_WORK_F
20. No missing values for column MAINWORK_M

21. No missing values for column MAINWORK_F
22. No missing values for column MAIN_CL_M
23. No missing values for column MAIN_CL_F
24. No missing values for column MAIN_AL_M
25. No missing values for column MAIN_AL_F
26. No missing values for column MAIN_HH_M
27. No missing values for column MAIN_HH_F
28. No missing values for column MAIN_OT_M
29. No missing values for column MAIN_OT_F
30. No missing values for column MARGWORK_M
31. No missing values for column MARGWORK_F
32. No missing values for column MARG_CL_M
33. No missing values for column MARG_CL_F
34. No missing values for column MARG_AL_M
35. No missing values for column MARG_AL_F
36. No missing values for column MARG_HH_M
37. No missing values for column MARG_HH_F
38. No missing values for column MARG_OT_M
39. No missing values for column MARG_OT_F
40. No missing values for column MARGWORK_3_6_M

```
41. No missing values for column MARGWORK_3_6_F
42. No missing values for column MARG_CL_3_6_M
43. No missing values for column MARG_CL_3_6_F
44. No missing values for column MARG_AL_3_6_M
45. No missing values for column MARG_AL_3_6_F
46. No missing values for column MARG_HH_3_6_M
47. No missing values for column MARG_HH_3_6_F
48. No missing values for column MARG_OT_3_6_M
49. No missing values for column MARG_OT_3_6_F
50. No missing values for column MARGWORK_0_3_M
51. No missing values for column MARGWORK_0_3_F
52. No missing values for column MARG_CL_0_3_M
53. No missing values for column MARG_CL_0_3_F
54. No missing values for column MARG_AL_0_3_M
55. No missing values for column MARG_AL_0_3_F
56. No missing values for column MARG_HH_0_3_M
57. No missing values for column MARG_HH_0_3_F
58. No missing values for column MARG_OT_0_3_M
59. No missing values for column MARG_OT_0_3_F
60. No missing values for column NON_WORK_M
61. No missing values for column NON_WORK_F
```

4. **Duplicate Data:** There are no duplicate values in the dataset

```
checking for duplicates
--------------------------------------------------------------------------------
number of dupliacte rows: 0
```

Table 14: Data duplicates

5. **Statistical Summary:**

```
statistical summary
--------------------------------------------------------------------------------
```

	count	mean	std	min	25%	50%	75%	max
State Code	640.0	17.114062	9.426486	1.0	9.00	18.0	24.00	35.0
Dist.Code	640.0	320.500000	184.896367	1.0	160.75	320.5	480.25	640.0
No_HH	640.0	51222.871875	48135.405475	350.0	19484.00	35837.0	68892.00	310450.0
TOT_M	640.0	79940.576563	73384.511114	391.0	30228.00	58339.0	107918.50	485417.0
TOT_F	640.0	122372.084375	113600.717282	698.0	46517.75	87724.5	164251.75	750392.0
M_06	640.0	12309.098438	11500.906881	56.0	4733.75	9159.0	16520.25	96223.0
F_06	640.0	11942.300000	11326.294567	56.0	4672.25	8663.0	15902.25	95129.0
M_SC	640.0	13820.946875	14426.373130	0.0	3466.25	9591.5	19429.75	103307.0
F_SC	640.0	20778.392188	21727.887713	0.0	5603.25	13709.0	29180.00	156429.0
M_ST	640.0	6191.807813	9912.668948	0.0	293.75	2333.5	7658.00	96785.0

F_ST	640.0	10155.640625	15875.701488	0.0	429.50	3834.5	12480.25	130119.0
M_LIT	640.0	57967.979688	55910.282466	286.0	21298.00	42693.5	77989.50	403261.0
F_LIT	640.0	66359.565625	75037.860207	371.0	20932.00	43796.5	84799.75	571140.0
M_ILL	640.0	21972.596875	19825.605268	105.0	8590.00	15767.5	29512.50	105961.0
F_ILL	640.0	56012.518750	47116.693769	327.0	22367.00	42386.0	78471.00	254160.0
TOT_WORK_M	640.0	37992.407813	36419.537491	100.0	13753.50	27936.5	50226.75	269422.0
TOT_WORK_F	640.0	41295.760938	37192.360943	357.0	16097.75	30588.5	53234.25	257848.0
MAINWORK_M	640.0	30204.446875	31480.915680	65.0	9787.00	21250.5	40119.00	247911.0
MAINWORK_F	640.0	28198.846875	29998.262689	240.0	9502.25	18484.0	35063.25	226166.0
MAIN_CL_M	640.0	5424.342188	4739.161969	0.0	2023.50	4160.5	7695.00	29113.0
MAIN_CL_F	640.0	5486.042188	5326.362728	0.0	1920.25	3908.5	7286.25	36193.0
MAIN_AL_M	640.0	5849.109375	6399.507966	0.0	1070.25	3936.5	8067.25	40843.0
MAIN_AL_F	640.0	8925.995312	12864.287584	0.0	1408.75	3933.5	10617.50	87945.0
MAIN_HH_M	640.0	883.893750	1278.642345	0.0	187.50	498.5	1099.25	16429.0
MAIN_HH_F	640.0	1380.773438	3179.414449	0.0	248.75	540.5	1435.75	45979.0
MAIN_OT_M	640.0	18047.101562	26068.480886	36.0	3997.50	9598.0	21249.50	240855.0
MAIN_OT_F	640.0	12406.035938	18972.202369	153.0	3142.50	6380.5	14368.25	209355.0
MARGWORK_M	640.0	7787.960938	7410.791691	35.0	2937.50	5627.0	9800.25	47553.0
MARGWORK_F	640.0	13096.914062	10996.474528	117.0	5424.50	10175.0	18879.25	66915.0
MARG_CL_M	640.0	1040.737500	1311.546847	0.0	311.75	606.5	1281.00	13201.0
MARG_CL_F	640.0	2307.682813	3564.626095	0.0	630.25	1226.0	2659.25	44324.0
MARG_AL_M	640.0	3304.326562	3781.555707	0.0	873.50	2062.0	4300.75	23719.0
MARG_AL_F	640.0	6463.281250	6773.876298	0.0	1402.50	4020.5	9089.25	45301.0
MARG_HH_M	640.0	316.742188	462.661891	0.0	71.75	166.0	356.50	4298.0
MARG_HH_F	640.0	786.626562	1198.718213	0.0	171.75	429.0	962.50	15448.0
MARG_OT_M	640.0	3126.154687	3609.391821	7.0	935.50	2036.0	3985.25	24728.0
MARG_OT_F	640.0	3539.323438	4115.191314	19.0	1071.75	2349.5	4400.50	36377.0
MARGWORK_3_6_M	640.0	41948.168750	39045.316918	291.0	16208.25	30315.0	57218.75	300937.0
MARGWORK_3_6_F	640.0	81076.323438	82970.406216	341.0	26619.50	56793.0	107924.00	676450.0
MARG_CL_3_6_M	640.0	6394.987500	6019.806644	27.0	2372.00	4630.0	8167.00	39106.0
MARG_CL_3_6_F	640.0	10339.864063	8467.473429	85.0	4351.50	8295.0	15102.00	50065.0
MARG_AL_3_6_M	640.0	789.848438	905.639279	0.0	235.50	480.5	986.00	7426.0
MARG_AL_3_6_F	640.0	1749.584375	2496.541514	0.0	497.25	985.5	2059.00	27171.0
MARG_HH_3_6_M	640.0	2743.635938	3059.586387	0.0	718.75	1714.5	3702.25	19343.0
MARG_HH_3_6_F	640.0	5169.850000	5335.640960	0.0	1113.75	3294.0	7502.25	36253.0

MARG_OT_3_6_M	640.0	245.362500	358.728567	0.0	58.00	129.5	276.00	3535.0
MARG_OT_3_6_F	640.0	585.884375	900.025817	0.0	127.75	320.5	719.25	12094.0
MARGWORK_0_3_M	640.0	2616.140625	3036.964381	7.0	755.00	1681.5	3320.25	20648.0
MARGWORK_0_3_F	640.0	2834.545312	3327.836932	14.0	833.50	1834.5	3610.50	25844.0
MARG_CL_0_3_M	640.0	1392.973438	1489.707052	4.0	489.50	949.0	1714.00	9875.0
MARG_CL_0_3_F	640.0	2757.050000	2788.776676	30.0	957.25	1928.0	3599.75	21611.0
MARG_AL_0_3_M	640.0	250.889062	453.336594	0.0	47.00	114.5	270.75	5775.0
MARG_AL_0_3_F	640.0	558.098438	1117.642748	0.0	109.00	247.5	568.75	17153.0
MARG_HH_0_3_M	640.0	560.690625	762.578991	0.0	136.50	308.0	642.00	6116.0
MARG_HH_0_3_F	640.0	1293.431250	1585.377936	0.0	298.00	717.0	1710.75	13714.0
MARG_OT_0_3_M	640.0	71.379688	107.897627	0.0	14.00	35.0	79.00	895.0
MARG_OT_0_3_F	640.0	200.742188	309.740854	0.0	43.00	113.0	240.00	3354.0
NON_WORK_M	640.0	510.014063	610.603187	0.0	161.00	326.0	604.50	6456.0
NON_WORK_F	640.0	704.778125	910.209225	5.0	220.50	464.5	853.50	10533.0

Table 15: Statistical Summary

6. Frequency Distribution of Categorical Columns:

```
frequency distribution of categorical columns
--------------------------------------------------------------------------------
value counts for State
--------------------------------------------------------------------------------
State
Uttar Pradesh               71
Madhya Pradesh              50
Bihar                       38
Maharashtra                 35
Rajasthan                   33
Tamil Nadu                  32
Karnataka                   30
Odisha                      30
Assam                       27
Gujarat                     26
Jharkhand                   24
Andhra Pradesh              23
Jammu & Kashmir             22
Haryana                     21
Punjab                      20
West Bengal                 19
Chhattisgarh                18
Arunachal Pradesh           16
Kerala                      14
Uttarakhand                 13
Himachal Pradesh            12
Nagaland                    11

Manipur                      9
NCT of Delhi                 9
Mizoram                      8
Meghalaya                    7
Tripura                      4
Sikkim                       4
Puducherry                   4
Andaman & Nicobar Island     3
Goa                          2
Daman & Diu                  2
Lakshadweep                  1
Chandigarh                   1
Dadara & Nagar Havelli       1
Name: count, dtype: int64
```

