Machine Learning -1 Project

Table of Contents:

List of Tables

List of Figures

Problem 1

Problem Statement:

Clustering:

Digital Ads Data:

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

The following three features are commonly used in digital marketing:

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

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

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

The Data Dictionary and the detailed description of the formulas for CPM, CPC and CTR are given in the sheet 2 of the Clustering Clean ads_data Excel File.

Perform the following in given order:

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

Problem 1 - Define the problem and perform Exploratory Data Analysis

Checking first 5 rows using head function

Timestamp	InventoryType	Ad - Length	Ad-Width	Ad Size	Ad Type	Platform	Device Type	Format	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
2020-9-2-17	Format1	300	250	75000	Inter222	Video	Desktop	Display	1806	325	323	1	0.0	0.35	0.0	0.0031	0.0	0.0
2020-9-2-10	Format1	300	250	75000	Inter227	App	Mobile	Video	1780	285	285	1	0.0	0.35	0.0	0.0035	0.0	0.0
2020-9-1-22	Format1	300	250	75000	Inter222	Video	Desktop	Display	2727	356	355	1	0.0	0.35	0.0	0.0028	0.0	0.0
2020-9-3-20	Format1	300	250	75000	Inter228	Video	Mobile	Video	2430	497	495	1	0.0	0.35	0.0	0.0020	0.0	0.0
2020-9-4-15	Format1	300	250	75000	Inter217	Web	Desktop	Video	1218	242	242	1	0.0	0.35	0.0	0.0041	0.0	0.0

Table1: Top 5 rows data

Checking last 5 rows using tail function

Timestamp	InventoryType	Ad - Length	Ad-Width	Ad Size	Ad Type	Platform	Device Type	Format	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
2020-9-13-7	Format5	720	300	216000	Inter220	Web	Mobile	Video	1	1	1	1	0.07	0.35	0.0455	NaN	NaN	NaN
2020-11-2-7	Format5	720	300	216000	Inter224	Web	Desktop	Video	3	2	2	1	0.04	0.35	0.0260	NaN	NaN	NaN
2020-9-14-22	Format5	720	300	216000	Inter218	App	Mobile	Video	2	1	1	1	0.05	0.35	0.0325	NaN	NaN	NaN
2020-11-18-2	Format4	120	600	72000	inter230	Video	Mobile	Video	7	1	1	1	0.07	0.35	0.0455	NaN	NaN	NaN
2020-9-14-0	Format5	720	300	216000	Inter221	App	Mobile	Video	2	2	2	1	0.09	0.35	0.0585	NaN	NaN	NaN

Table2: Last 5 rows data

Shape of the dataset

(23066, 19)

There are 23066 rows and 19 columns in the given dataset.

Data types and information of the dataset

```
<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
```

Table3: Data types and information

Insight

- There are 19 columns, 6 columns data types are float, 7 columns data types are integer and 6 columns data types are object.
- From the above table, we could see that columns CTR, CPM and CPC have the missing values. Will check them in the next steps.

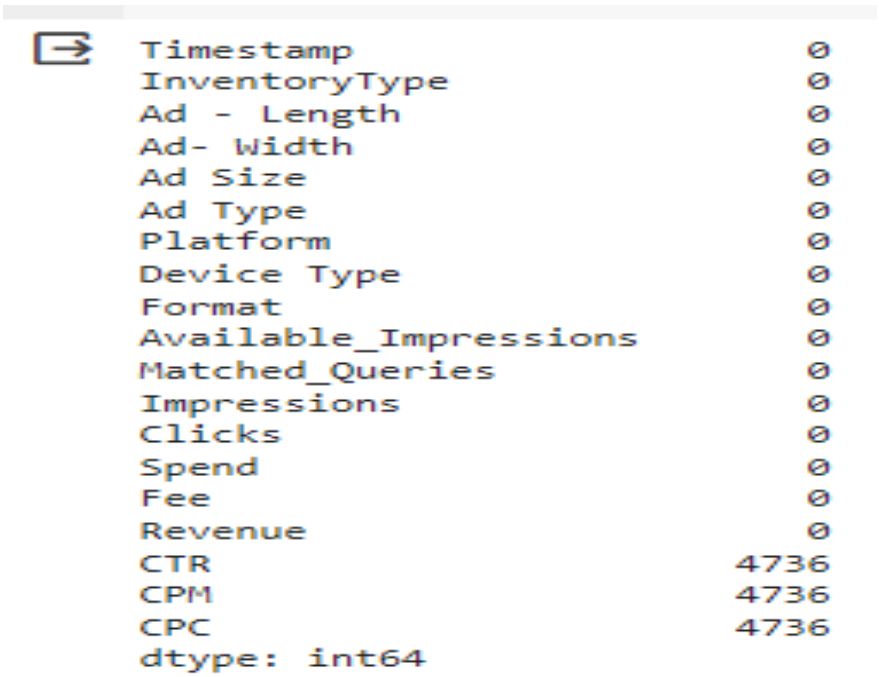
Statistical Summary

	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 4: Statistical summary

**Upon checking the duplicate rows, it was found that there are no duplicate rows.

Checking missing values, imputing the 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 5: Tables for missing values

Upon checking the missing values, there are 4736 missing values in column CTR, CPM and CPC. Let's impute them by given formulas.

CPM = (Spend/Impressions) * 1,000

CPC = Spend)/Clicks

CTR = Clicks / Impressions x 100

After applying the above formulas to impute the missing values, Let's check missing values and Statistical descriptive summary.

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

Table 6: Table for missing values after imputing.

	count	mean	std	min	25%	50%	75%	max
Ad - Length	23066.0	3.851631e+02	2.336514e+02	120.0000	120.000000	300.000000	7.200000e+02	728.00
Ad- Width	23066.0	3.378960e+02	2.030929e+02	70.0000	250.000000	300.000000	6.000000e+02	600.00
Ad Size	23066.0	9.667447e+04	6.153833e+04	33600.0000	72000.000000	72000.000000	8.400000e+04	216000.00
Available_Impressions	23066.0	2.432044e+06	4.742888e+06	1.0000	33672.250000	483771.000000	2.527712e+06	27592861.00
Matched_Queries	23066.0	1.295099e+06	2.512970e+06	1.0000	18282.500000	258087.500000	1.180700e+06	14702025.00
Impressions	23066.0	1.241520e+06	2.429400e+06	1.0000	7990.500000	225290.000000	1.112428e+06	14194774.00
Clicks	23066.0	1.067852e+04	1.735341e+04	1.0000	710.000000	4425.000000	1.279375e+04	143049.00
Spend	23066.0	2.706626e+03	4.067927e+03	0.0000	85.180000	1425.125000	3.121400e+03	26931.87
Fee	23066.0	3.351231e-01	3.196322e-02	0.2100	0.330000	0.350000	3.500000e-01	0.35
Revenue	23066.0	1.924252e+03	3.105238e+03	0.0000	55.365375	926.335000	2.091338e+03	21276.18
CTR	23066.0	2.614863e+00	7.853405e+00	0.0001	0.003400	0.112650	1.837777e-01	200.00
CPM	23066.0	8.396730e+00	9.057082e+00	0.0000	1.750000	8.370742	1.304000e+01	715.00
CPC	23066.0	3.366523e-01	3.412311e-01	0.0000	0.090000	0.140000	5.500000e-01	7.26

Table 7: Statistical summary after imputing missing values.

Univariate analysis

To perform the univariate analysis, we segregate the numerical columns and store them in a new dataframe.

Let's check the head of numerical column's dataframe.

	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
0	300	250	75000	1806	325	323	1	0.0	0.35	0.0	0.0031	0.0	0.0
1	300	250	75000	1780	285	285	1	0.0	0.35	0.0	0.0035	0.0	0.0
2	300	250	75000	2727	356	355	1	0.0	0.35	0.0	0.0028	0.0	0.0
3	300	250	75000	2430	497	495	1	0.0	0.35	0.0	0.0020	0.0	0.0
4	300	250	75000	1218	242	242	1	0.0	0.35	0.0	0.0041	0.0	0.0

Table 8: Top 5 rows of subset of dataset.

```
Description of Ad - Length
-------------------------------------------------------------------------------
--
count    23066.000000
mean       385.163097
std        233.651434
min        120.000000
25%        120.000000
50%        300.000000
75%        720.000000
max        728.000000
Name: Ad - Length, dtype: float64 Distribution of Ad - Length
```

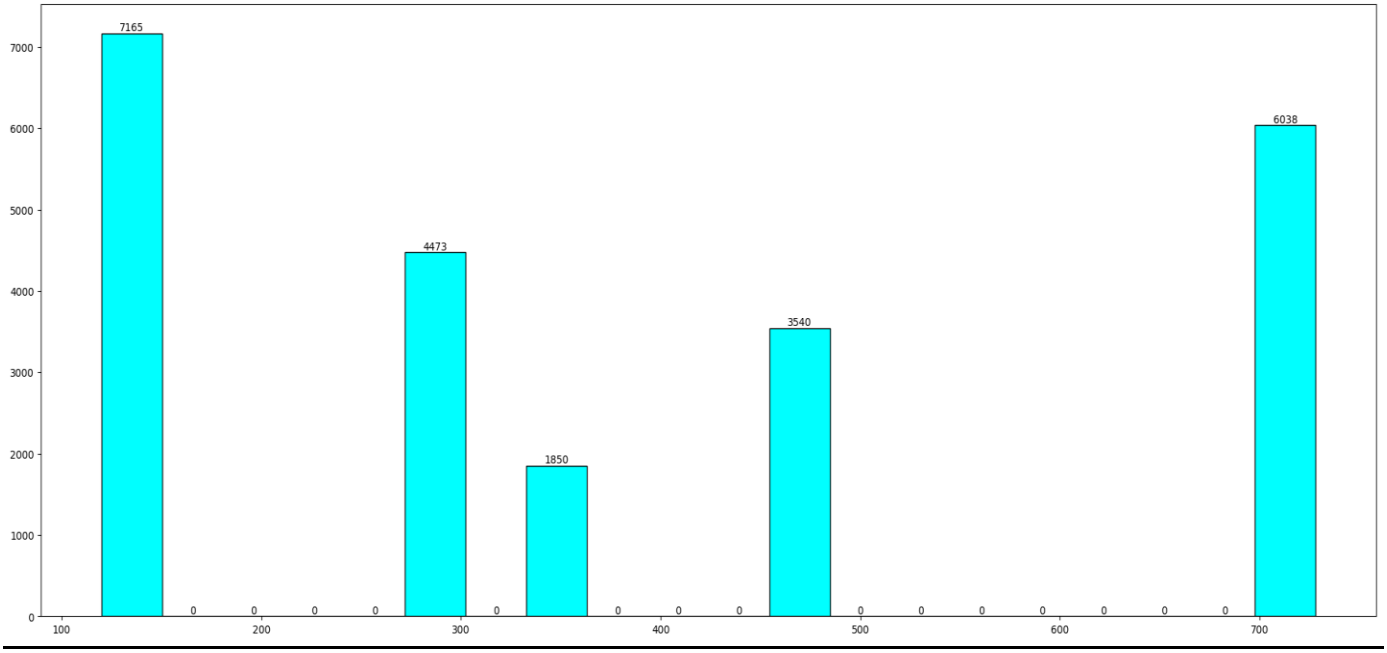


Figure 1: Plot of Ad-Length

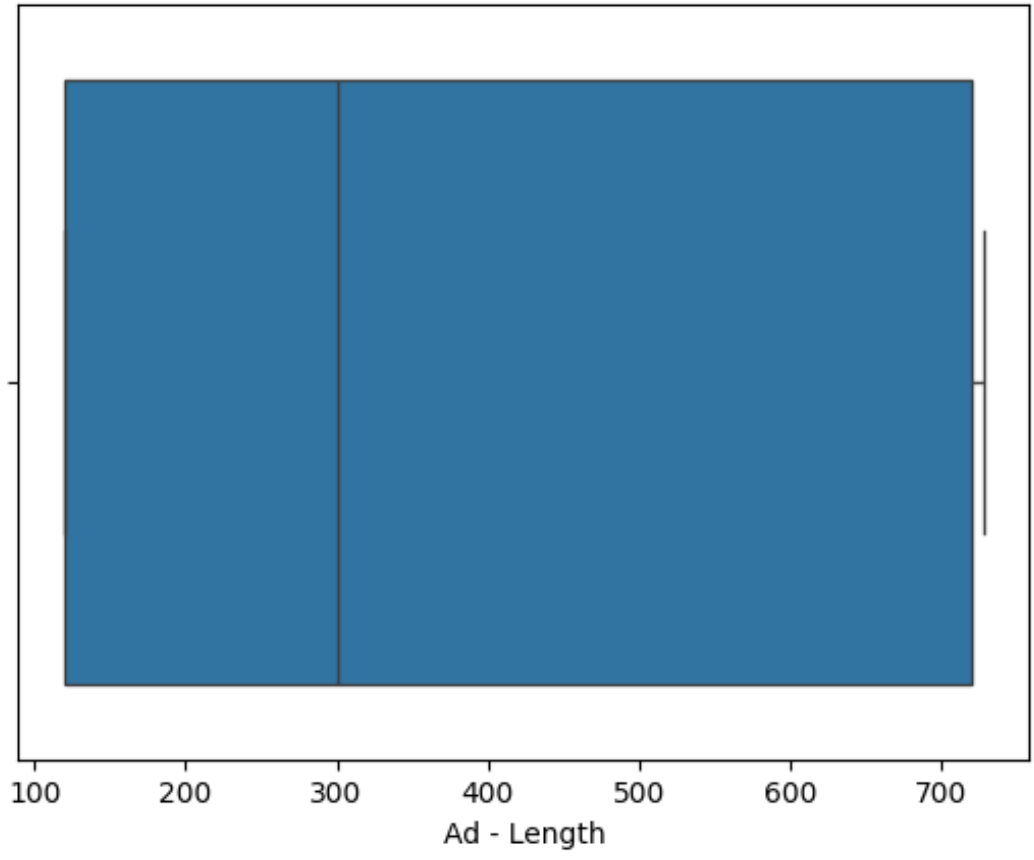


Figure 2: Boxplot of Ad-Length

Description of Ad- Width

```
--------------------------------------------------------------------------------
--
count    23066.000000
mean       337.896037
std        203.092885
min         70.000000
25%        250.000000
50%        300.000000
75%        600.000000
max        600.000000
```