Table 16: Frequency Distribution categorical columns

Key Observations

1. Dataset has 640 rows and 61 columns
2. Dataset has 59 numeric and 2 object type columns

3. In all the numeric columns the minimum value is greater than or equal to 1, there doesn't seem to be any bad data.
4. For missing values output is showing only for first five and last columns so we will have to check of it again.
5. There are no duplicates in the data.
6. There are 2 object type columns one represents the Area Name and another State, as per the census of 2011 there were 35 states and union territories in India comprising of total 640 districts represented by Area Name for our analysis, we will be using term District as it adds more context. States column has 35 unique states and 640 rows each representing a district, however, for some districts we can see that the value counts is more than one this is due to the fact that in some cases district with same name is present in multiple states like district named Bilaspur is present in 2 states namely Chhattisgarh and Himachal Pradesh.
7. Columns State code and Dist.Code were irrelevant for our analysis so they were dropped during Pre-processing phase.
   Updated shape of Data

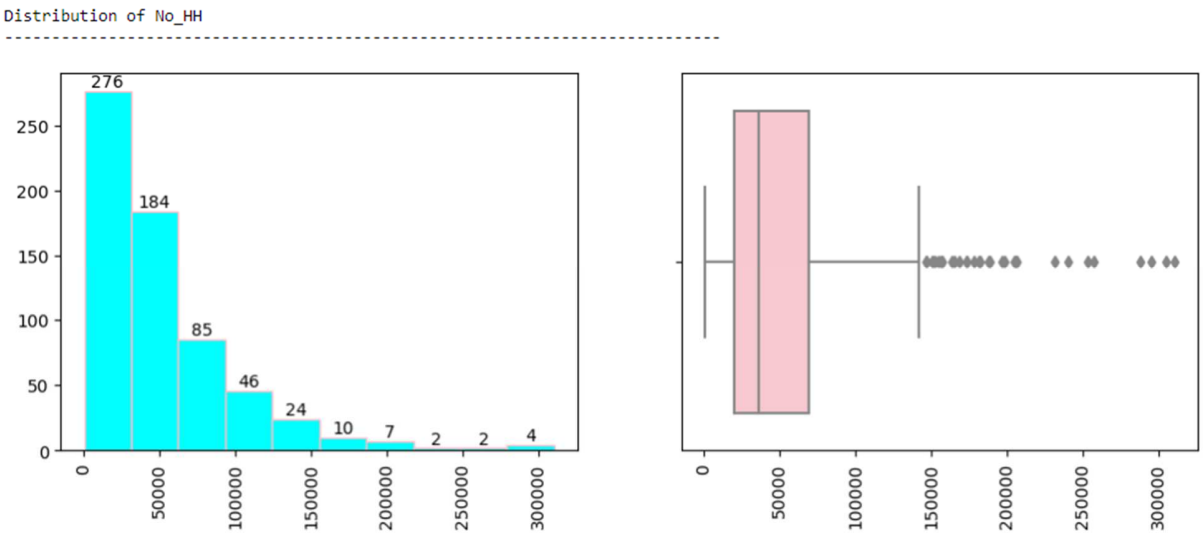
```
(640, 59)
```

2.5 Exploratory Data Analysis

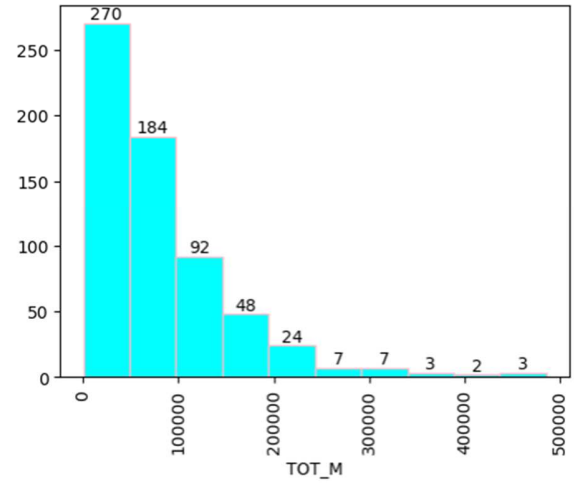
For EDA we have select only 5 columns namely No_HH, TOT_M, TOT_F, TOT_WORK_M and TOT_WORK_F along with State and Area Name columns.

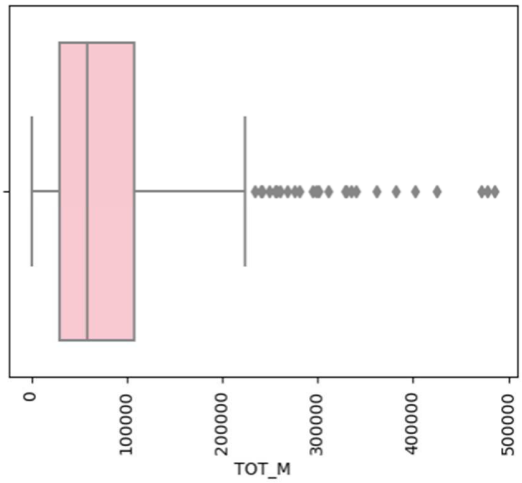
2.5.1 Univariate Analysis

Analysis of Numeric Columns

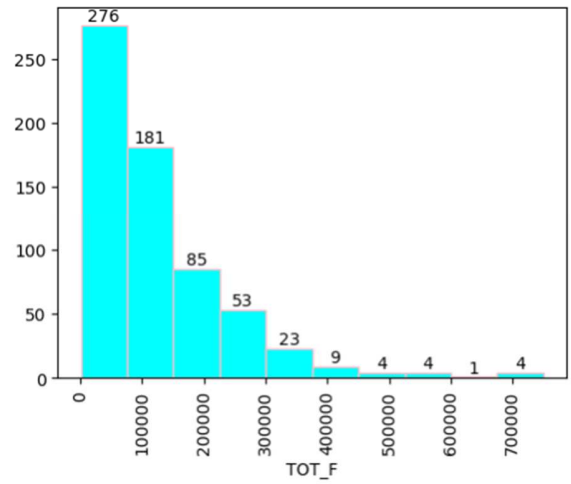


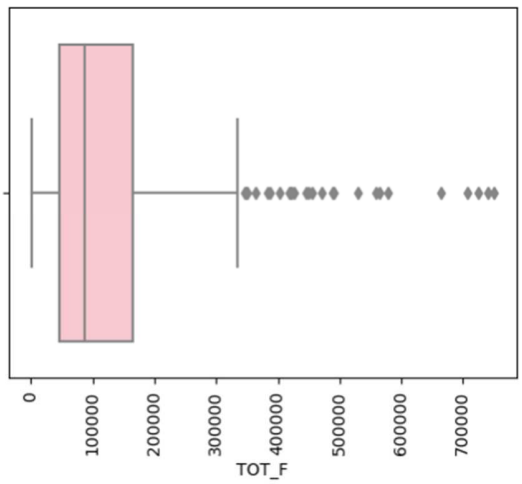
Distribution of TOT_M





Distribution of TOT_F





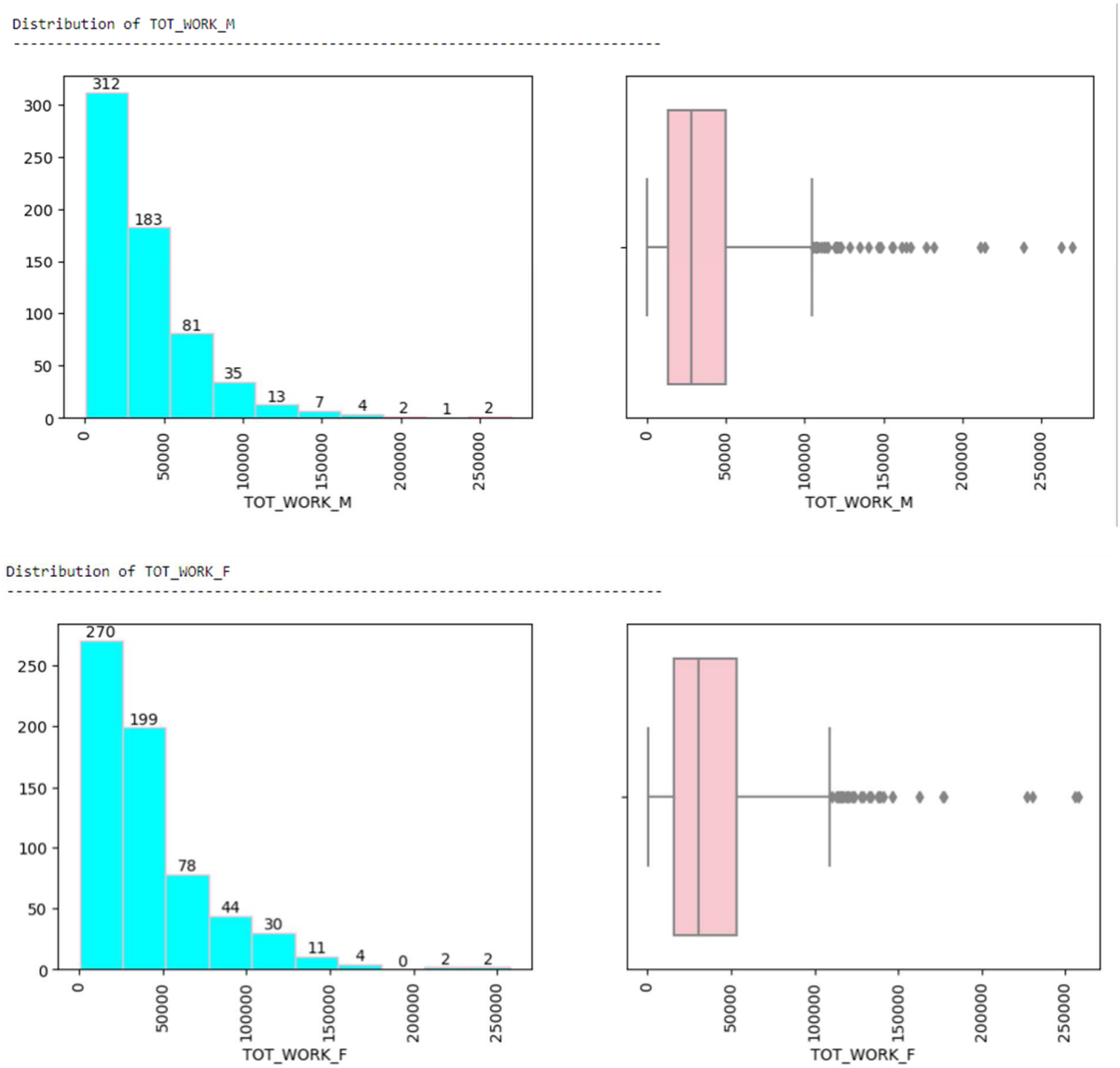


Figure 17: Univariate analysis numeric columns

Key Observations

1. Based on the univariate analysis we can conclude that all the five numeric columns are right skewed and have outliers.
2. In PCA, though the outliers have an impact on principal components, we will refrain from addressing outliers, as one of the fundamental objectives of the census is to capture the variance within the population. This variance is invaluable for both governmental and private organizations in understanding the population and make informed decisions.

Analysis of Categorical Column

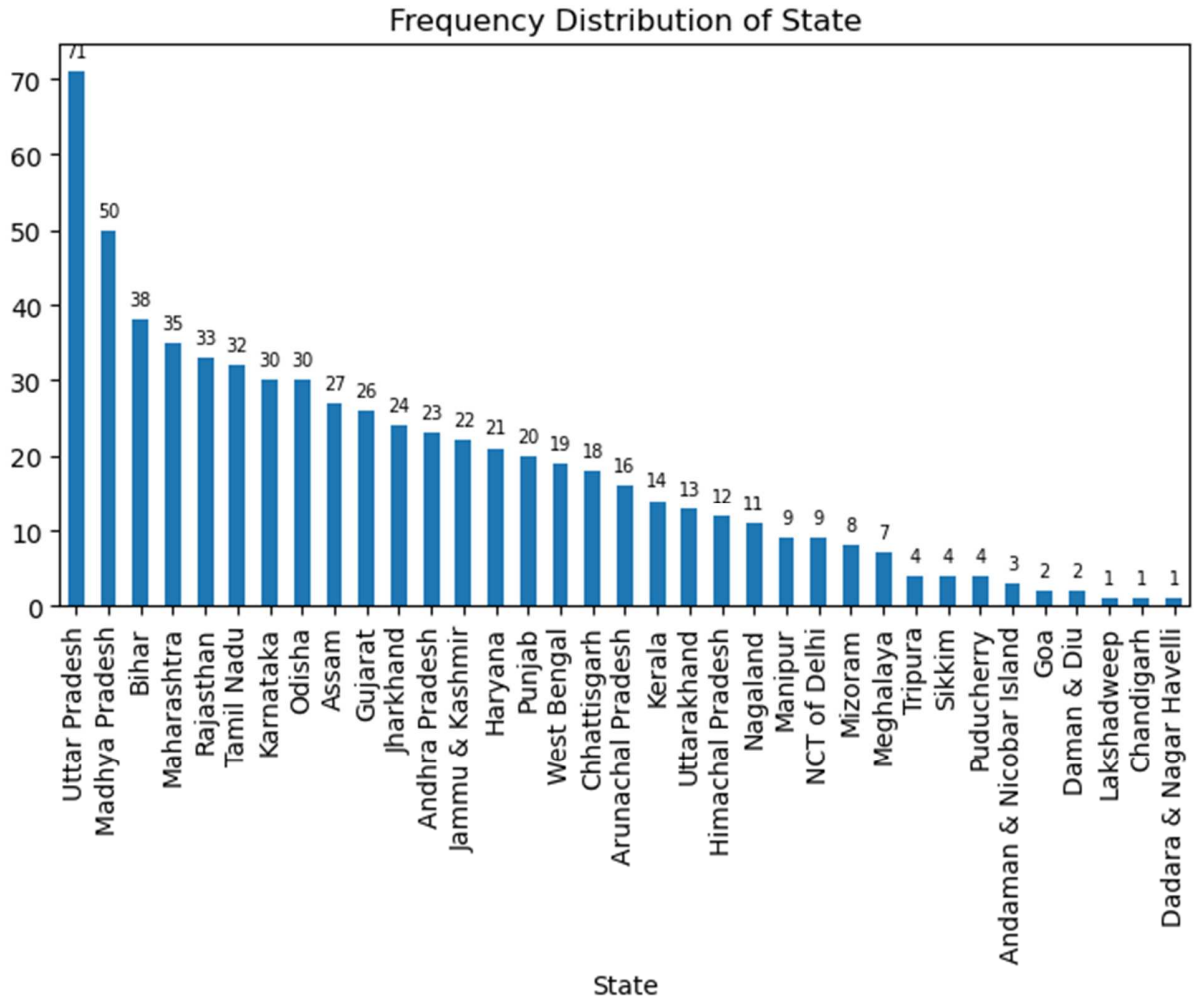


Figure 18: Univariate analysis categorical columns

Key Observations

1. In categorical columns we have done the univariate analysis for States column only since for district column there are 640 districts and visualizing it will not be helpful as the plot would be very cluttered due to presence of large number of elements
2. Amongst the states Uttar Pradesh has the highest number of districts at 70 while Lakshadweep, Chandigarh and Dadar & Nagar Haveli have least number of districts at one each

2.5.2 Bivariate Analysis

Relation between numeric variables

Pair plot showing data spread between numeric features

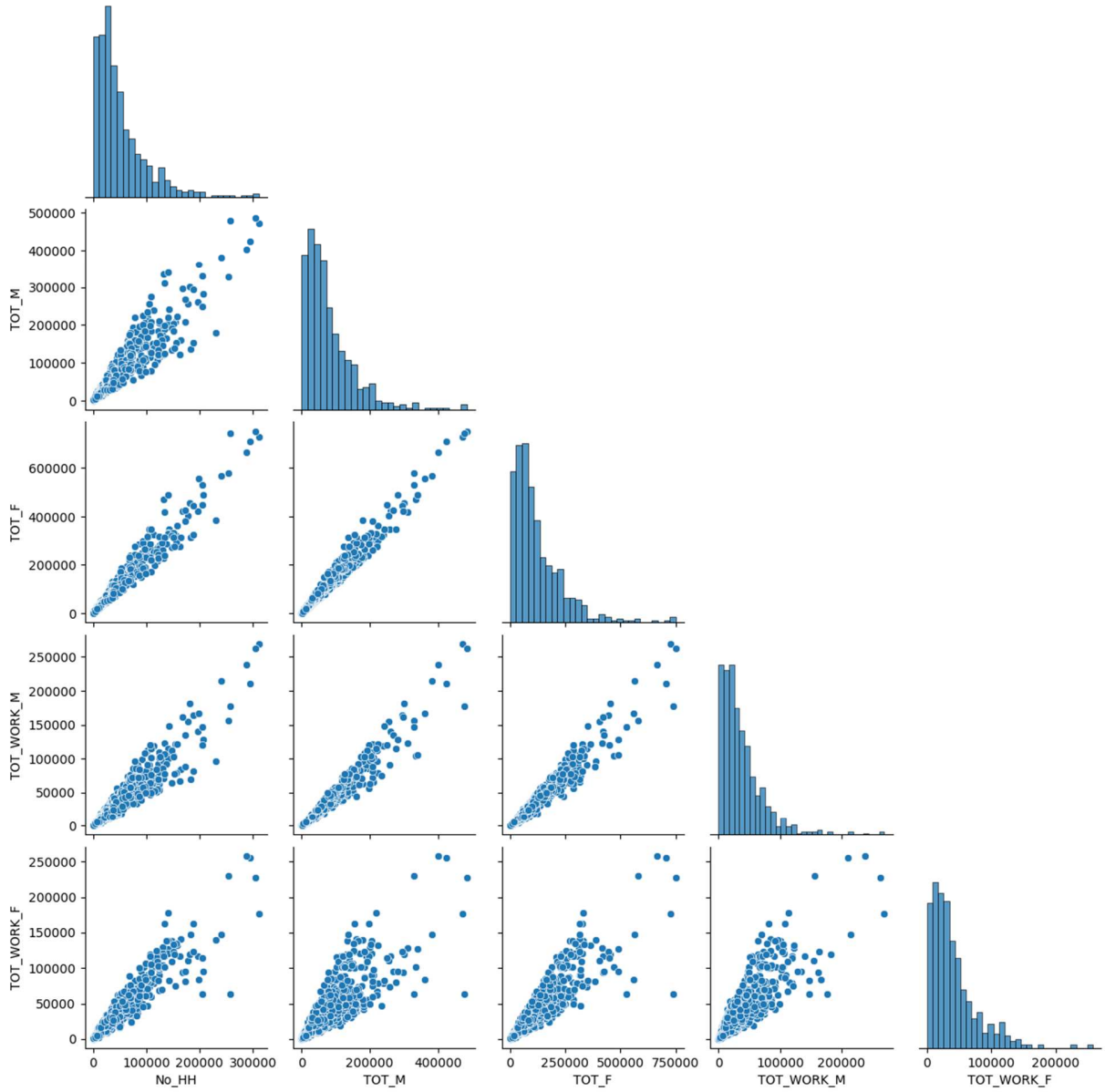


Figure 19: Pair plot

Heatmap showing correlation between numeric variables

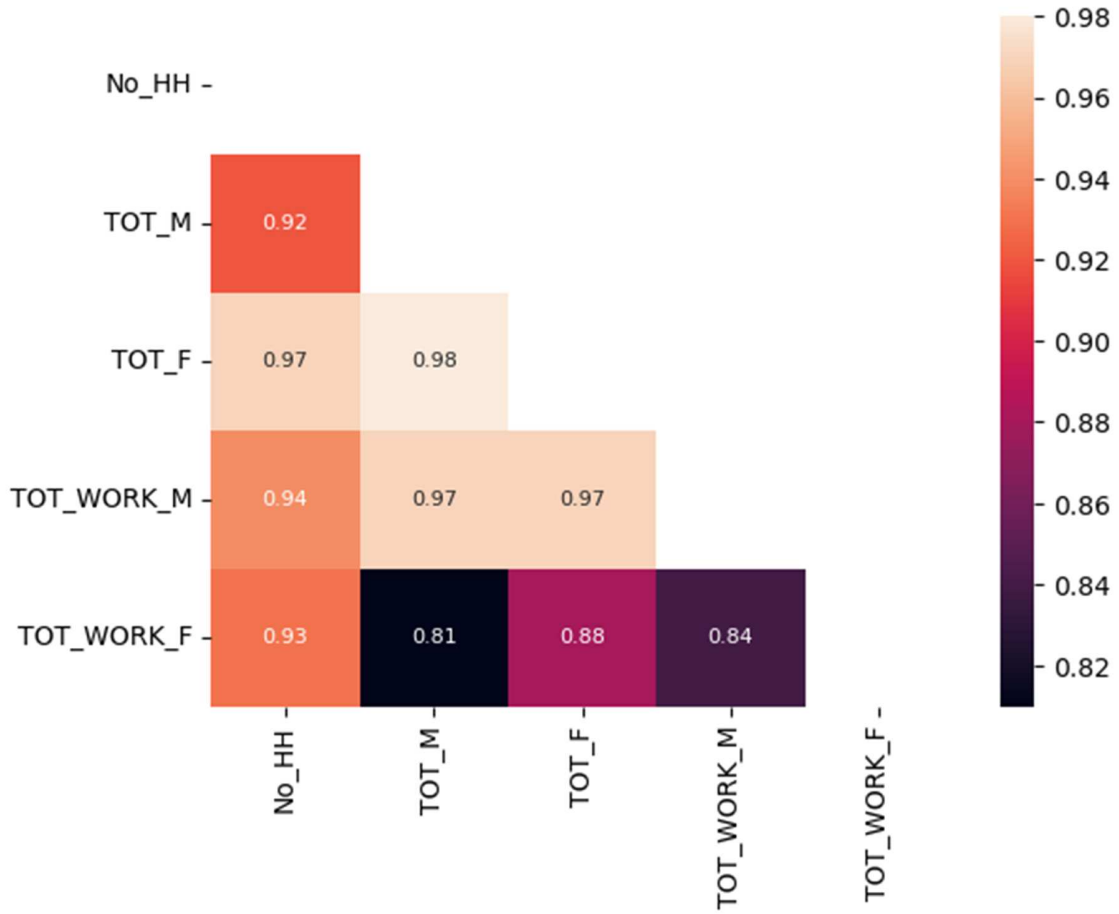


Figure 20: Heatmap

Key Observations

1. There is strong positive correlation between the numeric variables.
2. Applying PCA to this dataset could be beneficial as we might see significant dimension reduction.