Name: Ad- Width, dtype: float64 Distribution of Ad- Width

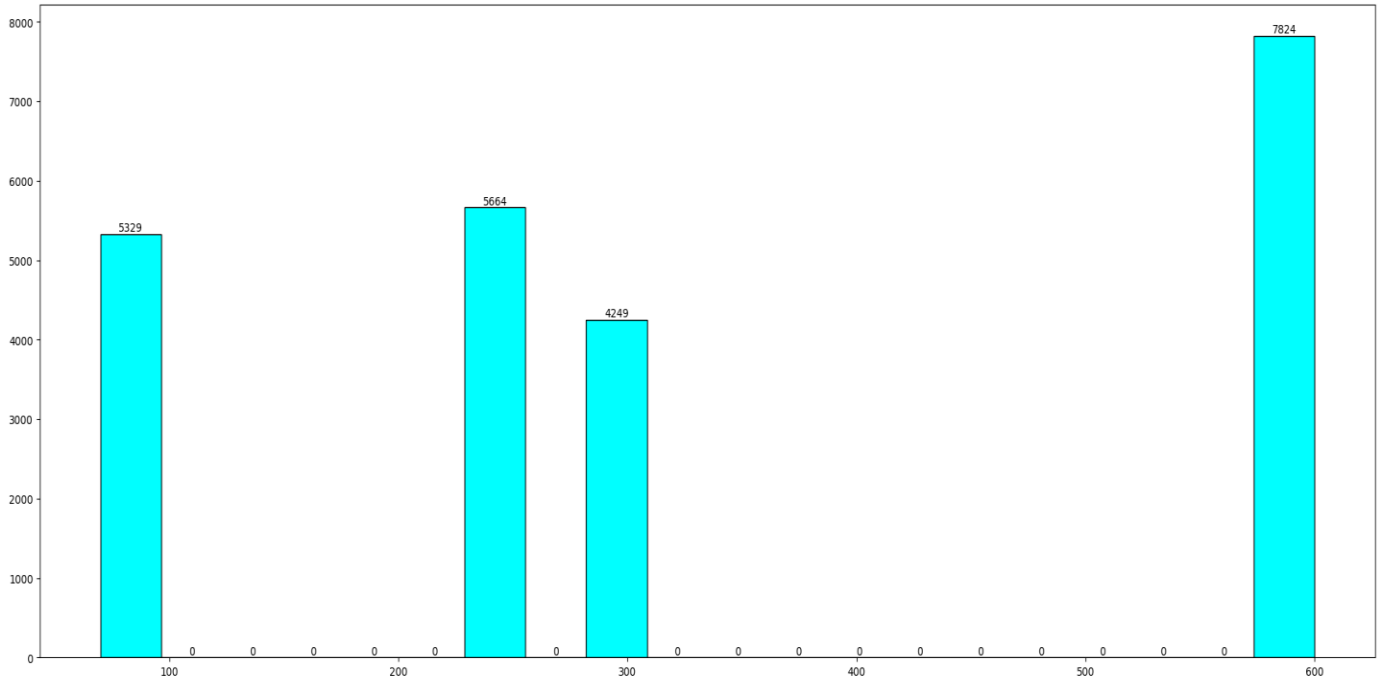


Figure 3: plot of Ad-width

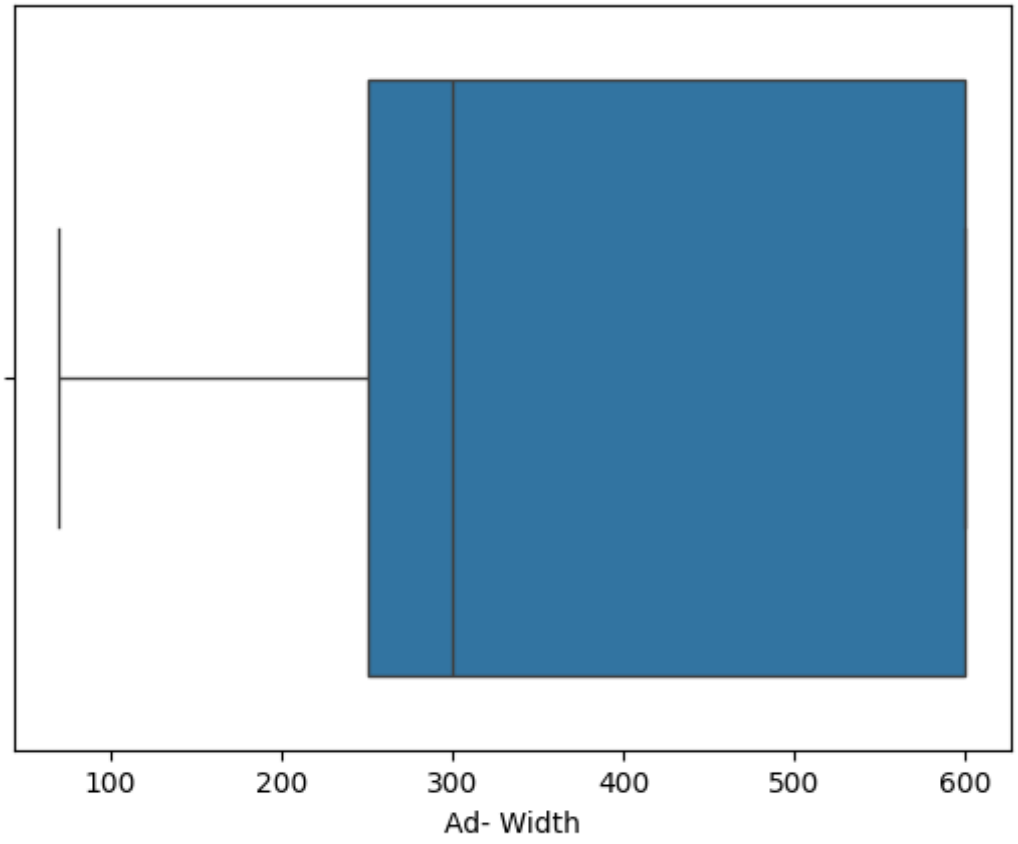


Figure 4: Boxplot of Ad-width

Description of Ad Size

```
--------------------------------------------------------------------------------
--
count     23066.000000
mean      96674.468048
std       61538.329557
min       33600.000000
25%       72000.000000
50%       72000.000000
75%       84000.000000
max      216000.000000
```

Name: Ad Size, dtype: float64 Distribution of Ad Size

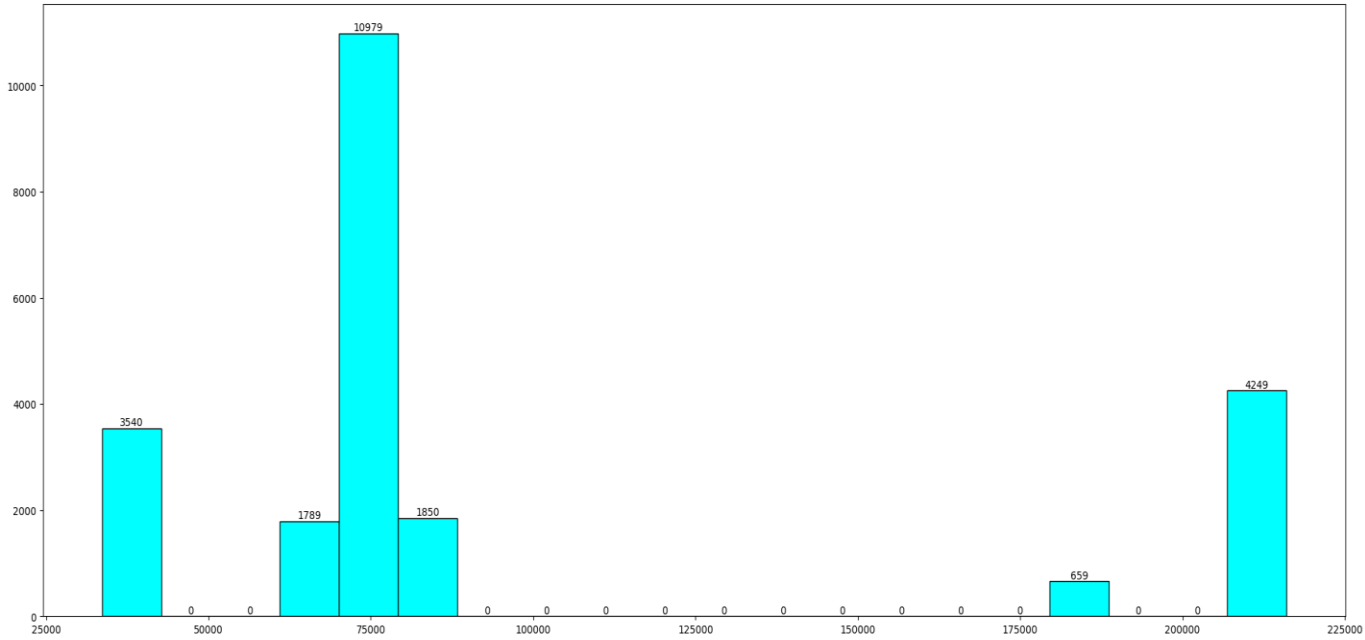


Figure 5: plot of Ad-Size

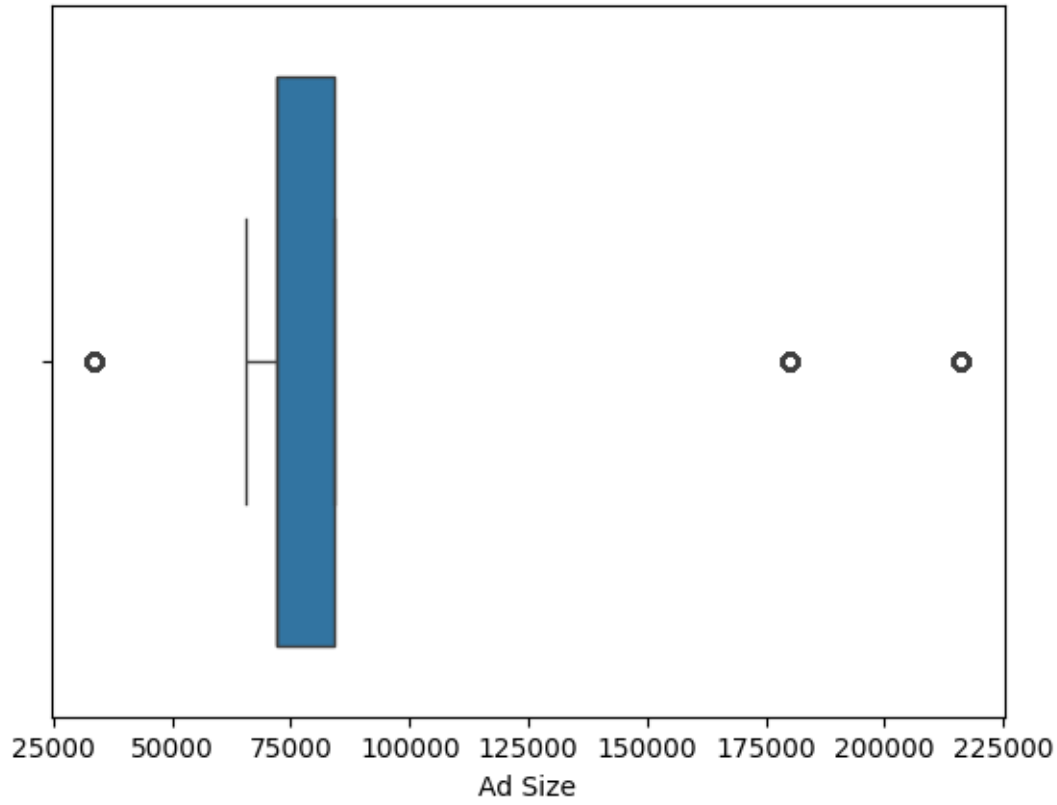


Figure 6: Boxplot of Ad-Size

```
Description of Available_Impressions
```

```
--------------------------------------------------------------------------------
--
count    2.306600e+04
mean     2.432044e+06
std      4.742888e+06
min      1.000000e+00
25%      3.367225e+04
50%      4.837710e+05
75%      2.527712e+06
max      2.759286e+07
```

Name: Available_Impressions, dtype: float64 Distribution of Available_Impressions

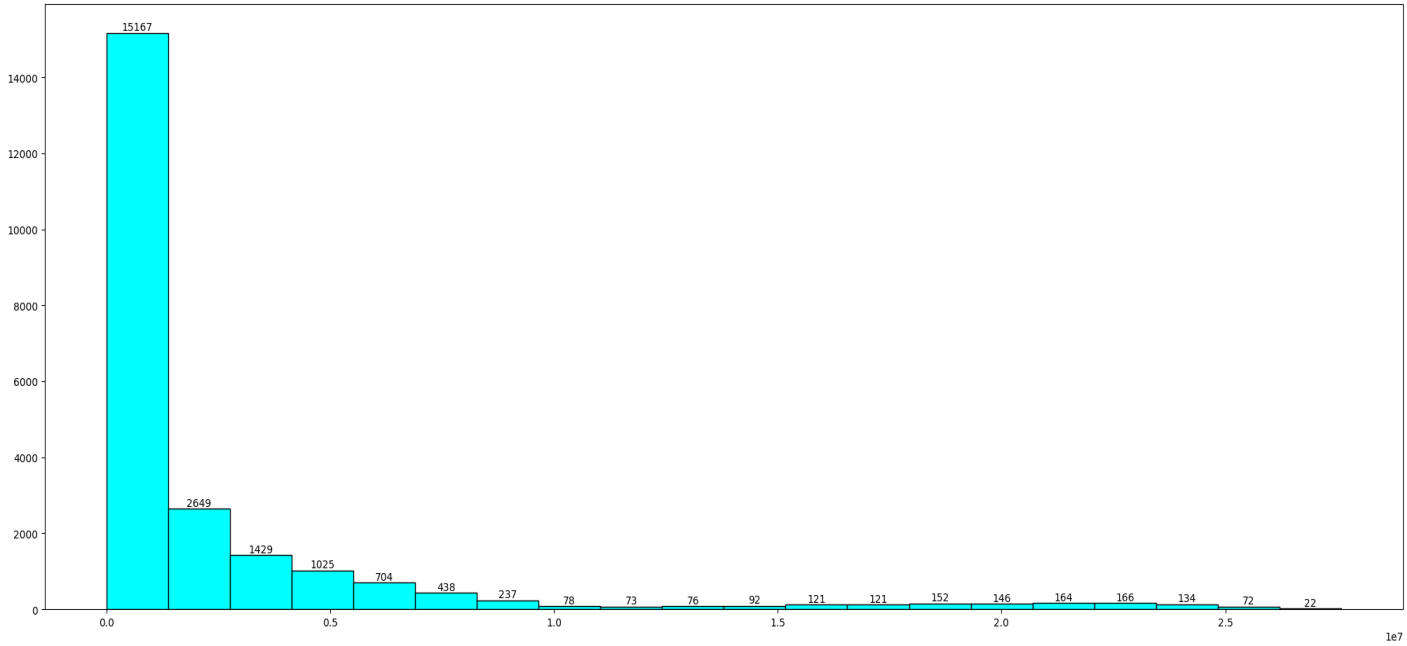


Figure 7: plot of Available Impressions

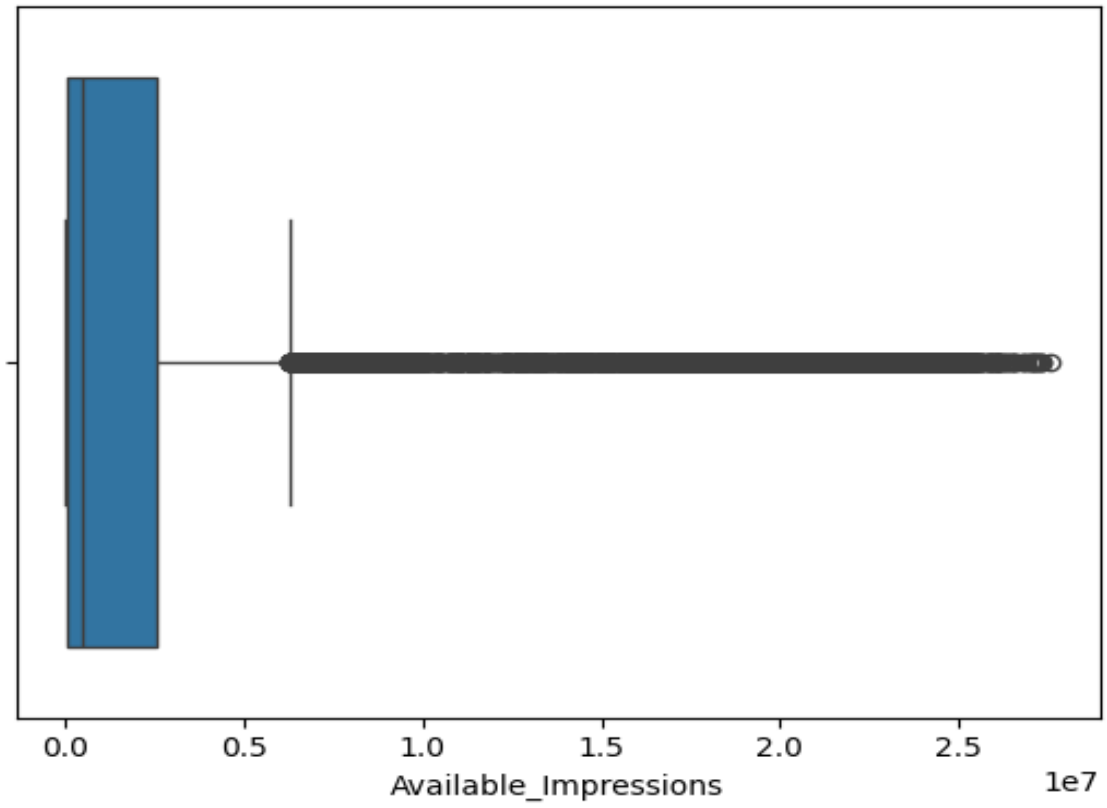


Figure 8: Boxplot of Available Impressions

Description of Matched_Queries

```
----------------------------------------------------------------------------
--
count    2.306600e+04
mean     1.295099e+06
std      2.512970e+06
min      1.000000e+00
25%      1.828250e+04
50%      2.580875e+05
75%      1.180700e+06
max      1.470202e+07
```

Name: Matched_Queries, dtype: float64 Distribution of Matched_Queries

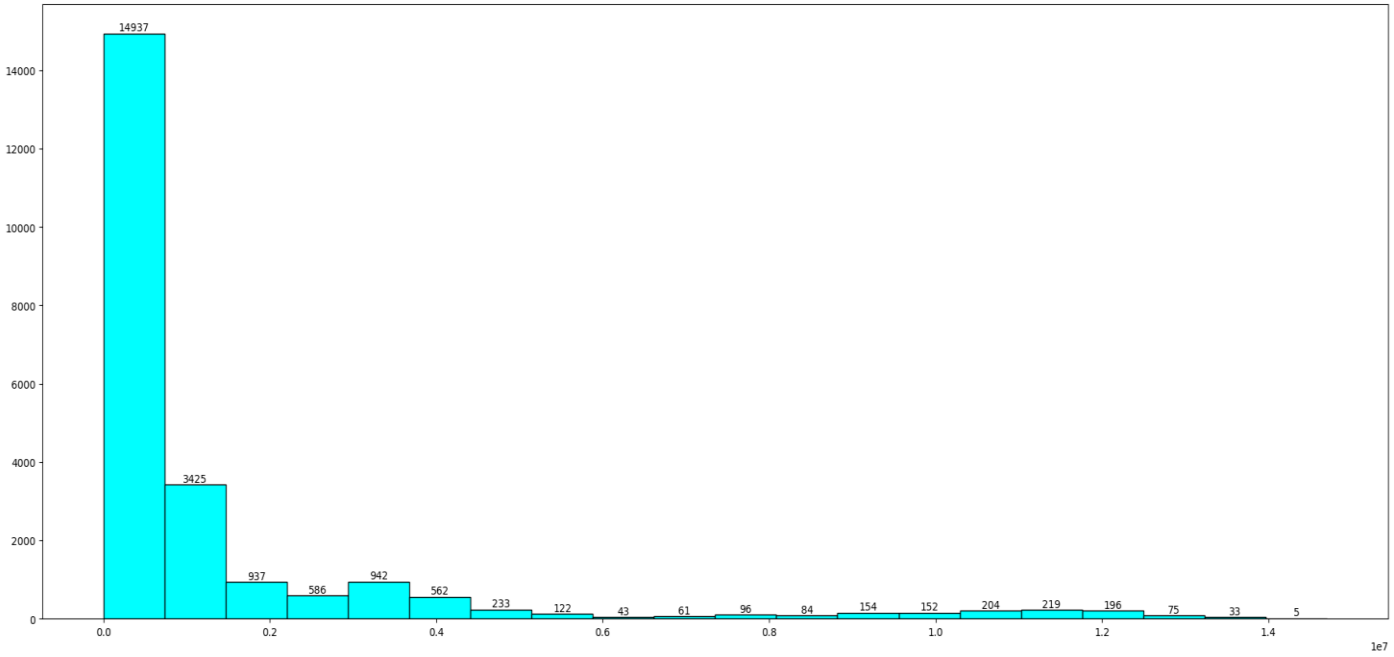


Figure 9: Plot of Matched queries

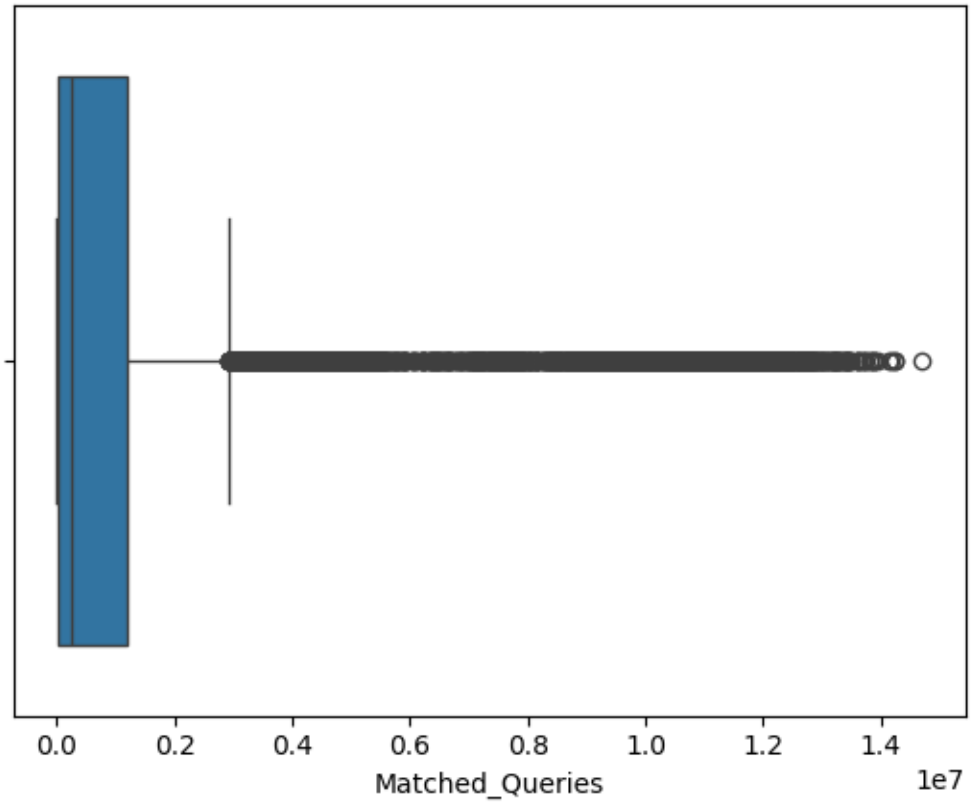


Figure 10: Boxplot of Matched queries

Description of Impressions

```
--------------------------------------------------------------------------------
--
count    2.306600e+04
mean     1.241520e+06
std      2.429400e+06
min      1.000000e+00
25%      7.990500e+03
50%      2.252900e+05
75%      1.112428e+06
max      1.419477e+07
```

Name: Impressions, dtype: float64 Distribution of Impressions

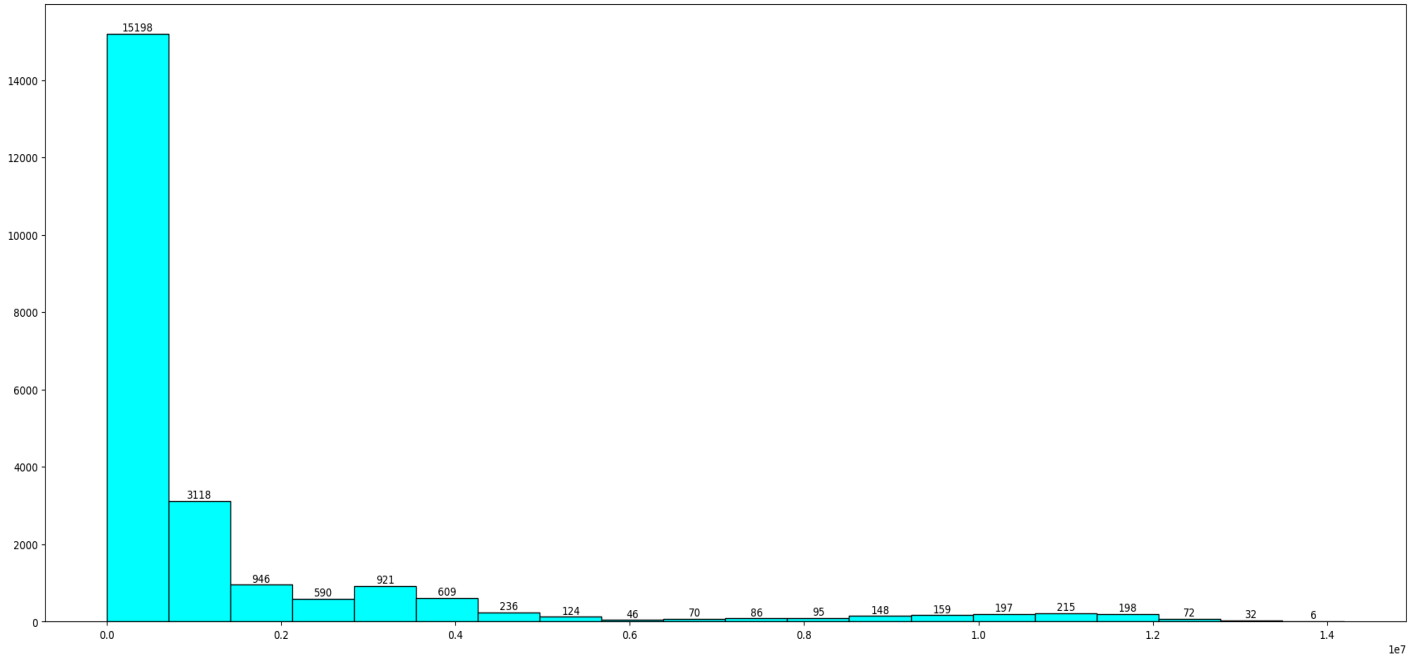


Figure 11: Plot of Impressions

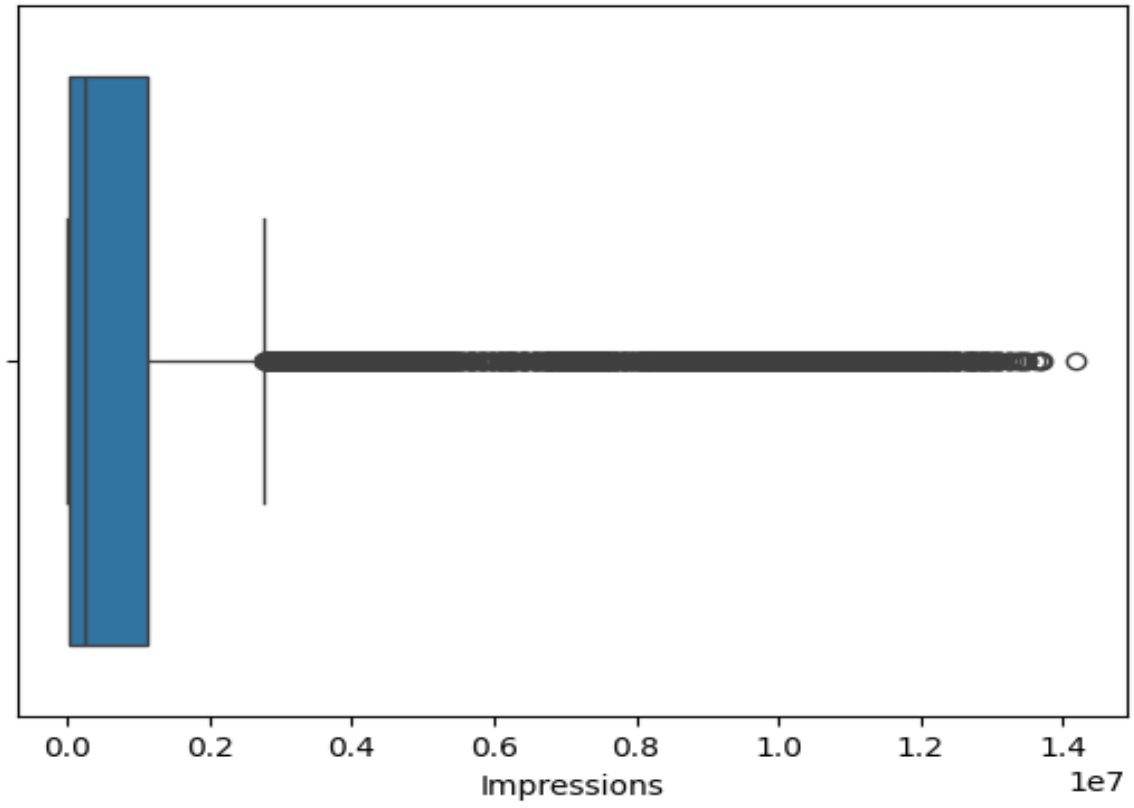


Figure 12: Boxplot of Impressions

Description of Clicks

```
----------------------------------------------------------------------------
--
count     23066.000000
mean      10678.518816
std       17353.409363
min           1.000000
25%         710.000000
50%        4425.000000
75%       12793.750000
max      143049.000000
```

Name: Clicks, dtype: float64 Distribution of Clicks

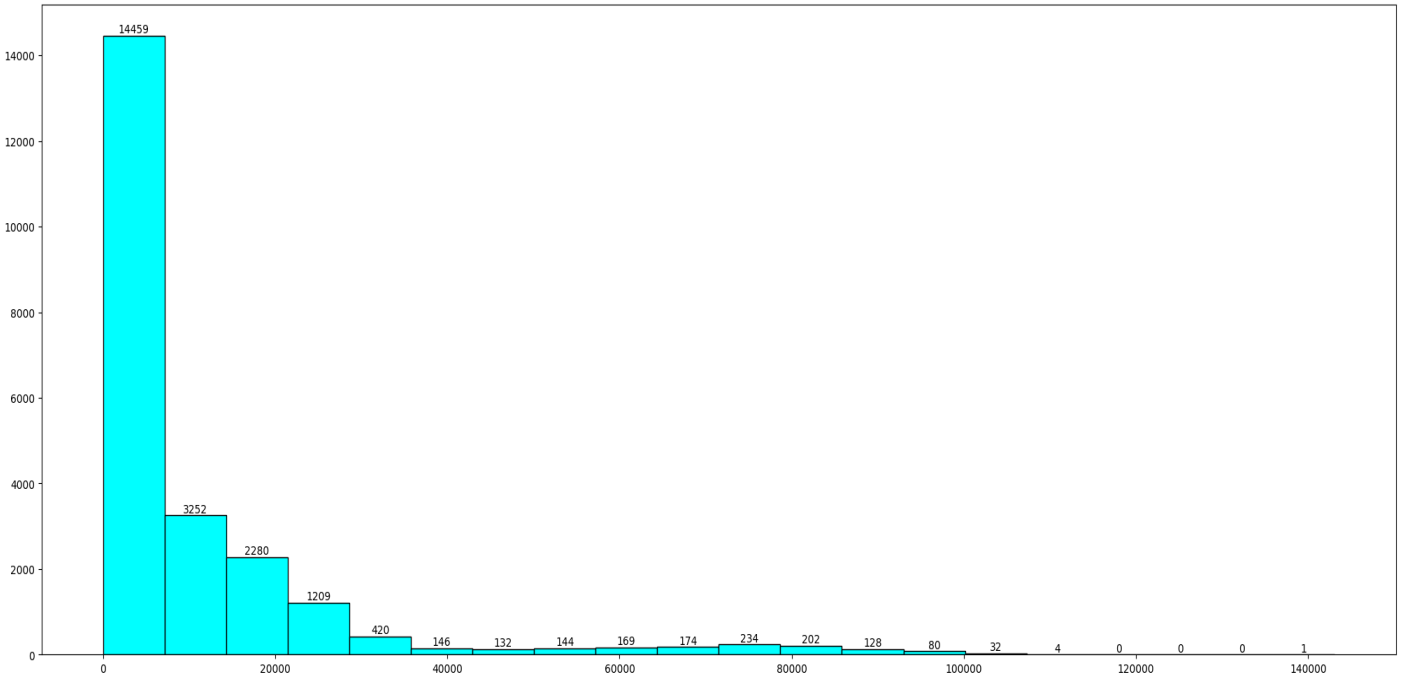


Figure 13: Plot of Clicks

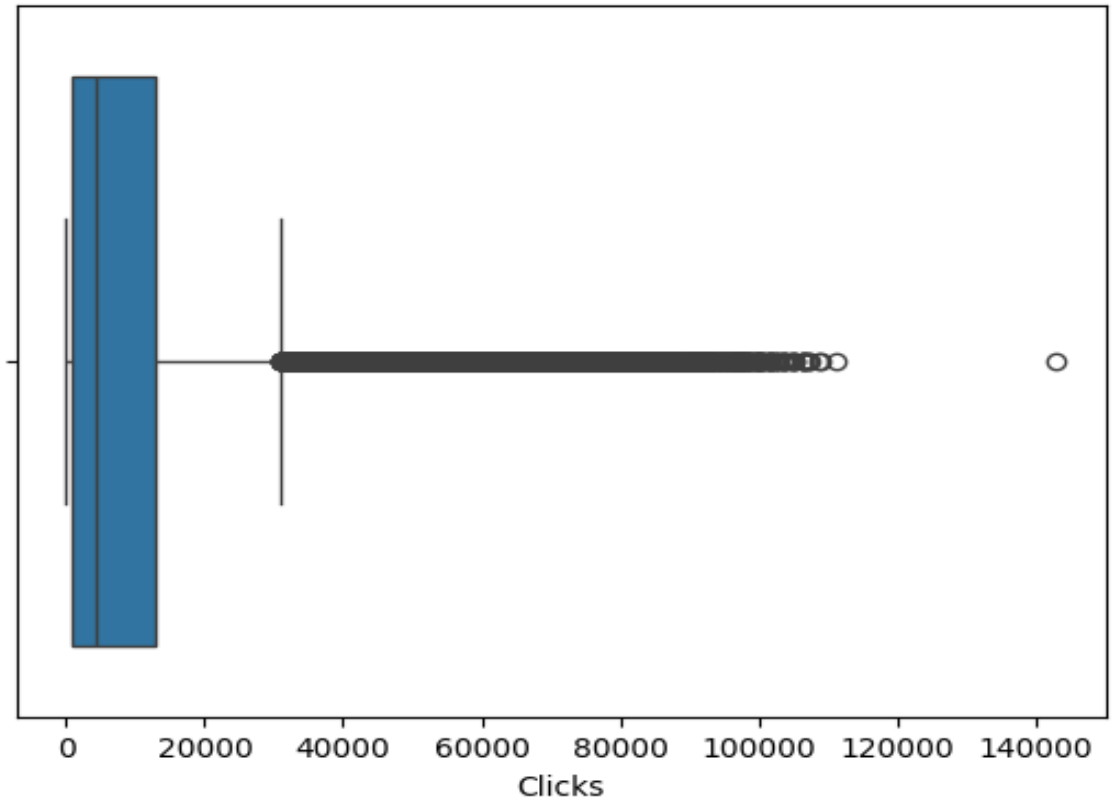


Figure 14: Boxplot of clicks

`Description of Spend`

```
--------------------------------------------------------------------------------
--
count    23066.000000
mean      2706.625689
std       4067.927273
min          0.000000
25%         85.180000
50%       1425.125000
75%       3121.400000
max      26931.870000
```