Relation between categorical and numeric variables

Bivariate Analysis for State

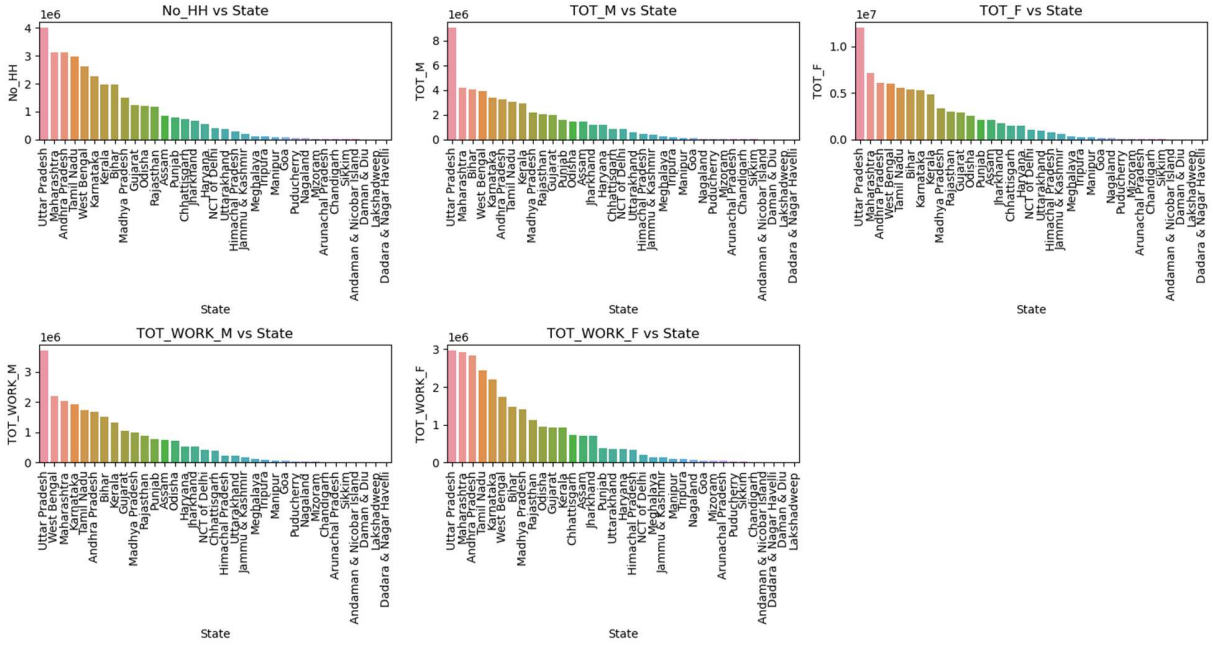


Figure 21: Bivariate Analysis for State

Bivariate Analysis for Area Name

```
top 3 areas with highest No_HH
--------------------------------------------------------------------------------

State        Area Name
West Bengal  North Twenty Four Parganas    310450
Maharashtra  Mumbai Suburban               304502
             Thane                         294698
Name: No_HH, dtype: int64

bottom 3 areas with lowest No_HH
--------------------------------------------------------------------------------

State              Area Name
Arunachal Pradesh  Upper Siang      929
                   Anjaw            783
                   Dibang Valley    350
Name: No_HH, dtype: int64
```

```
top 3 areas with highest TOT_M
--------------------------------------------------------------------------------

State        Area Name
Maharashtra  Mumbai Suburban               485417
Kerala       Malappuram                    477790
West Bengal  North Twenty Four Parganas    471482
Name: TOT_M, dtype: int64

bottom 3 areas with lowest TOT_M
--------------------------------------------------------------------------------

State              Area Name
Arunachal Pradesh  Upper Siang      1187
                   Anjaw             853
                   Dibang Valley     391
Name: TOT_M, dtype: int64


top 3 areas with highest TOT_F
--------------------------------------------------------------------------------

State        Area Name
Maharashtra  Mumbai Suburban               750392
Kerala       Malappuram                    739441
West Bengal  North Twenty Four Parganas    725514
Name: TOT_F, dtype: int64

bottom 3 areas with lowest TOT_F
--------------------------------------------------------------------------------

State              Area Name
Arunachal Pradesh  Upper Siang      2117
                   Anjaw            1688
                   Dibang Valley     698
Name: TOT_F, dtype: int64
```

```
top 3 areas with highest TOT_WORK_M
--------------------------------------------------------------------------------

State        Area Name
West Bengal  North Twenty Four Parganas    269422
Maharashtra  Mumbai Suburban               262638
Karnataka    Bangalore                     238323
Name: TOT_WORK_M, dtype: int64

bottom 3 areas with lowest TOT_WORK_M
--------------------------------------------------------------------------------

State              Area Name
Arunachal Pradesh  Upper Siang      460
                   Anjaw            257
                   Dibang Valley    100
Name: TOT_WORK_M, dtype: int64
```

```
top 3 areas with highest TOT_WORK_F
--------------------------------------------------------------------------------

State        Area Name
Karnataka    Bangalore    257848
Maharashtra  Thane        255770
             Pune         230024
Name: TOT_WORK_F, dtype: int64

bottom 3 areas with lowest TOT_WORK_F
--------------------------------------------------------------------------------

State                     Area Name
Andaman & Nicobar Island  Nicobars         1031
Arunachal Pradesh         Anjaw            1002
                          Dibang Valley     357
Name: TOT_WORK_F, dtype: int64
```

Table 17: Bivariate analysis with Area Name

Key Observations

1. Uttar Pradesh leads in all the metrics like number of households, Total Male and Female population as well as Total Worker Population Male and Female on other side Lakshadweep and Dadar & Nagar Haveli are at the bottom.
2. In district terms, North Twenty-Four Parganas from West Bengal has highest number of household while Dibang Valley from Arunachal Pradesh have least number of households.
3. Mumbai Suburban in Maharashtra has highest population for both Male and Female, Dibang Valley from Arunachal Pradesh has lowest population for both Male and Female.
4. North Twenty-Four Parganas from West Bengal has highest Total Worker Population Male while Bangalore, Karnataka has highest Total Worker Population Female. While here again Dibang Valley from Arunachal Pradesh has the lowest numbers.

Extracting Key proportions

While above we have analysed the data in absolute terms, we have seen that scales are different for different states and districts, digging deeper we will analyse in proportion and ratio terms for which some additional columns were extracted which are as follow

Column Name	Description
mem_hh	Number of people part of a household
sex_ratio	Number of females against every male
work_rat_M	Ratio of working male population in total male population
work_rat_F	Ratio of working female population in total female population
work_sex_ratio	Number of working females against every working male

Table 18: Additional Data Dictionary

Analysis with State column

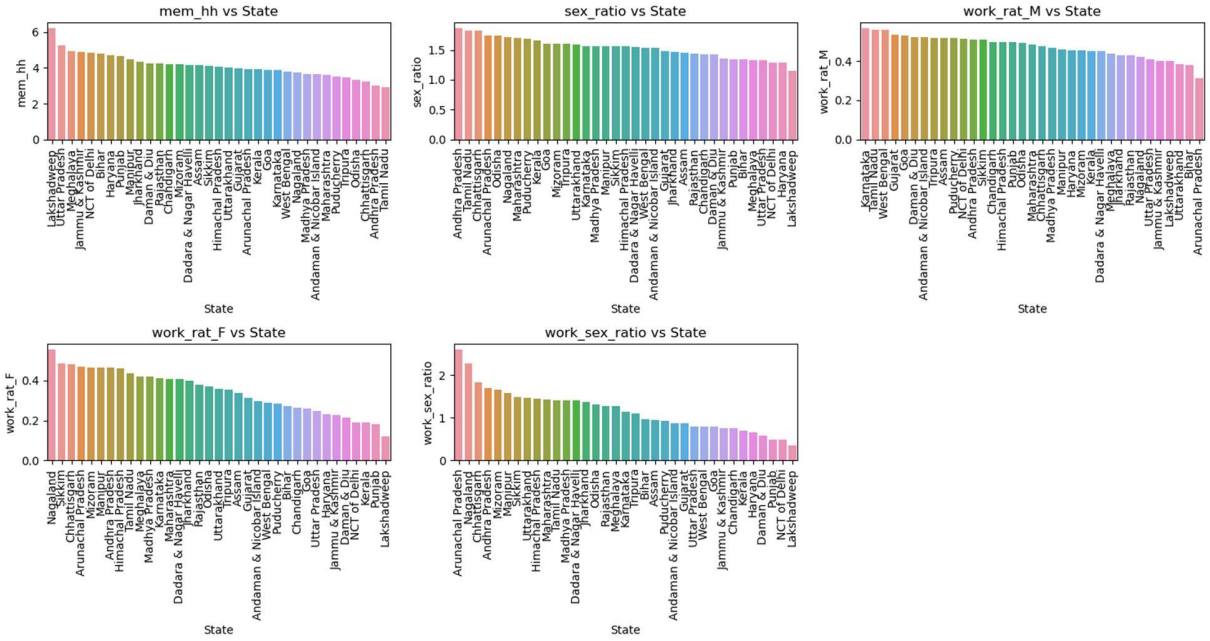


Figure 22: Bivariate Analysis for State

Key Observations

1. While in absolute terms for all features Uttar Pradesh had highest numbers and Lakshadweep and Dadar & Nagar Haveli had lowest number, as we dissect deeper, we can find some interesting trends as other states take lead here.
2. Lakshadweep has highest number of people per household while state of Tamil Nadu has lowest number of people per household.
3. In sex ration Andra Pradesh has highest number of females per male at almost 2.5 female for every male, while Lakshadweep this value is lowest at almost 1.25 female for every male. Though from this we can conclude that for households with female head number of female is higher than male.