Name: Spend, dtype: float64 Distribution of Spend

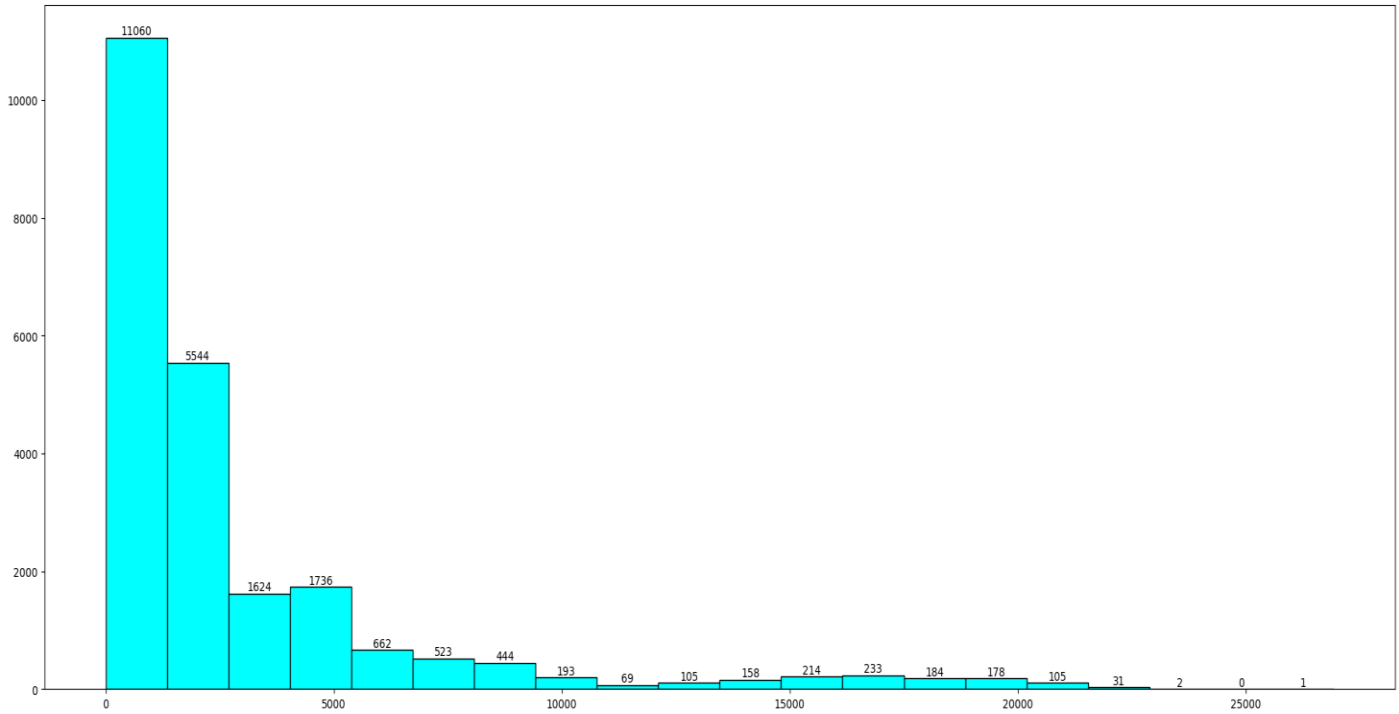


Figure 15: Plot of Spend

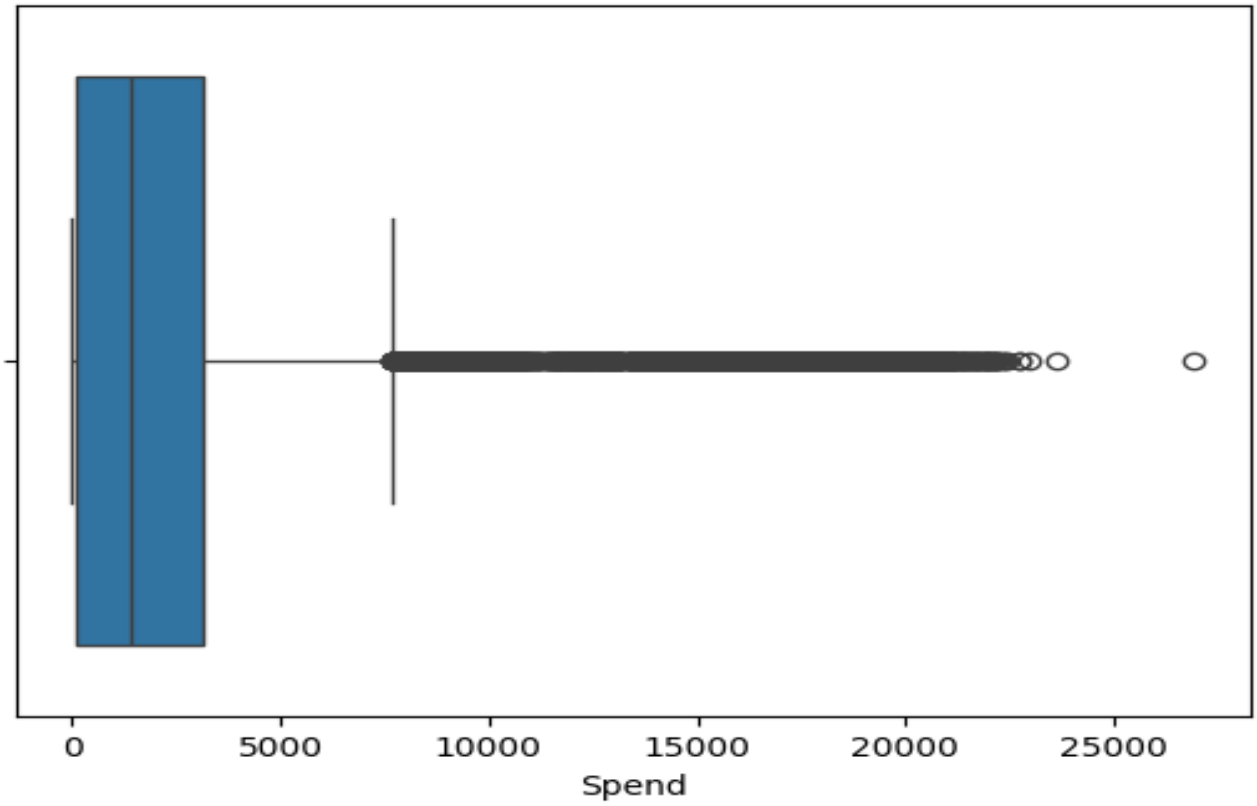


Figure 16: Boxplot of Spend

Description of Fee

```
----------------------------------------------------------------------------
--
count    23066.000000
mean         0.335123
std          0.031963
min          0.210000
25%          0.330000
50%          0.350000
75%          0.350000
max          0.350000
```

Name: Fee, dtype: float64 Distribution of Fee

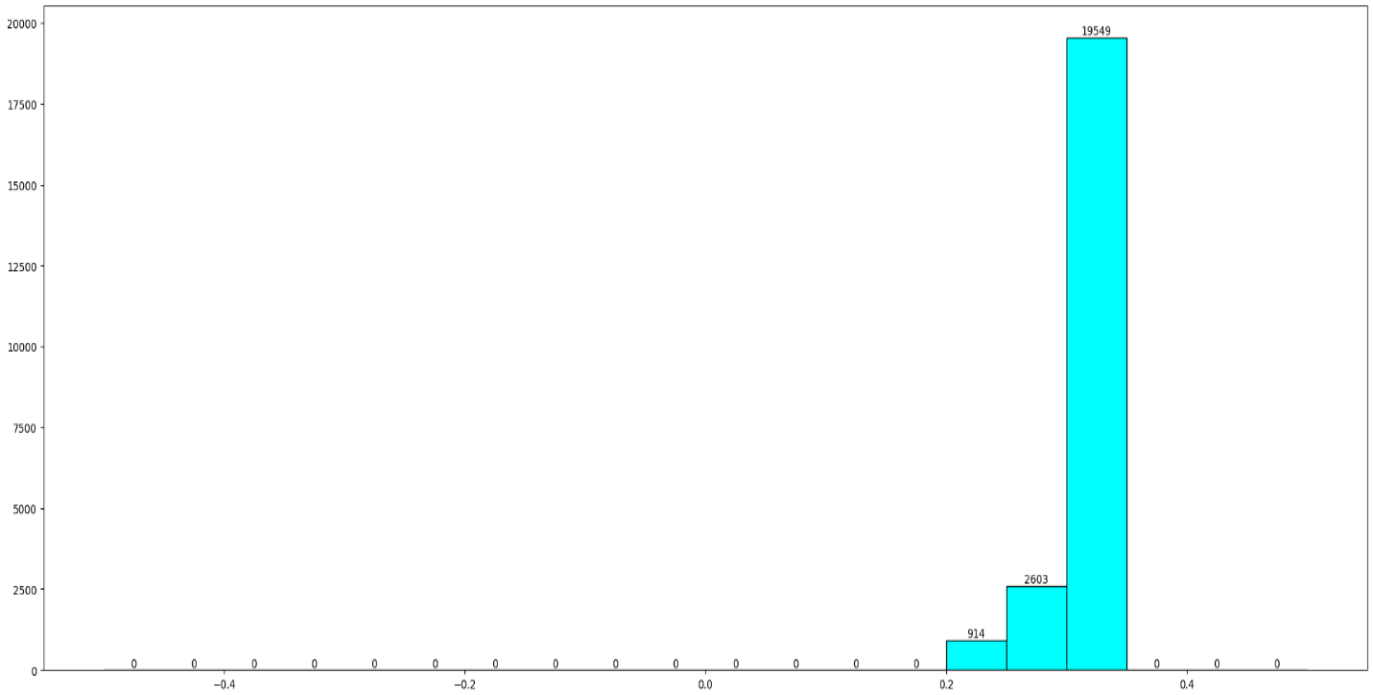


Figure 17: Plot of Fee

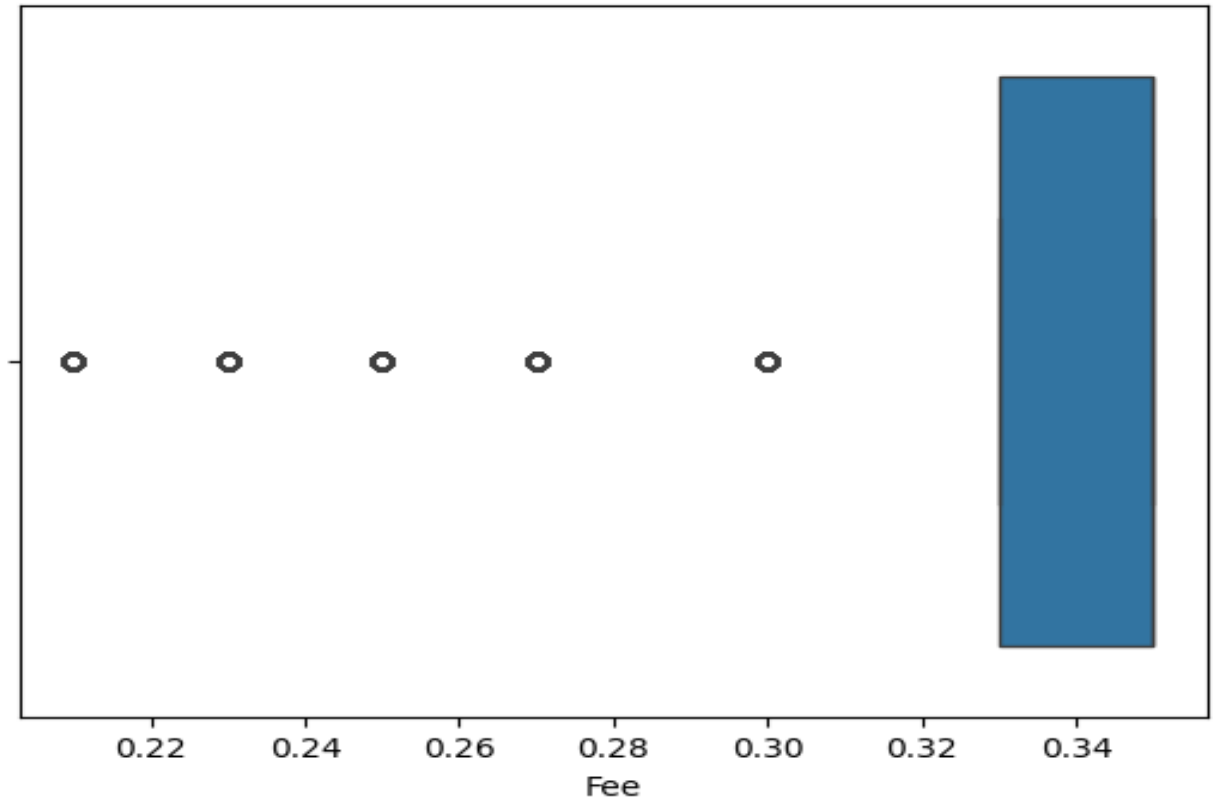


Figure 18: Boxplot of Fee

```
Description of Revenue
-------------------------------------------------------------------------------
--
count    23066.000000
mean      1924.252331
std       3105.238410
min          0.000000
25%         55.365375
50%        926.335000
75%       2091.338150
max      21276.180000

Name: Revenue, dtype: float64 Distribution of Revenue
```

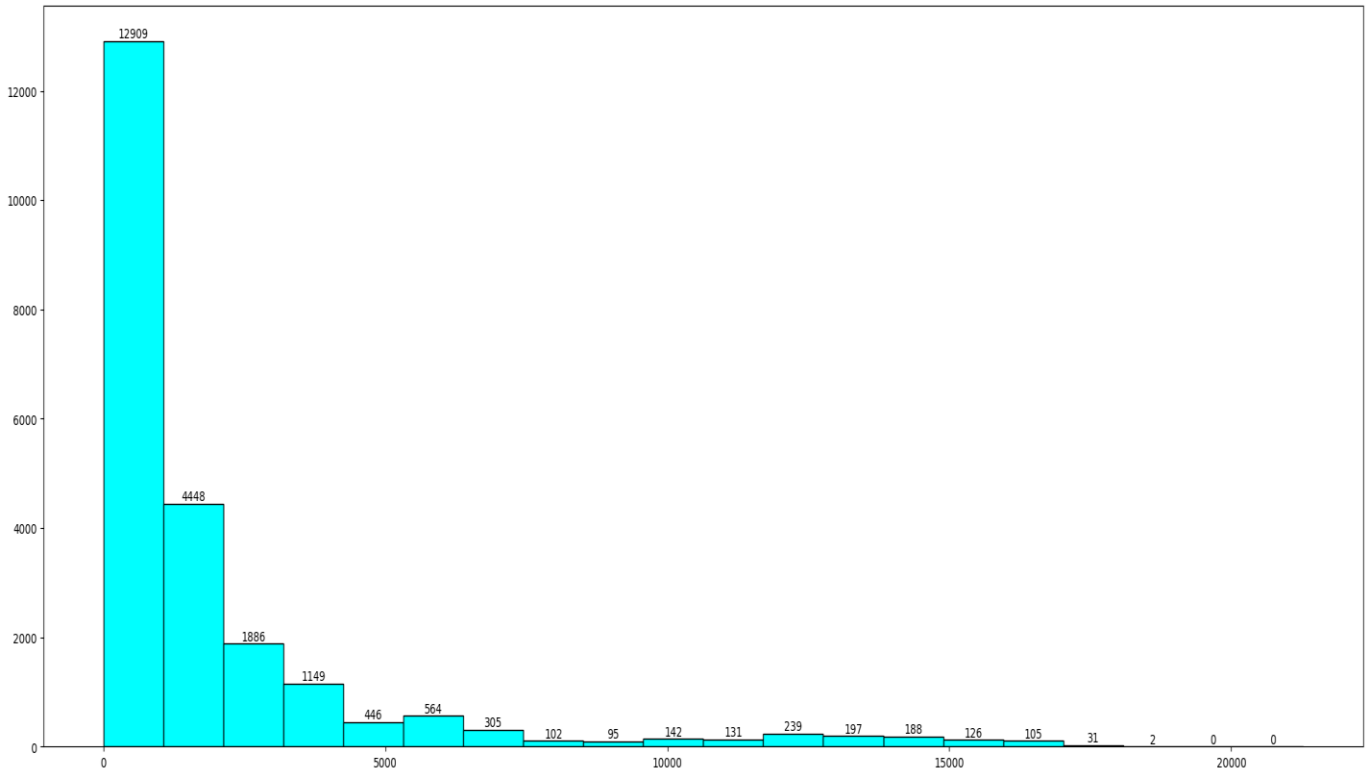


Figure 19: Plot of Revenue

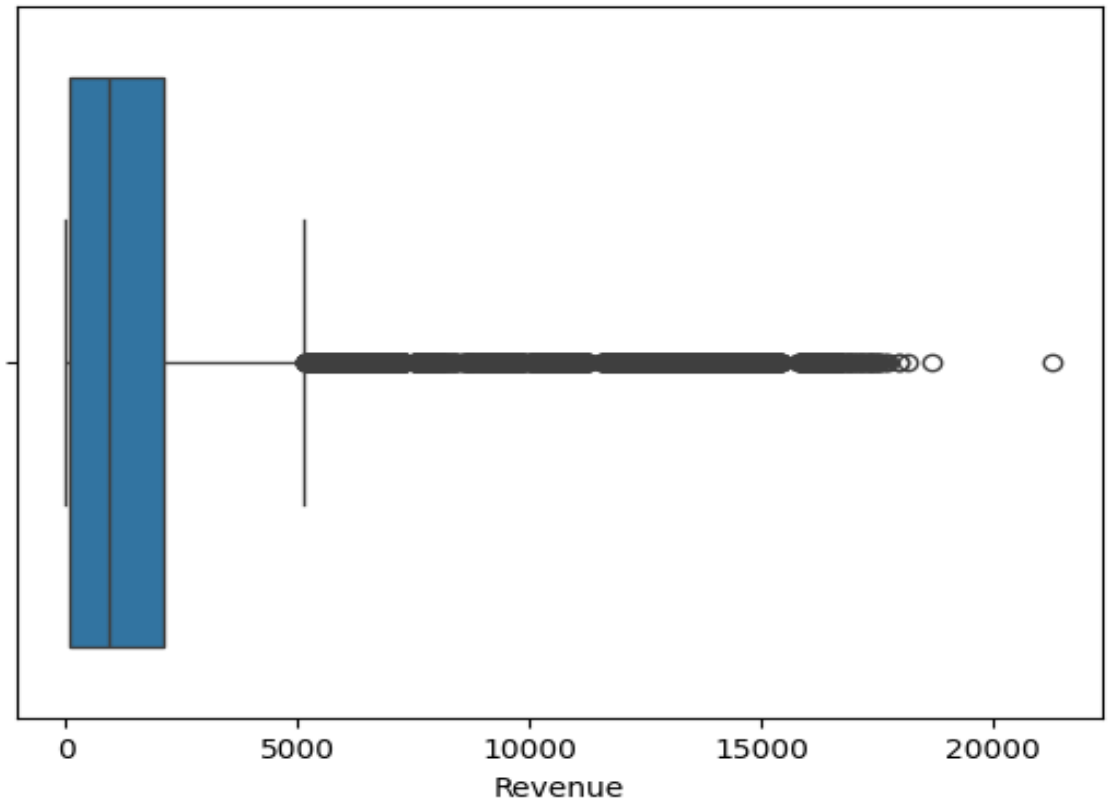


Figure 20: Boxplot of Revenue

Description of CTR

```
--------------------------------------------------------------------------------
--
count    23066.000000
mean         2.614863
std          7.853405
min          0.000100
25%          0.003400
50%          0.112650
75%          0.183778
max        200.000000
```

Name: CTR, dtype: float64 Distribution of CTR

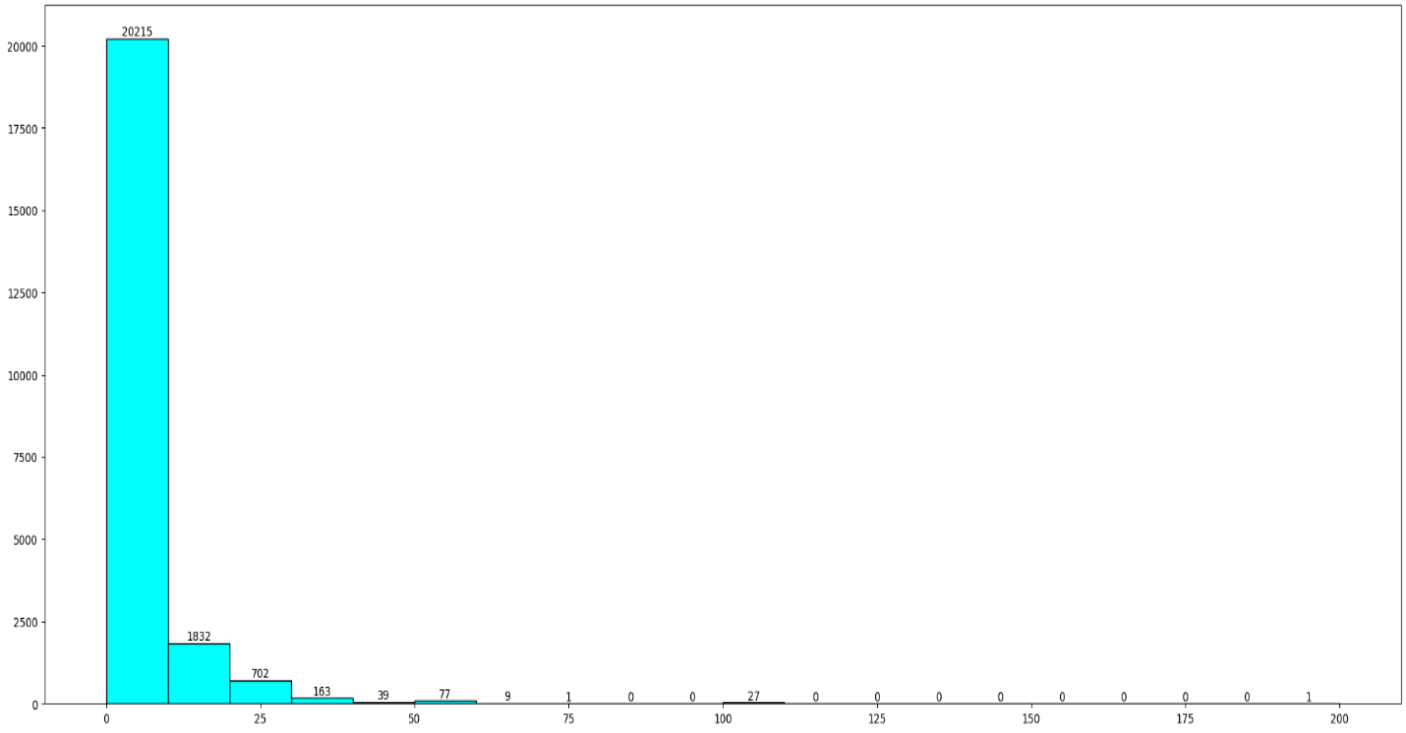


Figure 21: Plot of CTR

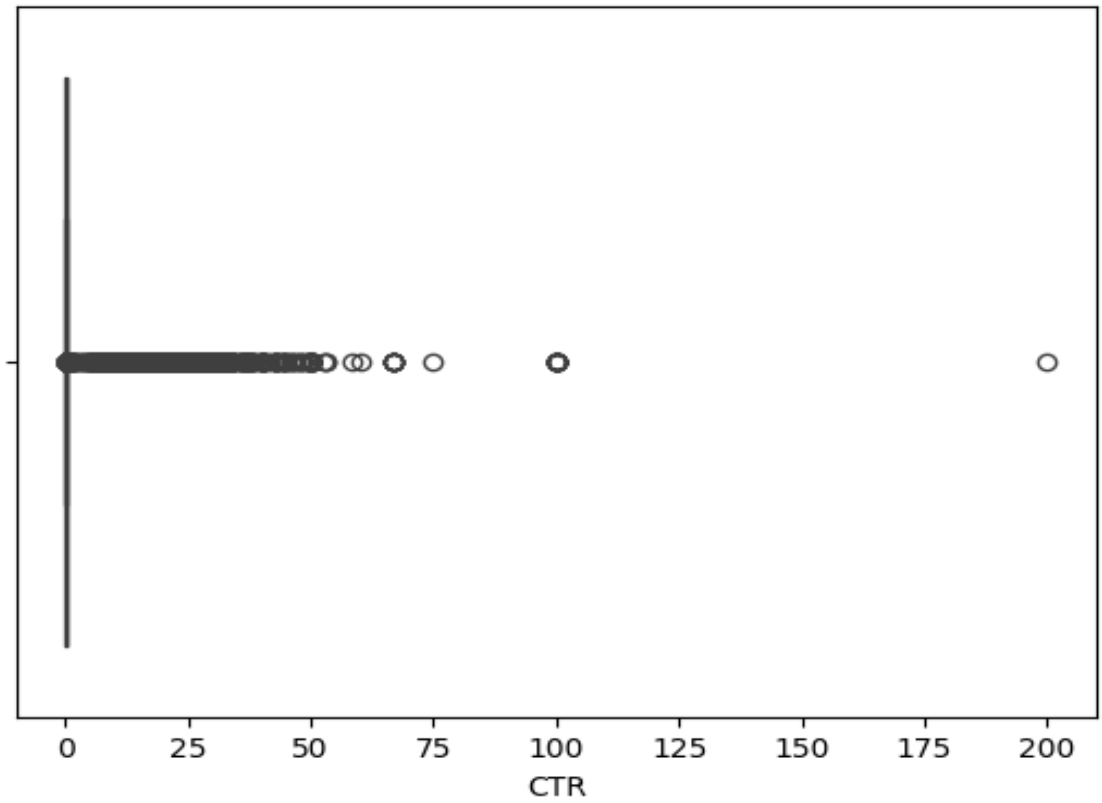


Figure 22: Boxplot of CTR

Description of CPM

```
--------------------------------------------------------------------------------
--
count    23066.000000
mean         8.396730
std          9.057082
min          0.000000
25%          1.750000
50%          8.370742
75%         13.040000
max        715.000000
```

Name: CPM, dtype: float64 Distribution of CPM

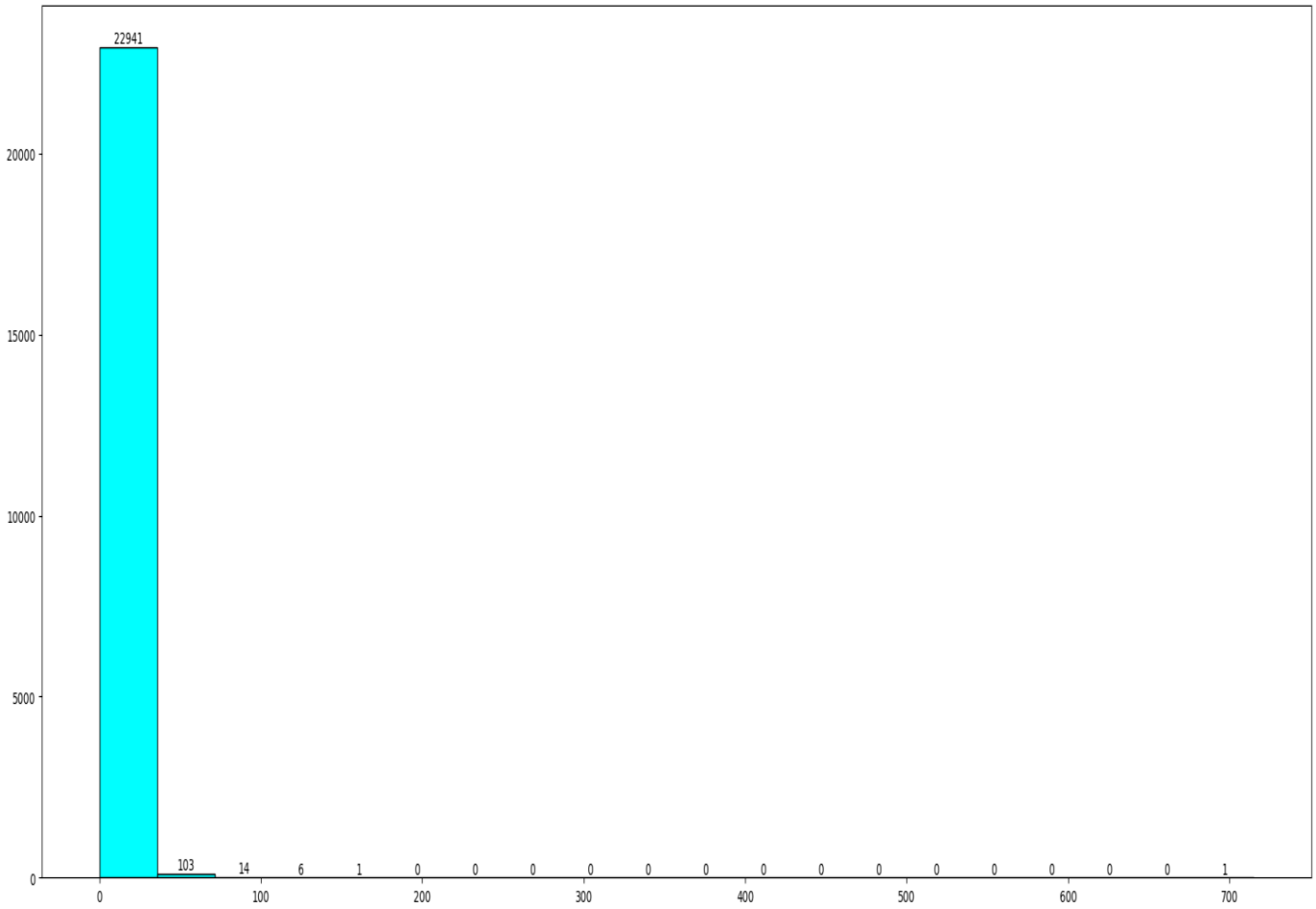


Figure 23: Plot of CPM

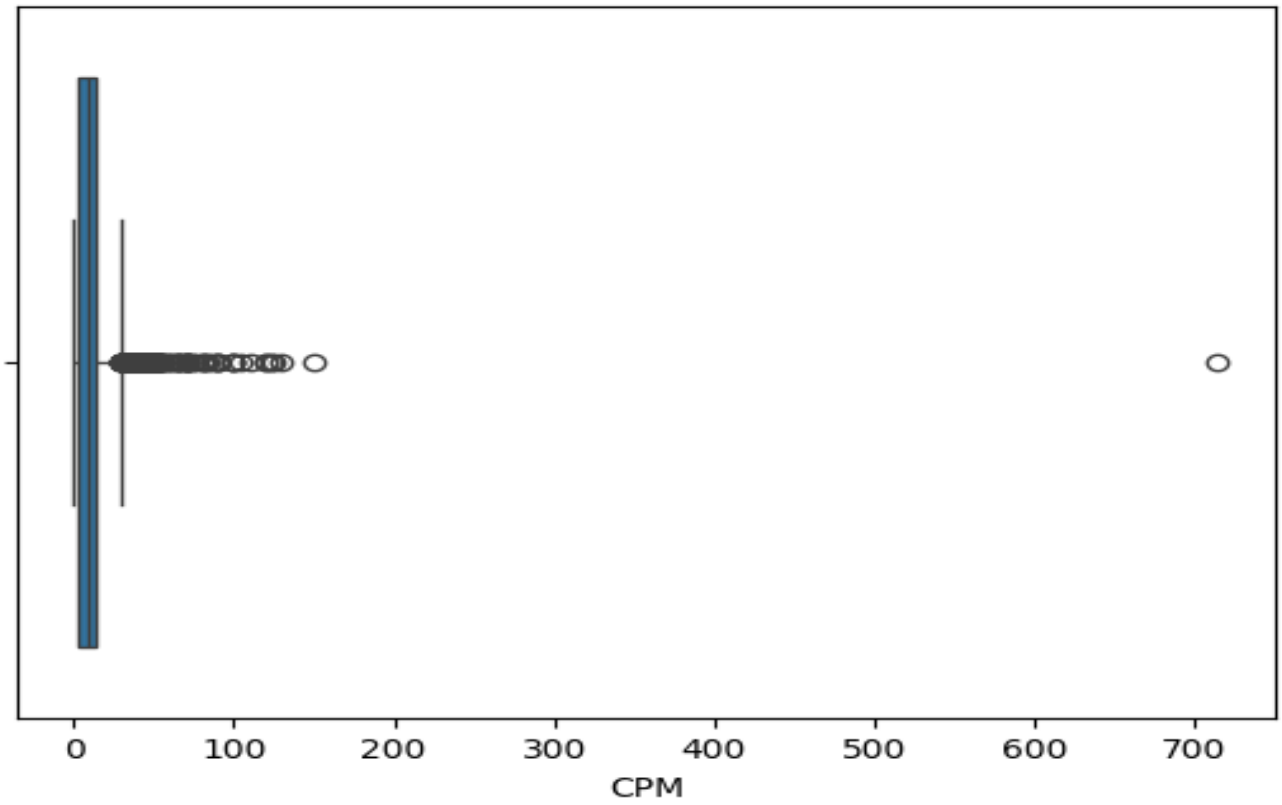


Figure 24: Boxplot of CPM

Description of CPC

```
----------------------------------------------------------------------------
--
count    23066.000000
mean         0.336652
std          0.341231
min          0.000000
25%          0.090000
50%          0.140000
75%          0.550000
max          7.260000
```