4. In Karnataka, approximately 60% of the male population is engaged in work, whereas in Arunachal Pradesh, this figure is at its lowest, standing at 35%. Conversely, Nagaland boasts the highest proportion of working females, accounting for nearly 60%, while Lakshadweep records the lowest, at around 15%. Additionally, the plot depicting the ratio of working males to the ratio of working females indicates that the percentage of working males compared to the total male population is higher than the percentage of working females compared to the total female population.
5. While for total population in all the states we found that there were at least 125 female for very male in all the states but when we compared the sex ratio for working population there are many states working female are less than working male with ratio being as low as below 1:2 for states like Lakshadweep.

Analysis with Area Name column

```
top 3 areas with highest mem_hh
--------------------------------------------------------------------------------

State            Area Name
Jammu & Kashmir  Kupwara      6.90
                 Badgam       6.87
                 Anantnag     6.54
Name: mem_hh, dtype: float64

bottom 3 areas with lowest mem_hh
--------------------------------------------------------------------------------

State       Area Name
Tamil Nadu  Virudhunagar    2.38
            Namakkal        2.35
            Erode           2.30
Name: mem_hh, dtype: float64
```

```
top 3 areas with highest sex_ratio
--------------------------------------------------------------------------------

State           Area Name
Andhra Pradesh  Krishna         2.283
Odisha          Koraput         2.269
Tamil Nadu      Virudhunagar    2.225
Name: sex_ratio, dtype: float64

bottom 3 areas with lowest sex_ratio
--------------------------------------------------------------------------------

State            Area Name
Jammu & Kashmir  Badgam           1.180
Uttar Pradesh    Mahamaya Nagar   1.180
Lakshadweep      Lakshadweep      1.152
Name: sex_ratio, dtype: float64
```

```
top 3 areas with highest work_rat_M
--------------------------------------------------------------------------------

State       Area Name
Tamil Nadu  Tiruppur     0.648
            Erode        0.631
Nagaland    Peren        0.628
Name: work_rat_M, dtype: float64

bottom 3 areas with lowest work_rat_M
--------------------------------------------------------------------------------

State              Area Name
Nagaland           Kiphire           0.243
Arunachal Pradesh  Kurung Kumey      0.240
                   Upper Subansiri   0.233
Name: work_rat_M, dtype: float64


top 3 areas with highest work_rat_F
--------------------------------------------------------------------------------

State     Area Name
Nagaland  Peren        0.760
          Longleng     0.707
          Zunheboto    0.677
Name: work_rat_F, dtype: float64

bottom 3 areas with lowest work_rat_F
--------------------------------------------------------------------------------

State            Area Name
Puducherry       Mahe          0.113
Jammu & Kashmir  Samba         0.106
Kerala           Malappuram    0.085
Name: work_rat_F, dtype: float64


top 3 areas with highest work_sex_ratio
--------------------------------------------------------------------------------

State              Area Name
Arunachal Pradesh  Anjaw           3.899
Nagaland           Kiphire         3.772
Arunachal Pradesh  Dibang Valley   3.570
Name: work_sex_ratio, dtype: float64

bottom 3 areas with lowest work_sex_ratio
--------------------------------------------------------------------------------

State         Area Name
Puducherry    Mahe           0.349
Lakshadweep   Lakshadweep    0.348
NCT of Delhi  North East     0.341
Name: work_sex_ratio, dtype: float64
```

Table 19: Analysis with Area Name column

Key Observations

1. Kupwara district of Jammu and Kashmir at 6.9 has most members per household while Virudhunagar district of Tamil Nadu at 2.38 has least member per household.
2. Krishna district in Andhra Pradesh has the highest female sex ratio at 2.283 female for every male, Lakshadweep in Lakshadweep has lowest female sex ratio at 1.152 female for every male.
3. In Tirupur, Tamil Nadu for every 64.8 male work for every 100 male which is highest in India, while the lowest is Upper Subansiri, Arunachal Pradesh where 23.3 male work for every 100 males.
4. In Peren, Nagaland 76 female work in every 100 female which is highest in India, while in Malappuram, Kerala only 8.5 female work in every 100 females.
5. Amongst the working population, Anjaw, Arunachal Pradesh has highest sex ratio that is for every working male here 3.899 female work while this ratio is lowest for North East of NCT of Delhi where for every working male there is only 0.341 female working.

2.6 Data Scaling

For scaling we will use Z score scaling on numeric columns of the data frame as PCA is applicable only to numeric columns, since data in features are of different scale and Z score method will help reduce this difference bring data on almost similar scales where mean for all the features is almost 0 and standard deviation is 1, ensuring that all features contribute equally to PCA.

Also, since PCA model gives more weightage to bigger values so if we do not treat the data then due to difference in scales features with larger values will have more weightage making the outcome biased, scaling reduces this variability between features resulting in equal weightage to all features.

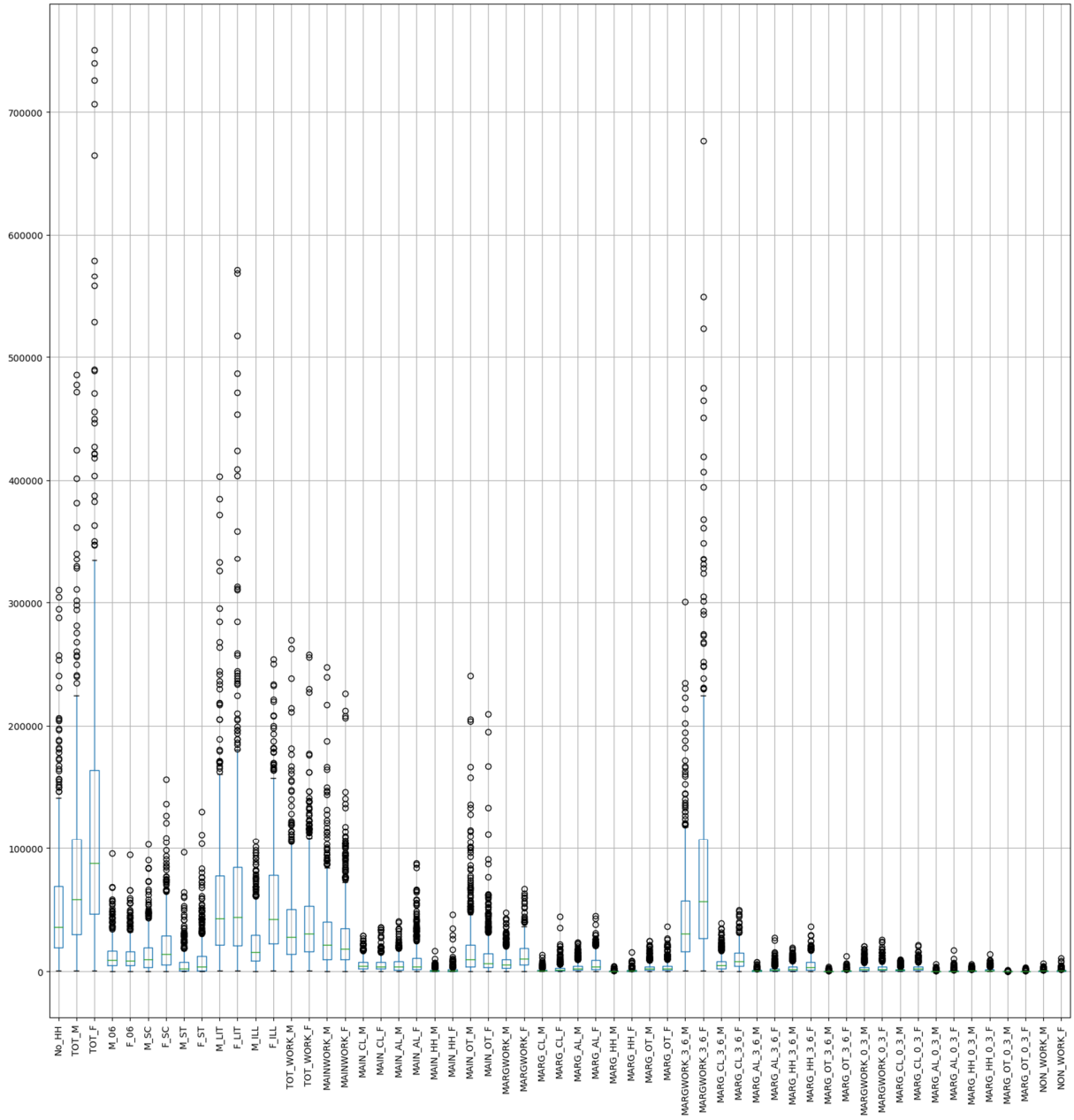


Figure 23: Boxplot unscaled data

In above plot which shows unscaled data of all numeric features, the difference in scale is clearly visible.

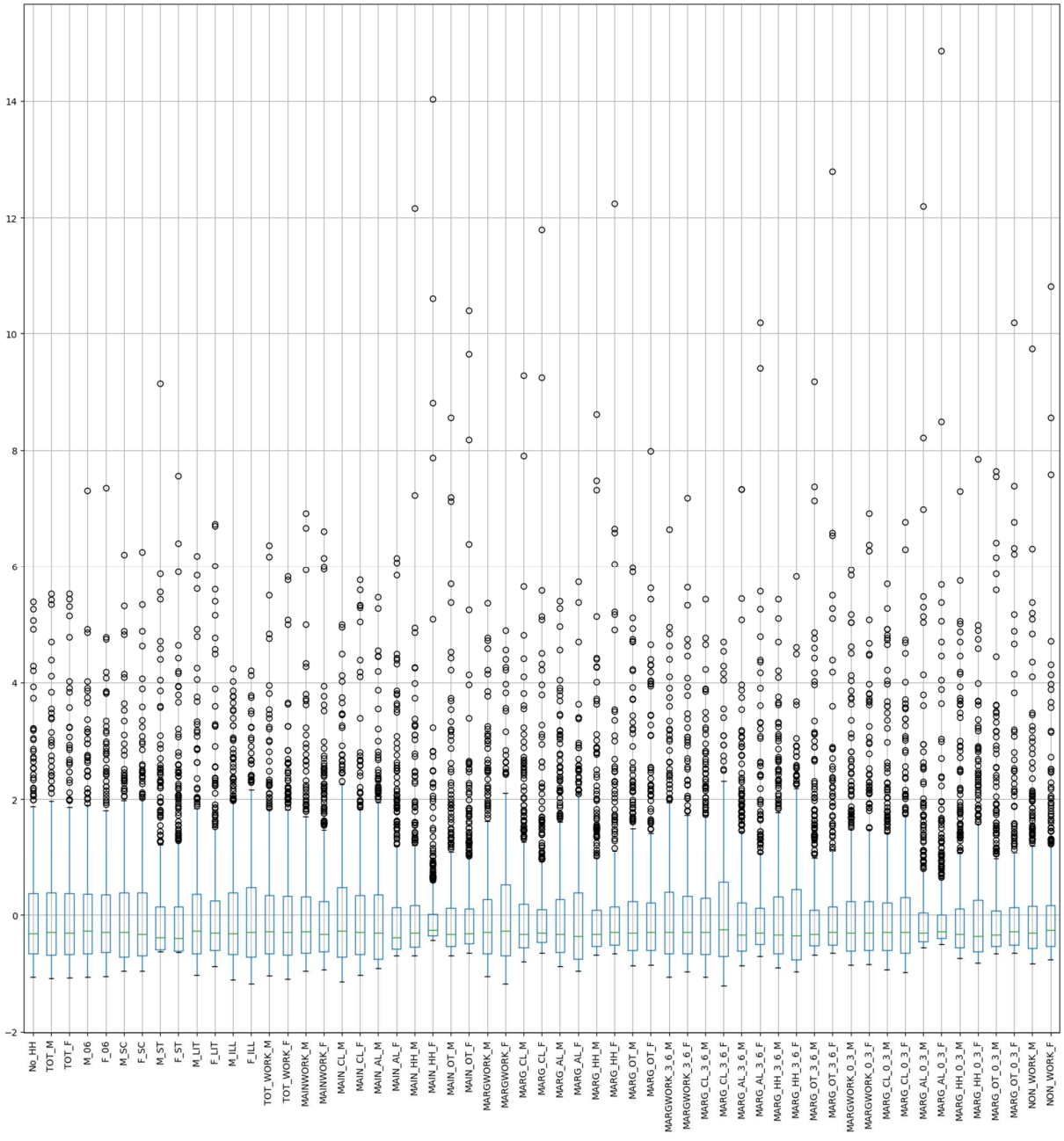


Figure 24: Boxplot scaled data

After scaling the range of data for all features have become almost similar though due to the presence of outliers the range varies but the extent has come down substantially on checking the statistical summary, we will observe that mean and standard deviation for all features have become almost similar.

Statistical summary of scaled data

	count	mean	std	min	25%	50%	75%	max
No_HH	640.0	4.440892e-17	1.000782	-1.057697	-0.659882	-0.319887	0.367358	5.389586
TOT_M	640.0	-8.881784e-17	1.000782	-1.084858	-0.677956	-0.294592	0.381549	5.529690
TOT_F	640.0	-4.440892e-17	1.000782	-1.071906	-0.668250	-0.305233	0.368945	5.532633
M_06	640.0	-5.551115e-17	1.000782	-1.066236	-0.659189	-0.274114	0.366445	7.301993
F_06	640.0	6.661338e-17	1.000782	-1.050264	-0.642376	-0.289756	0.349898	7.350309
M_SC	640.0	5.551115e-18	1.000782	-0.958783	-0.718323	-0.293404	0.389092	6.207800
F_SC	640.0	-5.551115e-17	1.000782	-0.957049	-0.698964	-0.325615	0.386976	6.248040
M_ST	640.0	-4.440892e-17	1.000782	-0.625124	-0.595467	-0.389534	0.148027	9.146281
F_ST	640.0	-2.220446e-17	1.000782	-0.640197	-0.613122	-0.398476	0.146540	7.562324
M_LIT	640.0	-4.440892e-17	1.000782	-1.032495	-0.656385	-0.273410	0.358381	6.180672
F_LIT	640.0	0.000000e+00	1.000782	-0.880091	-0.605869	-0.300924	0.245937	6.732272
M_ILL	640.0	3.885781e-17	1.000782	-1.103860	-0.675544	-0.313229	0.380609	4.239674
F_ILL	640.0	-4.440892e-17	1.000782	-1.182788	-0.714648	-0.289434	0.477029	4.208752
TOT_WORK_M	640.0	-4.440892e-17	1.000782	-1.041256	-0.666067	-0.276329	0.336191	6.359515
TOT_WORK_F	640.0	-8.881784e-17	1.000782	-1.101591	-0.678035	-0.288114	0.321244	5.827047
MAINWORK_M	640.0	-2.220446e-17	1.000782	-0.958137	-0.649073	-0.284647	0.315185	6.920918
MAINWORK_F	640.0	4.440892e-17	1.000782	-0.932745	-0.623743	-0.324100	0.229006	6.604449
MAIN_CL_M	640.0	-8.881784e-17	1.000782	-1.145474	-0.718165	-0.266889	0.479501	5.002401
MAIN_CL_F	640.0	-1.110223e-17	1.000782	-1.030785	-0.669985	-0.296408	0.338245	5.769599
MAIN_AL_M	640.0	0.000000e+00	1.000782	-0.914709	-0.747338	-0.299102	0.346882	5.472493
MAIN_AL_F	640.0	4.440892e-17	1.000782	-0.694401	-0.584807	-0.388393	0.131591	6.147314
MAIN_HH_M	640.0	1.665335e-17	1.000782	-0.691816	-0.545061	-0.301644	0.168557	12.167019
MAIN_HH_F	640.0	0.000000e+00	1.000782	-0.434625	-0.356326	-0.264492	0.017305	14.038154
MAIN_OT_M	640.0	0.000000e+00	1.000782	-0.691455	-0.539371	-0.324365	0.122942	8.553708
MAIN_OT_F	640.0	-4.440892e-17	1.000782	-0.646347	-0.488651	-0.317847	0.103507	10.389042
MARGWORK_M	640.0	-1.665335e-17	1.000782	-1.046990	-0.655025	-0.291825	0.271747	5.370026
MARGWORK_F	640.0	2.220446e-17	1.000782	-1.181294	-0.698262	-0.265922	0.526247	4.897950
MARG_CL_M	640.0	0.000000e+00	1.000782	-0.794140	-0.556257	-0.331347	0.183333	9.278947
MARG_CL_F	640.0	-5.551115e-17	1.000782	-0.647891	-0.470946	-0.303687	0.098704	11.796239
MARG_AL_M	640.0	1.110223e-17	1.000782	-0.874484	-0.643314	-0.328780	0.263702	5.402708
MARG_AL_F	640.0	2.220446e-17	1.000782	-0.954894	-0.747687	-0.360900	0.387964	5.737940
MARG_HH_M	640.0	-5.551115e-18	1.000782	-0.685144	-0.529942	-0.326070	0.086000	8.611844
MARG_HH_F	640.0	1.110223e-17	1.000782	-0.656736	-0.513346	-0.298574	0.146833	12.240442

MARG_OT_M	640.0	1.110223e-17	1.000782	-0.864853	-0.607407	-0.302269	0.238203	5.989580
MARG_OT_F	640.0	-4.440892e-17	1.000782	-0.856115	-0.600094	-0.289356	0.209431	7.985865
MARGWORK_3_6_M	640.0	7.216450e-17	1.000782	-1.067727	-0.659748	-0.298173	0.391405	6.638220
MARGWORK_3_6_F	640.0	-2.220446e-17	1.000782	-0.973823	-0.656854	-0.292903	0.323834	7.181348
MARG_CL_3_6_M	640.0	-2.220446e-17	1.000782	-1.058667	-0.668815	-0.293426	0.294594	5.438148
MARG_CL_3_6_F	640.0	-8.881784e-17	1.000782	-1.212036	-0.707773	-0.241685	0.562843	4.695168
MARG_AL_3_6_M	640.0	-4.440892e-17	1.000782	-0.872827	-0.612586	-0.341847	0.216758	7.333319
MARG_AL_3_6_F	640.0	4.440892e-17	1.000782	-0.701351	-0.502020	-0.306297	0.124035	10.190617
MARG_HH_3_6_M	640.0	-7.216450e-17	1.000782	-0.897436	-0.662335	-0.336627	0.313560	5.429606
MARG_HH_3_6_F	640.0	-6.661338e-17	1.000782	-0.969686	-0.760784	-0.351845	0.437478	5.830127
MARG_OT_3_6_M	640.0	-5.551115e-18	1.000782	-0.684513	-0.522705	-0.323234	0.085473	9.177442
MARG_OT_3_6_F	640.0	3.330669e-17	1.000782	-0.651473	-0.509422	-0.295094	0.148296	12.796429
MARGWORK_0_3_M	640.0	0.000000e+00	1.000782	-0.859800	-0.613309	-0.307996	0.232028	5.942106
MARGWORK_0_3_F	640.0	0.000000e+00	1.000782	-0.848224	-0.601775	-0.300744	0.233353	6.919646
MARG_CL_0_3_M	640.0	-2.775558e-17	1.000782	-0.933110	-0.606952	-0.298260	0.215665	5.698208
MARG_CL_0_3_F	640.0	-5.551115e-17	1.000782	-0.978631	-0.645877	-0.297513	0.302412	6.765940
MARG_AL_0_3_M	640.0	2.220446e-17	1.000782	-0.553861	-0.450104	-0.301091	0.043845	12.194982
MARG_AL_0_3_F	640.0	-2.220446e-17	1.000782	-0.499744	-0.402141	-0.278122	0.009538	14.859741
MARG_HH_0_3_M	640.0	4.440892e-17	1.000782	-0.735831	-0.556693	-0.331622	0.106708	7.290595
MARG_HH_0_3_F	640.0	-1.110223e-17	1.000782	-0.816489	-0.628374	-0.363877	0.263436	7.840581
MARG_OT_0_3_M	640.0	-2.775558e-17	1.000782	-0.662068	-0.532213	-0.337432	0.070681	7.639320
MARG_OT_0_3_F	640.0	0.000000e+00	1.000782	-0.648604	-0.509670	-0.283498	0.126843	10.188272
NON_WORK_M	640.0	-2.220446e-17	1.000782	-0.835916	-0.572036	-0.301600	0.154863	9.745505
NON_WORK_F	640.0	-6.661338e-17	1.000782	-0.769412	-0.532468	-0.264188	0.163521	10.806207

Table 20: Statistical Summary of scaled data

2.7 Principal Component Analysis(PCA)

The main objective of this analysis is to reduce the dimensions of the dataset using PCA, however, for PCA to be effective data has fulfilled conditions.

1. There should be significant correlation between different pair of variables
2. Sample size available should be adequate.

Correlation significance test

We will do Bartlett Sphericity Test to confirm that there is significant correlation between the features.

Hypothesis

Null Hypothesis (H0): All variables in data are uncorrelated.

Alternate Hypothesis (Ha): At least one pair of variables are significantly correlated.

If the null hypothesis cannot be rejected, then PCA is not advisable.

If the p-value is smaller than level of significance, then we can reject the null hypothesis and agree that there is at least one pair of variables in the data which are correlated hence PCA is recommended.

Level of significance (alpha): 0.05

```
p-value of test is 0.0
```

Since p-value of 0.0 is significantly lower than the level of significance of 0.05, hence, with 95% confidence we can reject the null hypothesis (H0) that all variables in the data are uncorrelated and can continue with the PCA.