Name: CPC, dtype: float64 Distribution of CPC

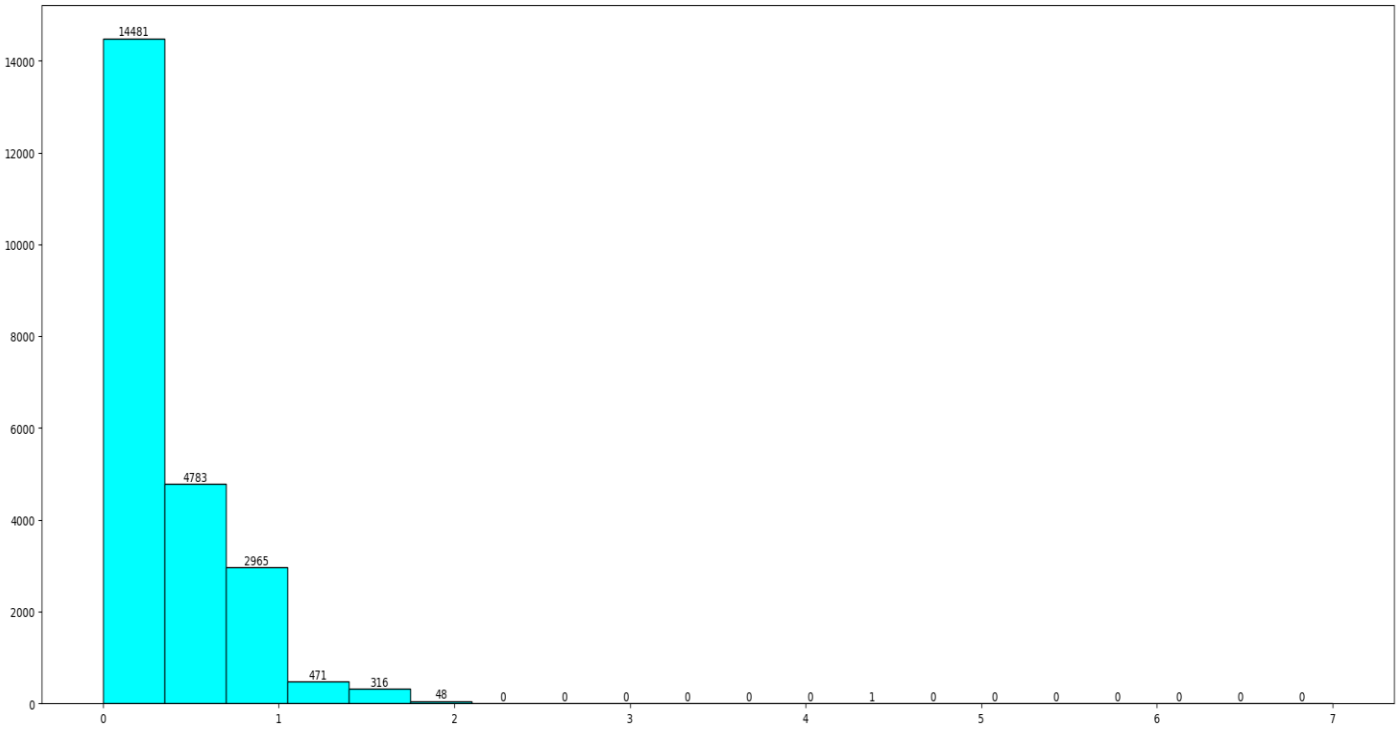


Figure 25: Plot of CPC

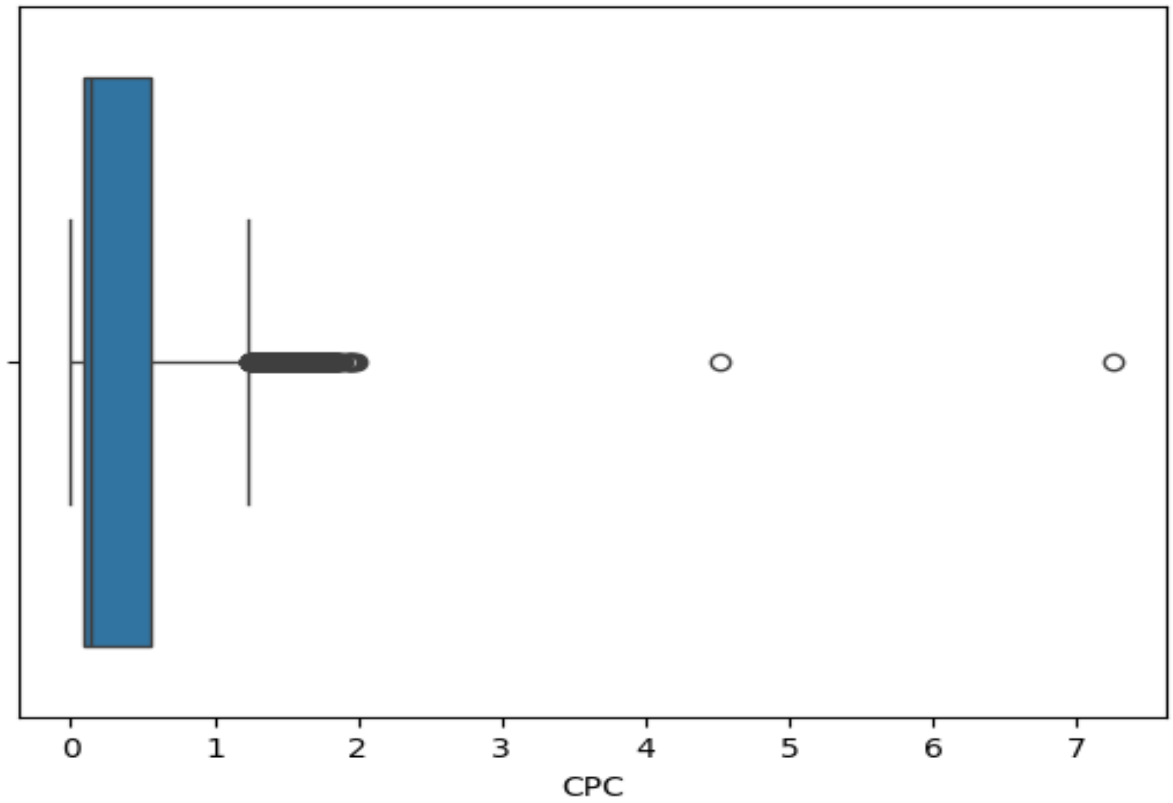


Figure 26: Boxplot of CPC

Univariate Categorical Analysis

Details of InventoryType

```
----------------------------------------------------------------
Format4    7165
Format5    4249
Format1    3814
Format3    3540
Format6    1850
Format2    1789
Format7     659
```

Name: InventoryType, dtype: int64

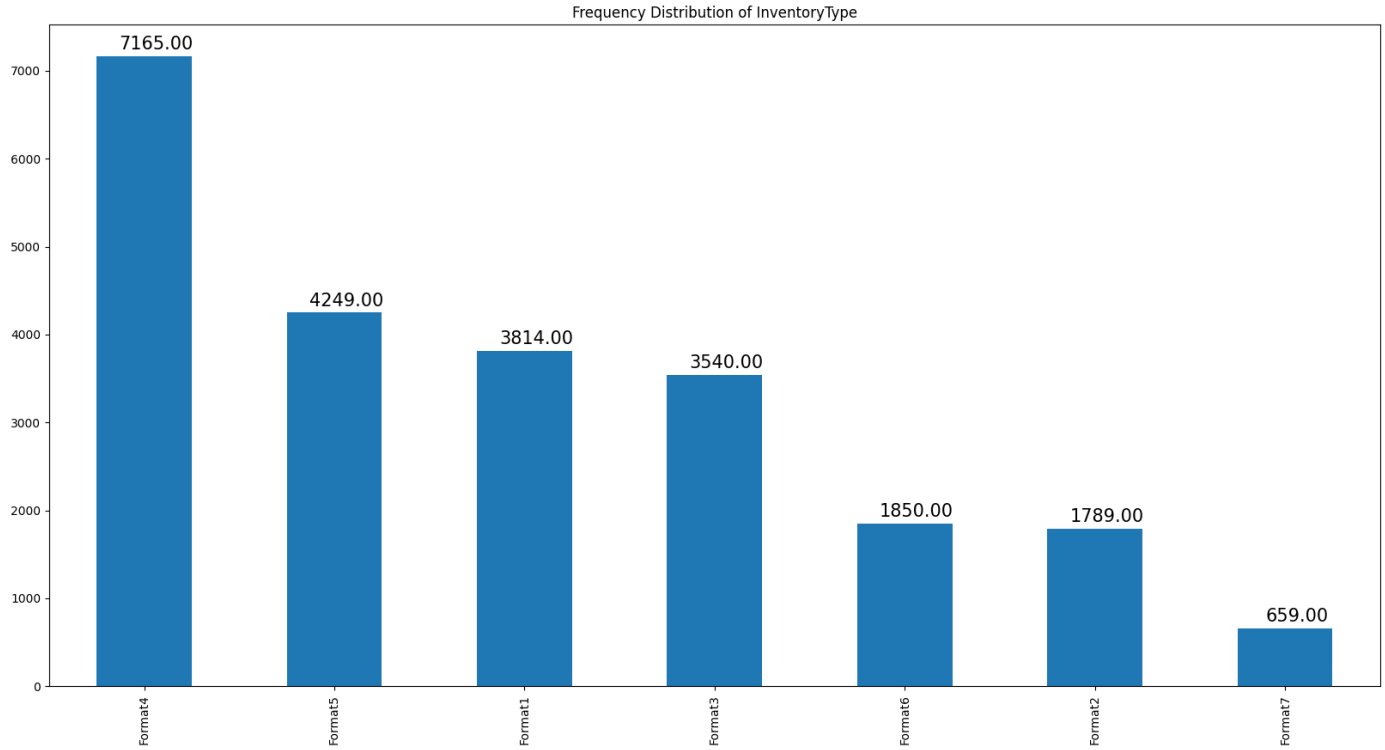


Figure 27: Plot of Inventory type

Details of 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: Ad Type, dtype: int64

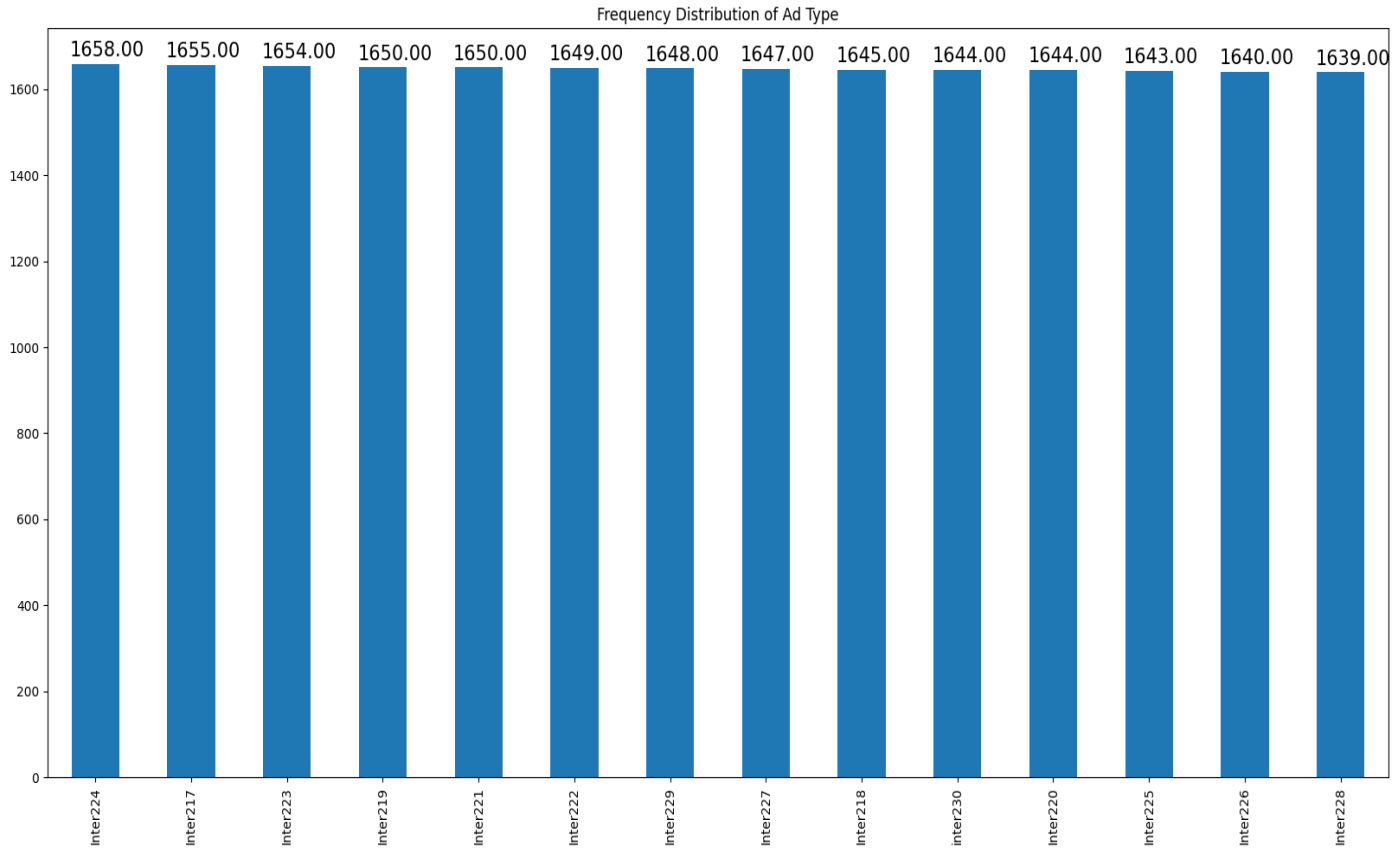


Figure 28: Plot of Ad type

Details of Platform

```
-----------------------------------------------------------------
Video     9873
Web       8251
App       4942
```

Name: Platform, dtype: int64

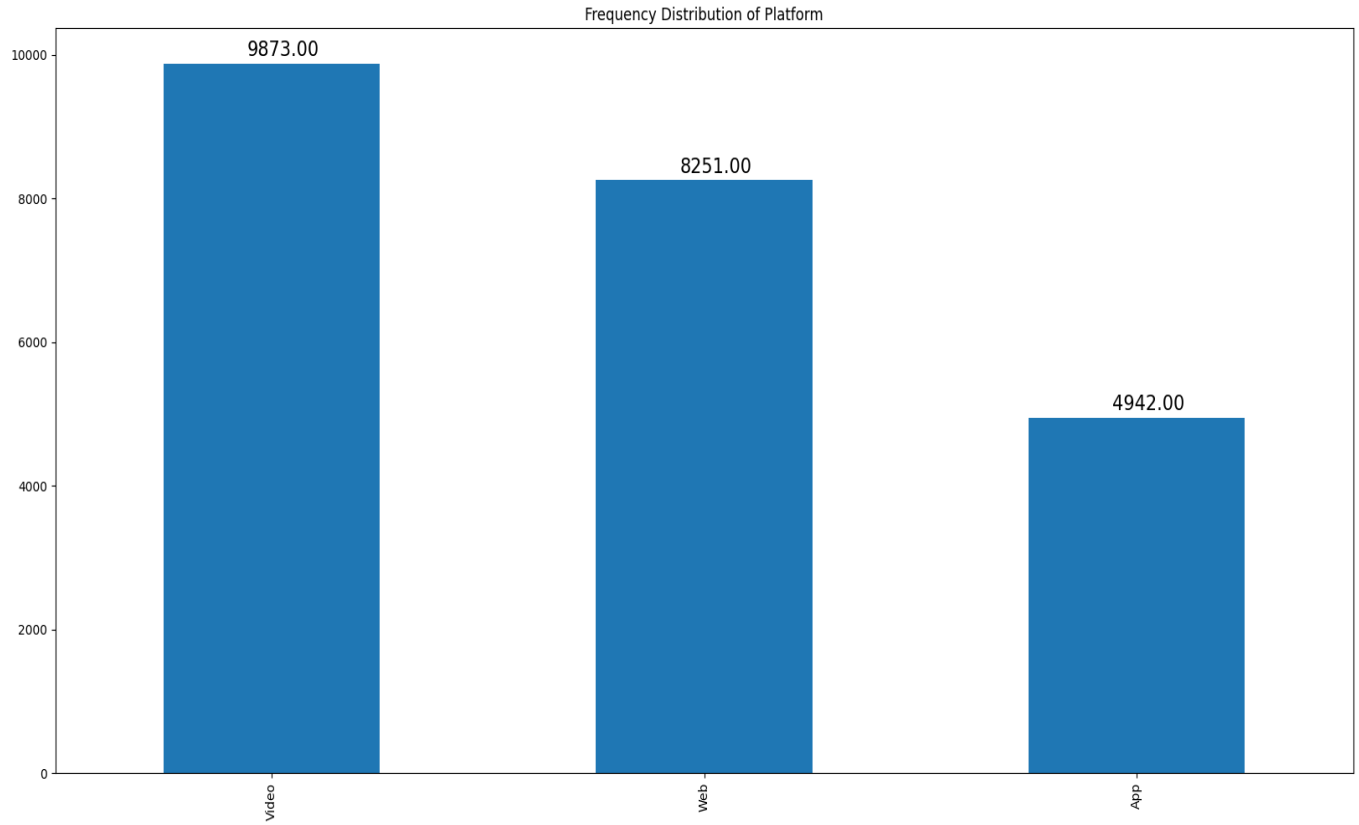


Figure 29: Plot of platforms

Details of Device Type

```
----------------------------------------------------------------------
Mobile     14806
Desktop     8260
Name: Device Type, dtype: int64
```

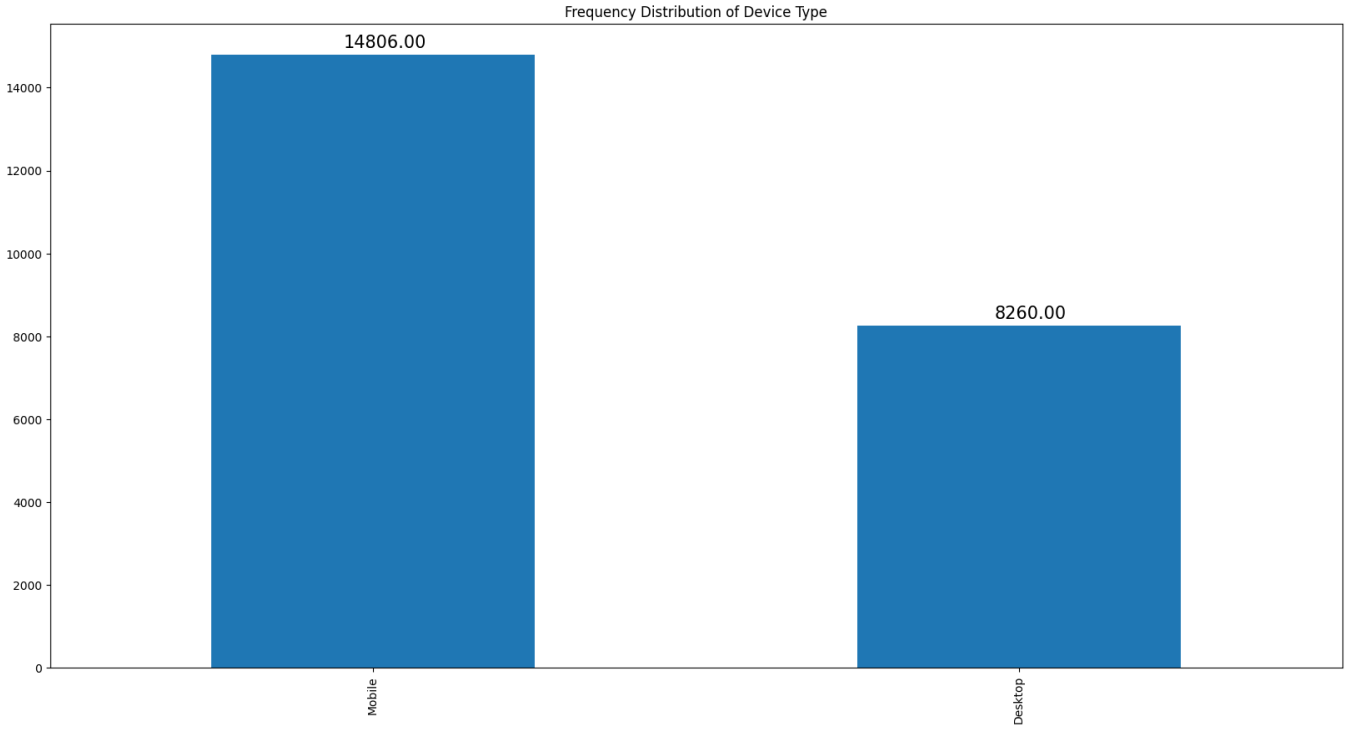


Figure 30: Plot of Device type

Details of Format

```
----------------------------------------------------------------
Video      11552
Display    11514
Name: Format, dtype: int64
<Figure size 640x480 with 0 Axes>
```

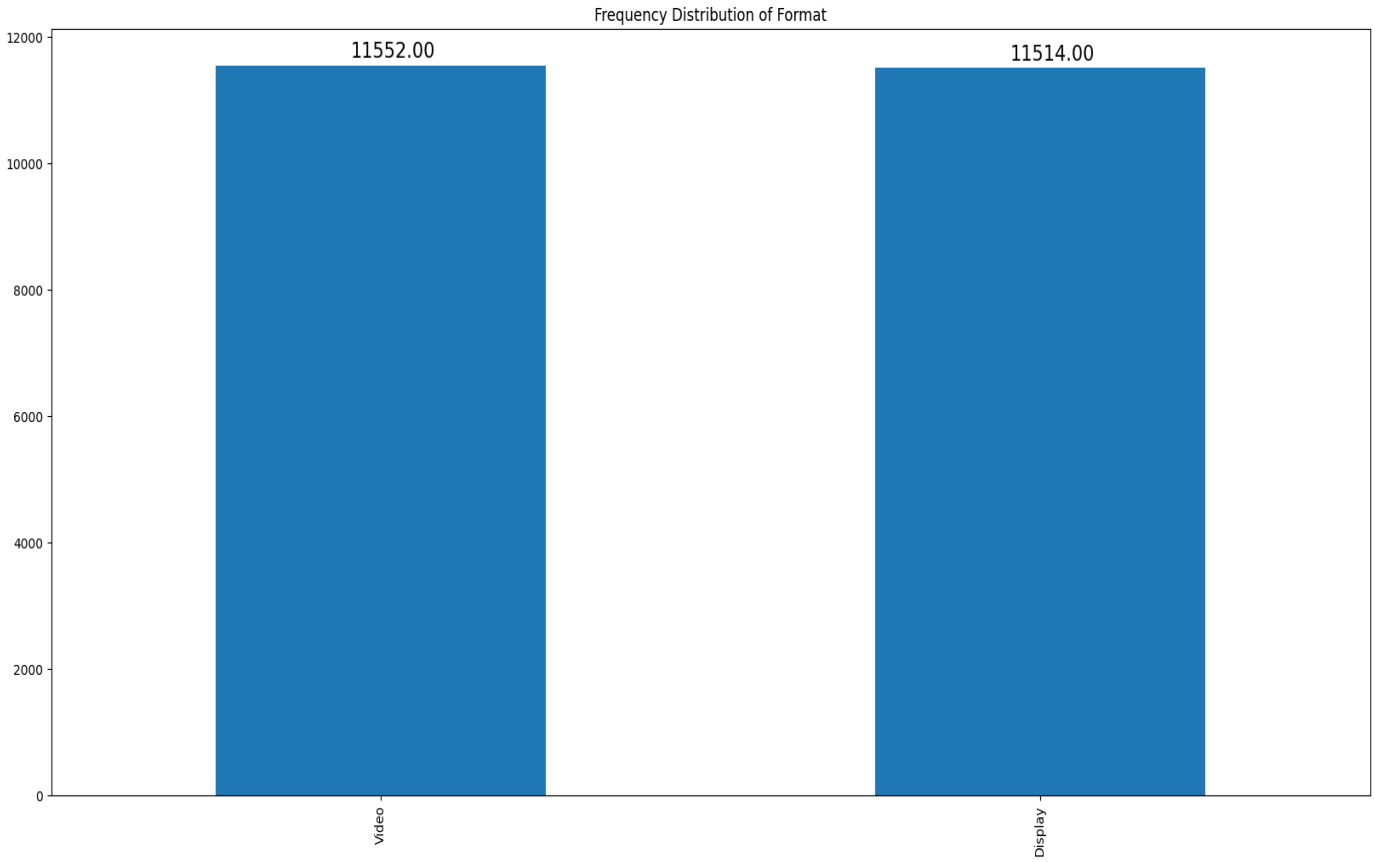


Figure 31: Plot of Format

Insight

- Format4 inventory type of the advertisement is higher and the lowest is Format7.
- There are no much differences in Ad-type. All are close to each other.
- There are video platforms more than web and Apps.
- The Ad are shown in mobile device more than desktops.
- The format of Ad is almost equal in both Display and Video.
- Minimum Ad – Length is 120 and maximum is 728.
- Minimum Ad – width is 70 and maximum is 600.
- The average size of Ad-Size is 96674.47.
- There are some Ad shown only once and some are shown more than 200 lacs time on a search result page or other site on a Network.
- The minimum exact search type of a particular Ad is one and the maximum is approx. 140 lacs.
- There are outlier presents in all features except Ad-length and Ad-width.
- Feature Fee has left skewness and Ad – Size has both side skewness.

Bivariate Analysis

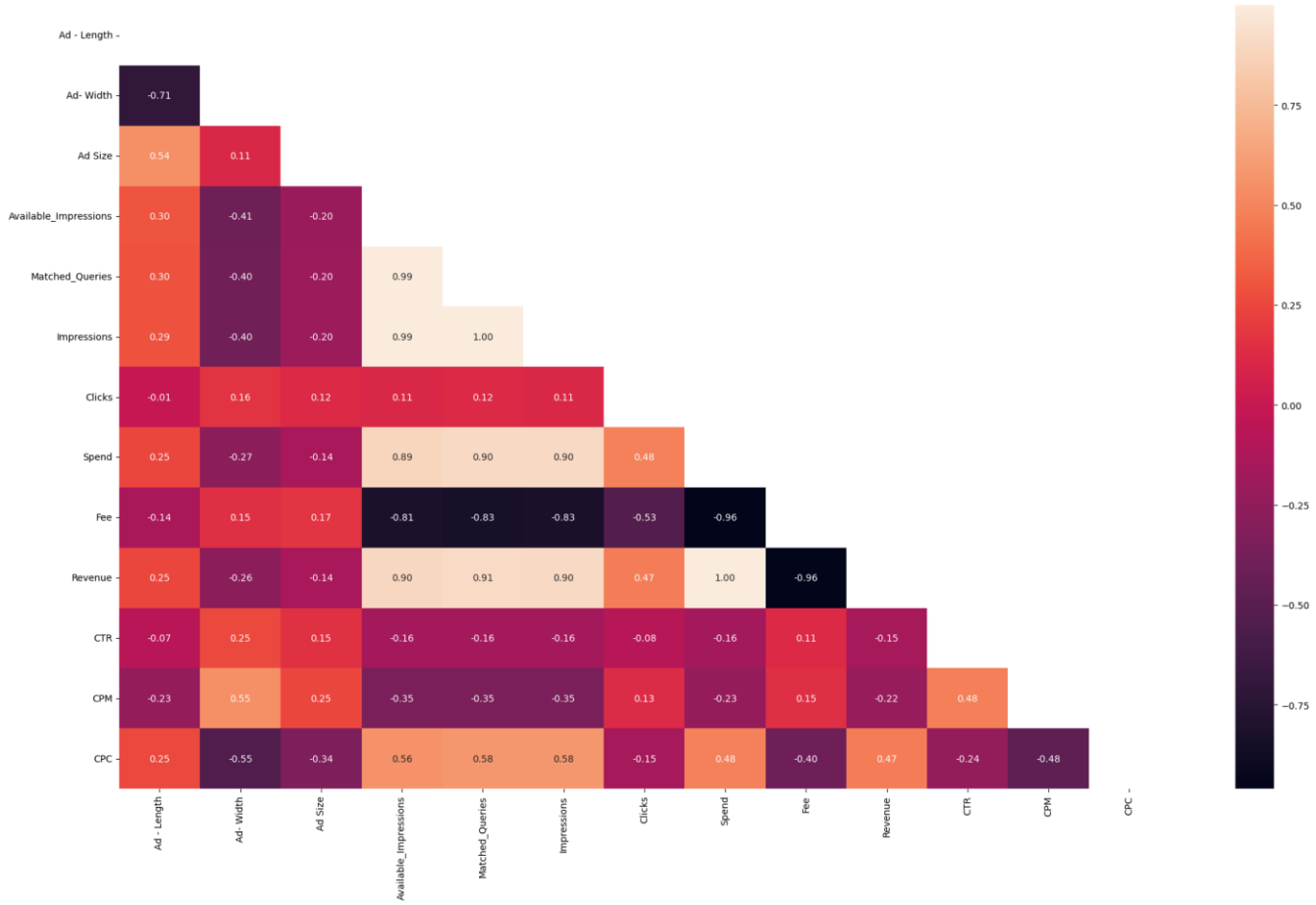


Figure 32: Heatmap for Correlation

Observation

- 'Impressions' shows high correlation with 'Available_Impressions' and 'Matched_Queries'
- 'Matched queries' has high correlation with 'Available_Impressions'
- 'Spend' has correlation with 'Available_Impressions','Matched_Queries' and 'Impressions'
- 'Revenue' has correlation with 'Spend', 'Available_Impressions','Matched_Queries' and has high correlation with'Impressions'.
- There are also negative correlations, Revenue has high negative correlation with Fee.
- Fee has high negative correlation with Spend, Available Impressions, Match_Queries and Impressions.

Outlier Treatment

Let's check the outliers in numerical variables.

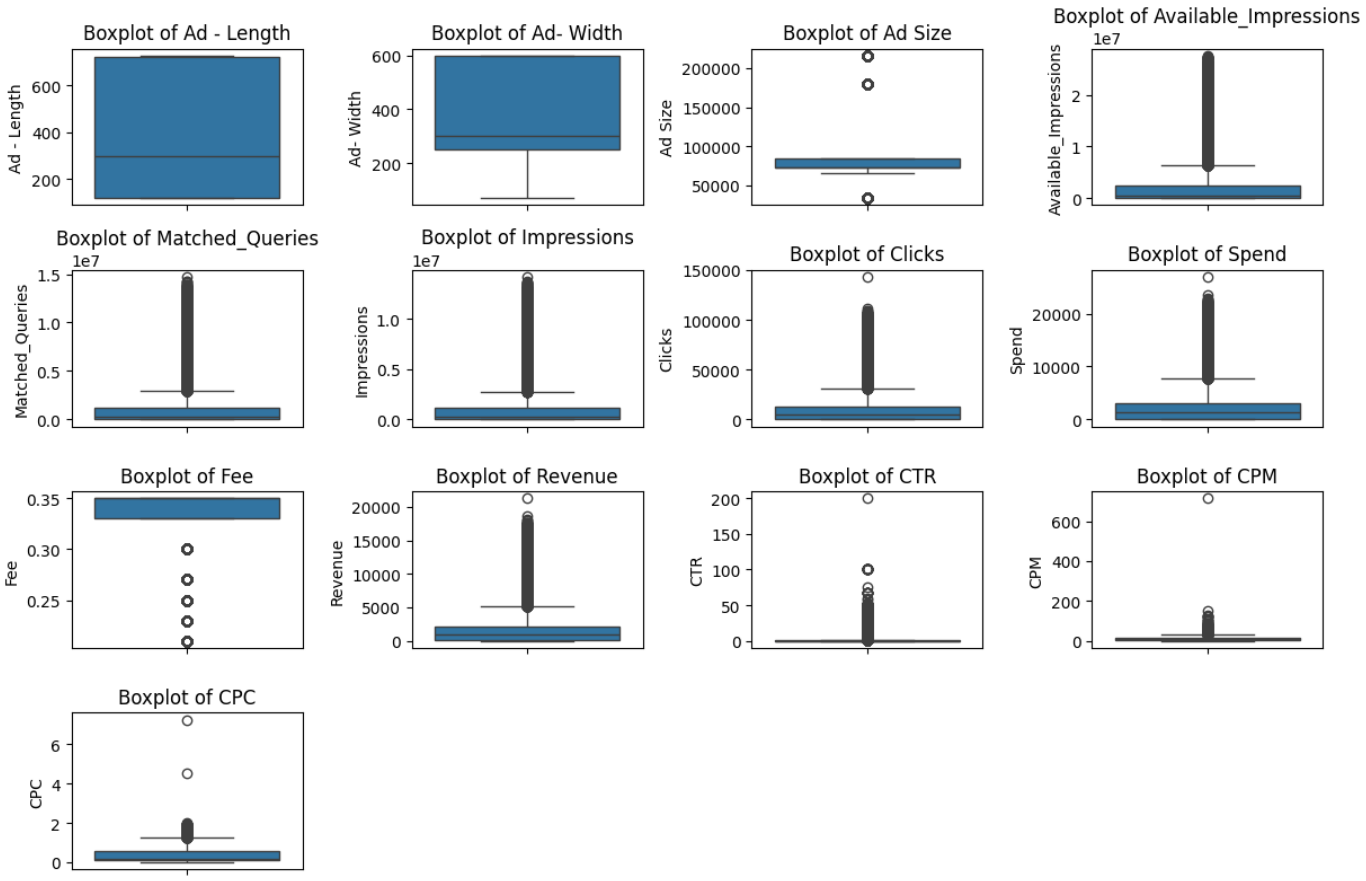


Figure 33: Boxplot of numerical variables

If we set Q1 at 0.25 and Q3 at 0.75 then we have outliers in 11 variables out of 13. Let's look at the table below which will indicate how much data needs to be treated for the outliers.

Ad - Length	0
Ad- Width	0
Ad Size	8448
Available_Impressions	2378
Matched_Queries	3192
Impressions	3269
Clicks	1691
Spend	2081
Fee	3517
Revenue	2325
CTR	3487
CPM	208
CPC	568

Table 9: checking outliers in variables.

In the above table, it looks like that Ad-Size has higher outliers and if we treat the outliers at the range of Q1 at 0.25 and Q3 at 0.75 then we manipulate approx. 36% original data. Similarly, CTR, Fee,

Impressions, Matched queries columns also have approx. 10-15% data. Therefore, we will check with different ranges of Q1 and Q2.

Let's set if number is **greater than Q3+1.5*IQR and less than 0.95,** See below table for the result.

```
Ad - Length                0
Ad- Width                  0
Ad Size                  659
Available_Impressions   1224
Matched_Queries         2038
Impressions             2115
Clicks                   537
Spend                    927
Fee                        0
Revenue                 1171
CTR                     2332
CPM                        0
CPC                        0
dtype: int64
```

Table 10: checking outliers in variables.

Compared to the previous table, in the above table we have 8 variables where we have the outliers. Their ratio is also less; therefore, we will treat the outlier with **greater than Q3+1.5*IQR and less than 0.95.**

Let's drop the column which does not have the outliers as per the above table.

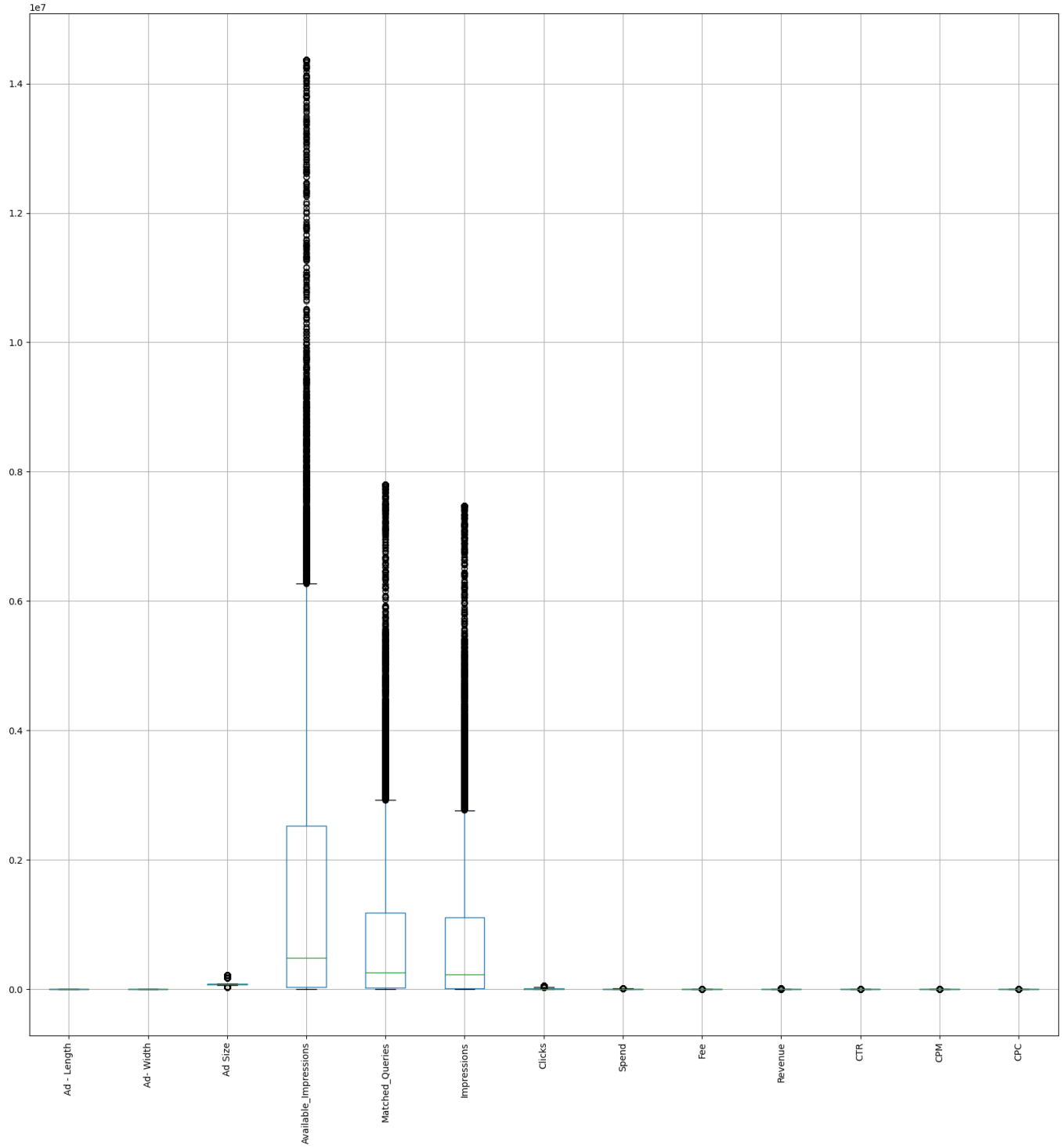


Figure 34: Boxplot of numerical variables after outliers treatment

Z-score scaling

Scaling the data with Z, below is the 5 rows table.