Adequacy of sample size

To check that we have adequate samples to carry out PCA we will do KMO test which will provide us measure of sampling adequacy (MSA) which is used to examine how appropriate PCA is.

If MSA is less than 0.5, PCA is not recommended, since no significant reduction in dimensions is expected. On the other hand, MSA > 0.7 is expected to provide a considerable reduction is the dimension and extraction of meaningful components.

```
Measure of sampling adequacy is 0.804
```

Since MSA value of 0.804 is greater than 0.7, hence, we can continue with PCA.

Applying PCA

On applying the PCA we get eigen vector and eigen value where eigen vector represents the direction of maximum variance in the data explained by a PC for each feature and eigen value represents the percentage of variance in data explained by a PC.

Initially for applying we take principal components equal to the number of numeric columns which in this case is 57 and get the eigen vector and eigen values, the output received is in the form of array.

```
array([[-4.62, -4.77, -5.96, ..., -6.29, -6.22, -5.9 ],
       [ 0.14, -0.11, -0.29, ..., -0.64, -0.67, -0.94],
       [ 0.33,  0.24,  0.37, ...,  0.11,  0.27,  0.35],
       ...,
       [-0.  ,  0.  ,  0.  , ...,  0.  , -0.  , -0.  ],
       [ 0.  ,  0.  , -0.  , ...,  0.  , -0.  ,  0.  ],
       [ 0.  ,  0.  ,  0.  , ..., -0.  ,  0.  , -0.  ]])
```

Table 21: Eigen vectors

Eigen vector for 57 principal components, to choose the optimum numbers of PC's that represents most variance in the data we have to plot a spree plot for eigen values where number PC's with eigen value greater than 1 are considered optimum number of PC's

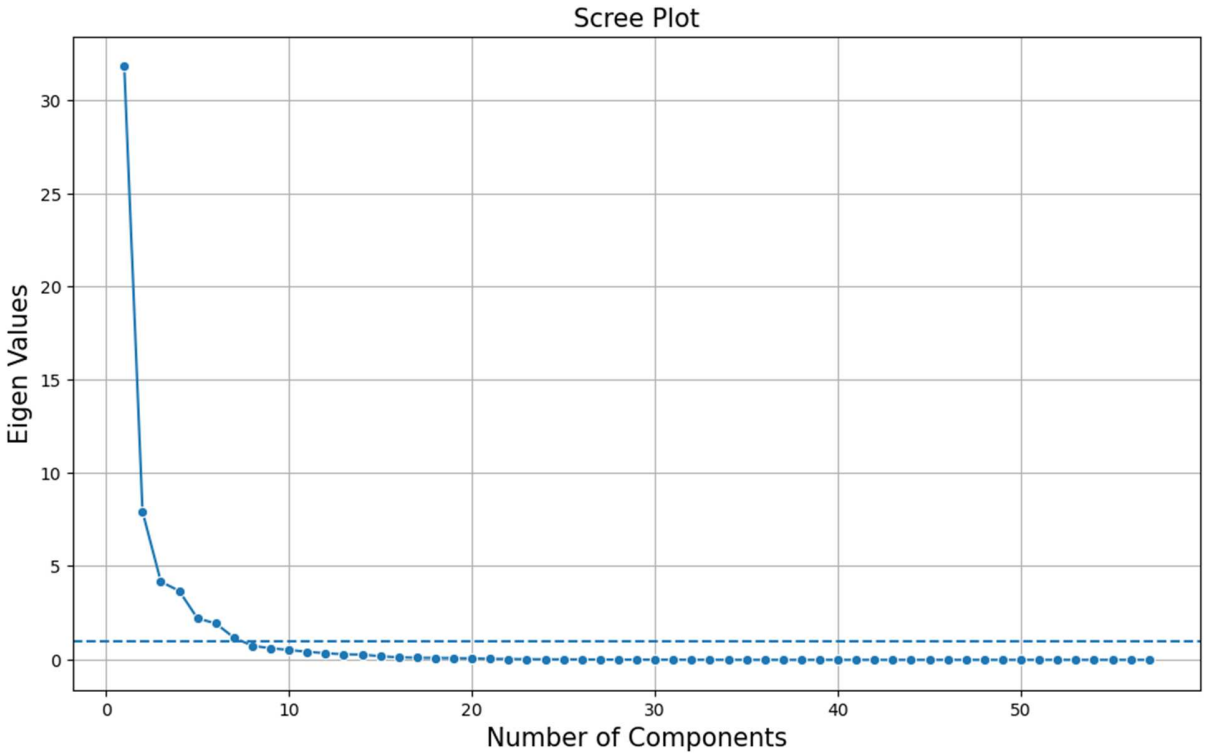


Figure 25: Scree plot

In the above plot eigen value one is marked with dotted blue line and 7 PC's have their eigen value greater than 1. However, for this project we have to ensure that PCs explain at least 90% of the variance and to check how many PCs together explain 90% variance we will have to use cumulative variance.

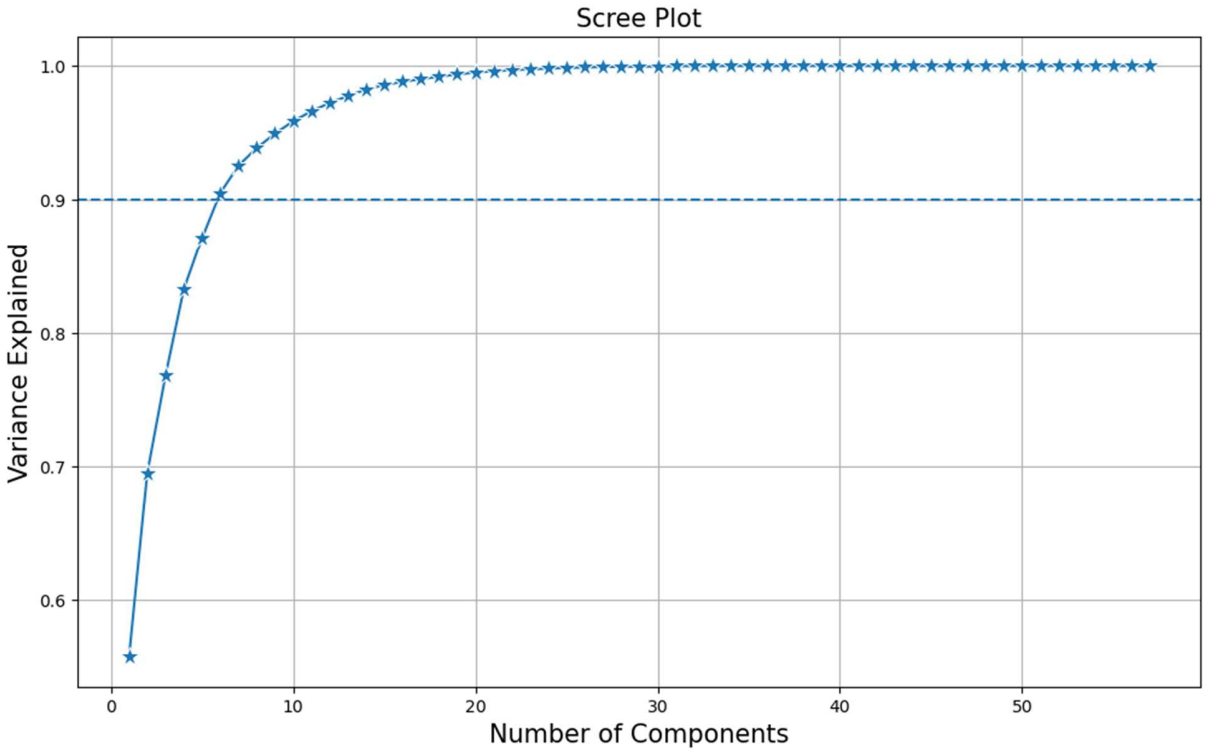


Figure 26: Scree plot

As per the scree plot above for cumulative variance, 6 PC's together can explain more than 90% variance where 90% mark is depicted by horizontal dashed line in the plot, so for this PCA we will take 6 as the optimum number of PC's.

To confirm that 6 PCs together account for over 90% variance we have displayed the cumulative variance in the array form also.

```
Cumulative variance for first six PCs

array([0.55726063, 0.69510499, 0.76785794, 0.83212212, 0.87077261,
       0.9047243 ])
```

Table 22: Cumulative variance of PCs

Based on the above results we again applied PCA this time taking number of components as six for which we got the following eigen vectors

```
array([[-4.62, -4.77, -5.96, ..., -6.29, -6.22, -5.9 ],
       [ 0.14, -0.11, -0.29, ..., -0.64, -0.67, -0.94],
       [ 0.33,  0.24,  0.37, ...,  0.11,  0.27,  0.35],
       [ 1.54,  1.96,  0.62, ...,  1.37,  1.14,  1.11],
       [ 0.35, -0.15,  0.48, ...,  0.15,  0.06,  0.15],
       [-0.42,  0.42,  0.28, ...,  0.14, -0.12, -0.15]])
```

Table 23: Eigen vectors

We loaded the components of six PCs for each of the columns of the original data set where values in cells of the table below represents how much variance of a columns is explained by that PC.

	PC1	PC2	PC3	PC4	PC5	PC6
No_HH	0.16	-0.13	-0.00	-0.13	-0.01	0.00
TOT_M	0.17	-0.09	0.06	-0.02	-0.03	-0.07
TOT_F	0.17	-0.10	0.04	-0.07	-0.01	-0.04
M_06	0.16	-0.02	0.06	0.01	-0.05	-0.16
F_06	0.16	-0.02	0.05	0.01	-0.04	-0.15
M_SC	0.15	-0.05	0.00	0.01	-0.17	-0.06
F_SC	0.15	-0.05	-0.03	-0.03	-0.16	-0.04
M_ST	0.03	0.03	-0.12	-0.22	0.43	0.22
F_ST	0.03	0.03	-0.14	-0.23	0.44	0.23
M_LIT	0.16	-0.12	0.08	-0.04	-0.01	-0.06
F_LIT	0.15	-0.15	0.12	-0.06	0.06	-0.05
M_ILL	0.16	-0.01	-0.02	0.03	-0.10	-0.12
F_ILL	0.17	-0.01	-0.09	-0.08	-0.12	-0.03
TOT_WORK_M	0.16	-0.13	0.05	-0.04	-0.02	-0.00
TOT_WORK_F	0.15	-0.09	-0.06	-0.23	-0.04	0.11
MAINWORK_M	0.15	-0.18	0.05	-0.07	-0.04	0.02
MAINWORK_F	0.12	-0.15	-0.06	-0.25	-0.08	0.12

MAIN_CL_M	0.10	0.06	-0.07	-0.09	-0.29	-0.01
MAIN_CL_F	0.07	0.09	-0.01	-0.29	-0.24	0.10
MAIN_AL_M	0.11	-0.03	-0.25	-0.14	-0.21	-0.03
MAIN_AL_F	0.07	-0.06	-0.25	-0.29	-0.18	0.02
MAIN_HH_M	0.13	-0.08	0.03	0.15	-0.13	0.17
MAIN_HH_F	0.08	-0.08	-0.06	0.05	-0.14	0.42
MAIN_OT_M	0.12	-0.21	0.14	-0.04	0.06	0.02
MAIN_OT_F	0.11	-0.21	0.10	-0.12	0.08	0.08
MARGWORK_M	0.16	0.09	-0.01	0.09	0.06	-0.09
MARGWORK_F	0.16	0.13	-0.05	-0.09	0.09	0.02
MARG_CL_M	0.08	0.27	0.20	-0.06	-0.02	0.03
MARG_CL_F	0.05	0.25	0.27	-0.17	-0.06	0.09
MARG_AL_M	0.13	0.17	-0.19	0.09	0.02	-0.14
MARG_AL_F	0.11	0.14	-0.27	-0.11	0.08	-0.09
MARG_HH_M	0.14	0.07	-0.02	0.24	-0.06	0.09
MARG_HH_F	0.13	0.02	-0.08	0.20	-0.03	0.37
MARG_OT_M	0.16	-0.09	0.11	0.09	0.12	-0.06
MARG_OT_F	0.15	-0.12	0.10	0.03	0.17	0.00
MARGWORK_3_6_M	0.16	-0.04	0.06	-0.00	-0.04	-0.14
MARGWORK_3_6_F	0.16	-0.11	0.08	0.00	0.00	-0.11
MARG_CL_3_6_M	0.17	0.08	-0.02	0.09	0.05	-0.10
MARG_CL_3_6_F	0.16	0.10	-0.07	-0.11	0.07	0.02
MARG_AL_3_6_M	0.09	0.26	0.15	-0.04	-0.01	0.01
MARG_AL_3_6_F	0.05	0.24	0.26	-0.18	-0.06	0.09
MARG_HH_3_6_M	0.13	0.16	-0.20	0.08	0.01	-0.14
MARG_HH_3_6_F	0.11	0.13	-0.28	-0.14	0.06	-0.08
MARG_OT_3_6_M	0.14	0.06	-0.02	0.24	-0.07	0.10
MARG_OT_3_6_F	0.12	0.01	-0.08	0.19	-0.04	0.38
MARGWORK_0_3_M	0.15	-0.09	0.11	0.09	0.11	-0.06
MARGWORK_0_3_F	0.15	-0.13	0.10	0.03	0.14	0.01
MARG_CL_0_3_M	0.15	0.15	0.05	0.09	0.08	-0.06

MARG_CL_0_3_F	0.14	0.18	0.02	-0.02	0.13	-0.00
MARG_AL_0_3_M	0.05	0.25	0.27	-0.10	-0.05	0.07
MARG_AL_0_3_F	0.04	0.24	0.28	-0.14	-0.05	0.08
MARG_HH_0_3_M	0.12	0.19	-0.14	0.13	0.06	-0.12
MARG_HH_0_3_F	0.12	0.18	-0.20	0.00	0.13	-0.11
MARG_OT_0_3_M	0.14	0.08	-0.02	0.23	-0.04	0.06
MARG_OT_0_3_F	0.13	0.05	-0.08	0.21	0.00	0.30
NON_WORK_M	0.15	-0.07	0.11	0.08	0.16	-0.05
NON_WORK_F	0.13	-0.07	0.10	0.02	0.24	-0.02

Table 24: Components of selected PCs for original dataset columns

we can represent each PC as a linear combination of original features. For example, we can write the equation for PC1 in the following manner:

PC1 = 0.16No_HH + 0.17TOT_M + 0.17TOT_F + 0.16M_06 + 0.16F_06 + 0.15M_SC + 0.15F_SC+.....+0.15NON_WORK_M + 0.13*NON_WORK_F

Equation 4: PC1 linear equation

now we will identify which features have maximum loading across features.

1. We will plot the component loading table above as a heatmap
2. We will highlight the features with rectangular red box representing the PC that explains most variance for that feature. These marked features decide the context that the component represents.

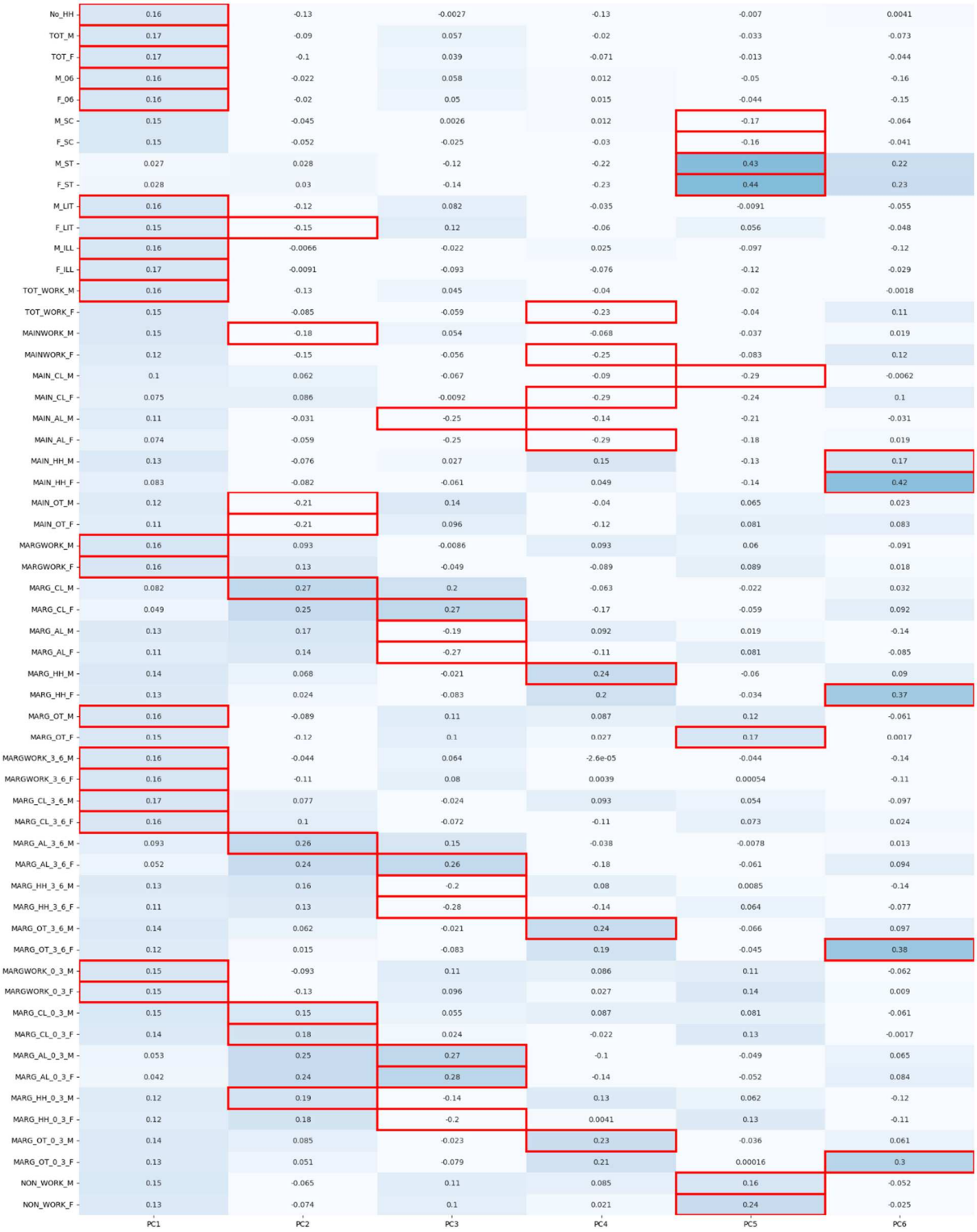


Figure 27: Heatmap

2.8 Inference

1. Based on the above heatmap different PC's explain highest variance for different features which could be understood better from the table below

PC1	PC2	PC3	PC4	PC5	PC6
No_HH	F_LIT	MAIN_AL_M	TOT_WORK_F	M_SC	MAIN_HH_M
TOT_M	MAINWORK_M	MARG_CL_F	MAINWORK_F	F_SC	MAIN_HH_F
TOT_F	MAIN_OT_M	MARG_AL_M	MAIN_CL_F	M_ST	MARG_HH_F
M_06	MAIN_OT_F	MARG_AL_F	MAIN_AL_F	F_ST	MARG_OT_3_6_F
F_06	MARG_CL_M	MARG_AL_3_6_F	MARG_HH_M	MAIN_CL_M	MARG_OT_0_3_F
M_LIT	MARG_AL_3_6_M	MARG_HH_3_6_M	MARG_OT_3_6_M	MARG_OT_F	
M_ILL	MARG_CL_0_3_M	MARG_HH_3_6_F	MARG_OT_0_3_M	NON_WORK_M	
F_ILL	MARG_CL_0_3_F	MARG_AL_0_3_M		NON_WORK_F	
TOT_WORK_M	MARG_HH_0_3_M	MARG_AL_0_3_F			
MARGWORK_M		MARG_HH_0_3_F			
MARGWORK_F					
MARG_OT_M					
MARGWORK_3_6_M					
MARGWORK_3_6_F					
MARG_CL_3_6_M					
MARG_CL_3_6_F					
MARGWORK_0_3_M					
MARGWORK_0_3_F					

Table 25: Features highest explaining PCs

2. PC1 captures most variance for 19 features and most of these features are related to overall male and female population demographics like total number of households, total male and female population, total male and female population between 0 and 6 years and so on. We can name this as PC_Population_demography as it explains key population questions like total population, literacy and illiteracy in population, sex ratio etc.
3. PC2 explains most variance for columns F_LIT, MAINWORK_M, MAIN_OT_M, MAIN_OT_F, MARG_CL_M, MARG_AL_3_6_M, MARG_CL_0_3_M, MARG_CL_0_3_F, MARG_HH_0_3_M, most of the features in this PC is related to working male population so we can name it as PC_Male_Workers.
4. PC3 has explain most variance for columns related to employment in farming sector providing information regarding main and marginal employment in farming sector, we can name it as PC_Farming_Workforce.
5. In PC4 most of the features are about female employment covering total working female, main working female etc so we can name it as PC_Female_Workers
6. PC5 captures variance for backward classes that is for SC and ST population so we can name it as PC_Backward_classes.
7. PC6 which captures the lowest variance in 6 PC's covers mostly columns related to workers in household industries so we can name it as PC_household_industries.