	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
0	-0.364496	-0.432797	-0.352218	-0.592761	-0.586089	-0.580978	-0.737121	-0.754487	0.465447	-0.712603	-0.400200	-0.92711	-0.986603
1	-0.364496	-0.432797	-0.352218	-0.592768	-0.586109	-0.580998	-0.737121	-0.754487	0.465447	-0.712603	-0.400122	-0.92711	-0.986603
2	-0.364496	-0.432797	-0.352218	-0.592505	-0.586073	-0.580961	-0.737121	-0.754487	0.465447	-0.712603	-0.400258	-0.92711	-0.986603
3	-0.364496	-0.432797	-0.352218	-0.592587	-0.586001	-0.580887	-0.737121	-0.754487	0.465447	-0.712603	-0.400414	-0.92711	-0.986603
4	-0.364496	-0.432797	-0.352218	-0.592925	-0.586131	-0.581021	-0.737121	-0.754487	0.465447	-0.712603	-0.400005	-0.92711	-0.986603

Table 11: Top 5 rows after scale the data

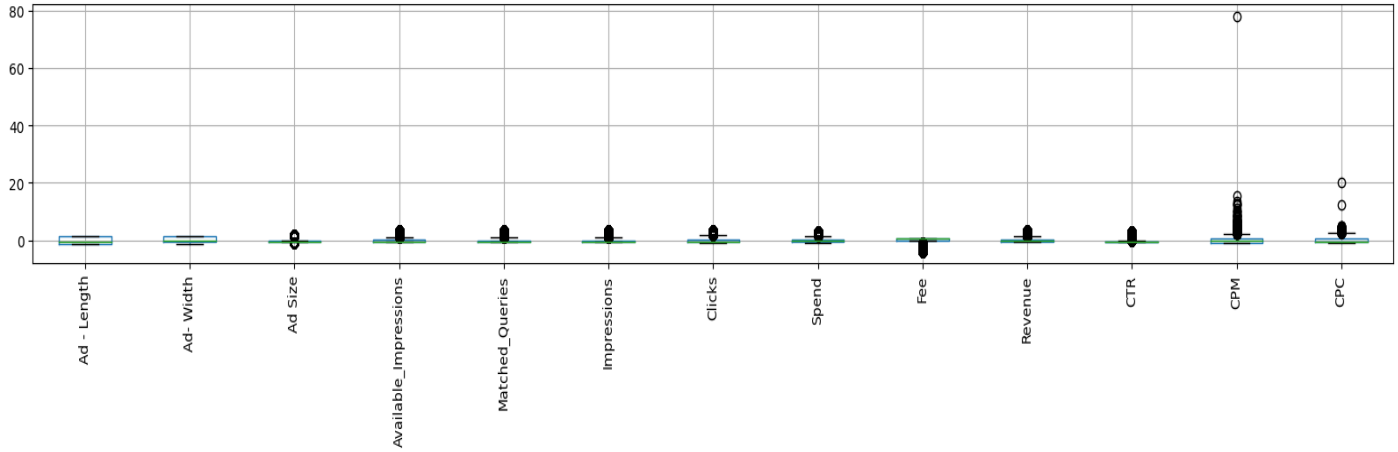


Figure 35: Boxplot of scaled data

Problem 1 - Hierarchical Clustering

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

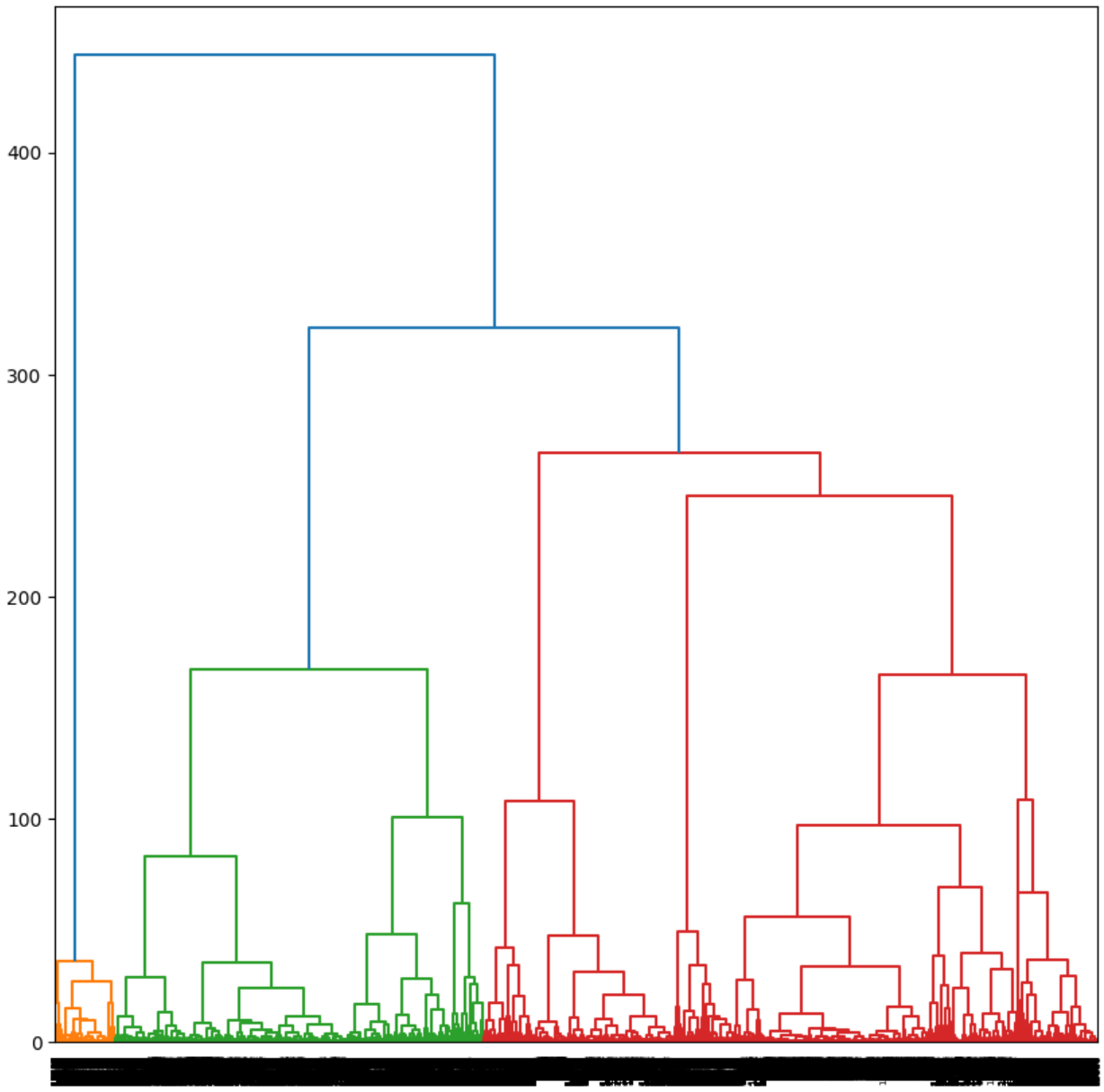


Figure 36: Dendrogram with all clusters

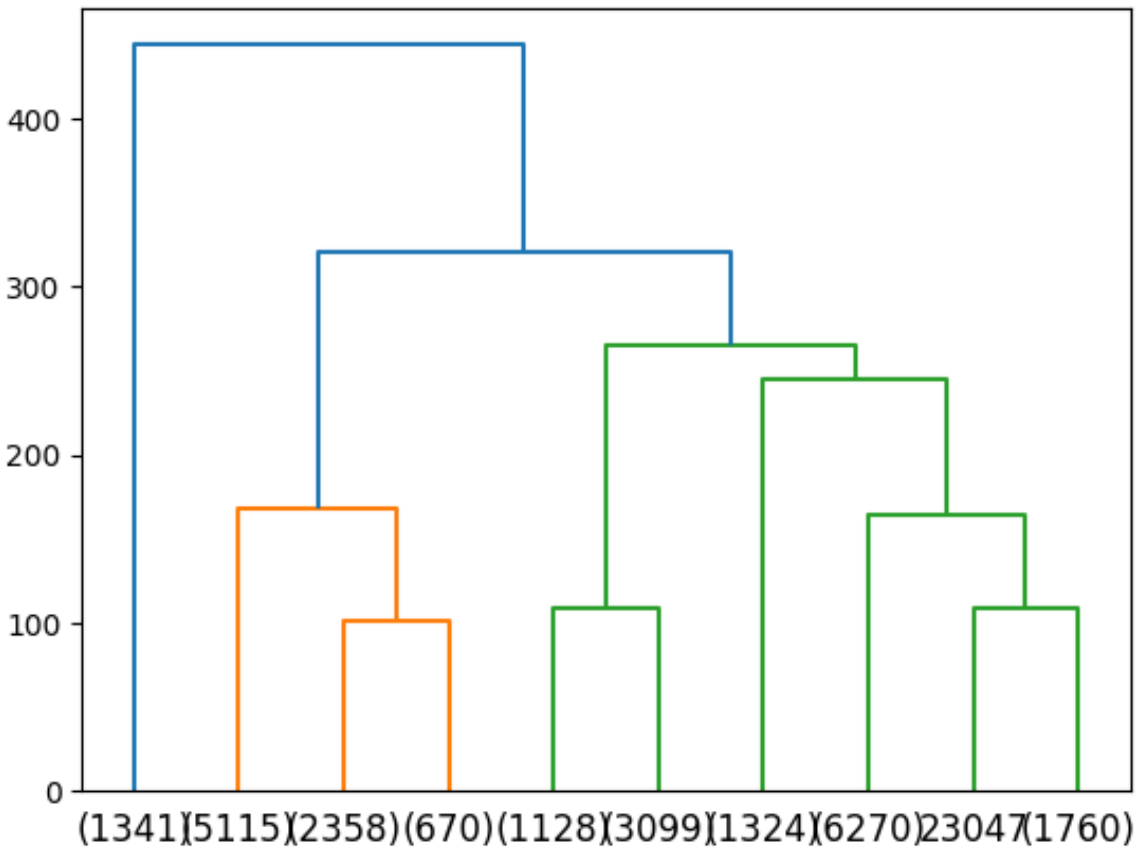


Figure 37: Cutting the dendrogram with 10 clusters.

The optimum number of Clusters are 2.

With the optimum clusters, let us apply agglomerative clustering and predict clusters for the given dataset.

0 21725
1 1341

Clusters	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC	Freq
0	366.803406	351.498274	98309.136939	1.394187e+06	7.440640e+05	7.104620e+05	10175.513878	1865.262293	0.341308	1268.944858	2.774454	8.811612	0.303137	21725
1	682.601044	117.531693	70191.856823	1.924595e+07	1.022220e+07	9.844968e+06	18827.498881	16337.214631	0.234922	12540.624321	0.029392	1.675405	0.879615	1341

Table 12: Table of hierarchical clusters with dataset

Insight

- There are 21275 rows belongs to cluster 0 and 1341 belongs to cluster 1.
- The CTR value of cluster 0 is higher than 1. Hence, the CPC value of cluster 0 is less than cluster 1. There is an inverse relationship between CTR and CPC. This means that as CTR increases, CPC tends to decrease. This is because a higher CTR indicates that the ad is more relevant and engaging to the audience, which means that the advertiser is getting more value from their ad spend.

Problem 1 - K-means Clustering

Forming clusters with K = 1,2,3,4,5,6 and comparing the WSS

- K=1, Inertia 299858.00000000047
- K=2, Inertia 196861.54754355873
- K=3, Inertia 146514.9854398781
- K=4, Inertia 110554.55510725029
- K=5, Inertia 80016.44156075134
- K=6, Inertia 65768.58471487366

```
[299858.00000000047,
 196861.54754355873,
 146514.9854398781,
 110554.55510725029,
 80016.44156075134,
 65768.58471487366,
 53437.02058520891,
 47464.97993942308,
 41512.23895409236,
 36475.77094042392]
```

Table 13: Forming clusters with K = 1,2,3,4,5,6 and comparing the WSS

Elbow curve

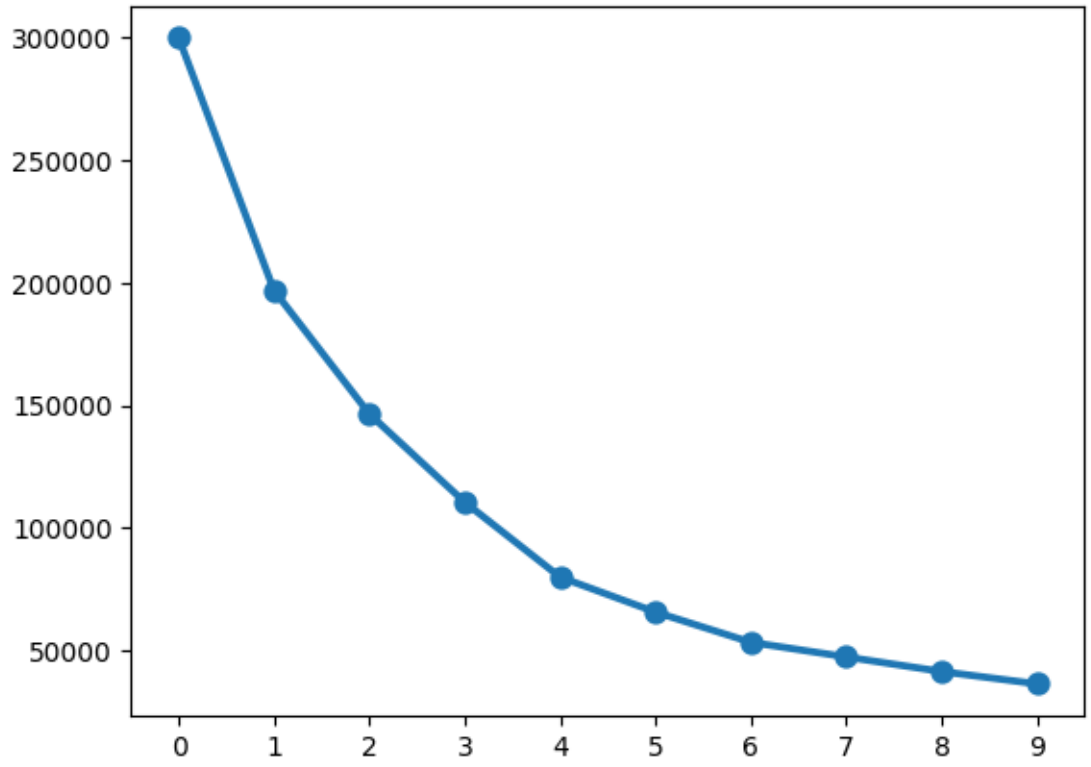


Figure 38: Elbow curve

Observations

- As we can see in the graph that from 1 to 2, 2 to 3 are sharp drop. Either we could say that there are 2 clusters or 3 clusters.

Silhouette Scores

```
For n_clusters=2, the silhouette score is 0.5562692971581535
For n_clusters=3, the silhouette score is 0.33516019446462325
For n_clusters=4, the silhouette score is 0.4406457912438506
For n_clusters=5, the silhouette score is 0.4937828331508482
For n_clusters=6, the silhouette score is 0.4857123627066932
For n_clusters=7, the silhouette score is 0.5194514232284791
For n_clusters=8, the silhouette score is 0.5290239340640361
For n_clusters=9, the silhouette score is 0.5291837875748412
For n_clusters=10, the silhouette score is 0.5219900671829257
```

Figure 39: Silhouette scores

Silhouette score is better for 2 clusters than for 3 and 4 clusters. So, final clusters will be 2.

Appending Clusters to the original dataset for Cluster Profiling.

Cluster	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC	freq
0	364.287867	354.430303	98735.316083	1.277661e+06	6.783578e+05	6.466786e+05	10197.351785	1778.509725	0.342239	1205.457931	2.819732	8.931070	0.292862	21371
1	648.363422	129.427729	70690.761062	1.698680e+07	9.071134e+06	8.741428e+06	16745.197640	14408.552684	0.245404	10986.998714	0.031815	1.659643	0.888768	1695

Table 14: Clusters with dataset

Insights

- There are 21371 belonging to cluster 0 and 1695 belongs to cluster 1.
- The CTR is higher of cluster 0 than 1.

Let's check more insight with the help of visualization.

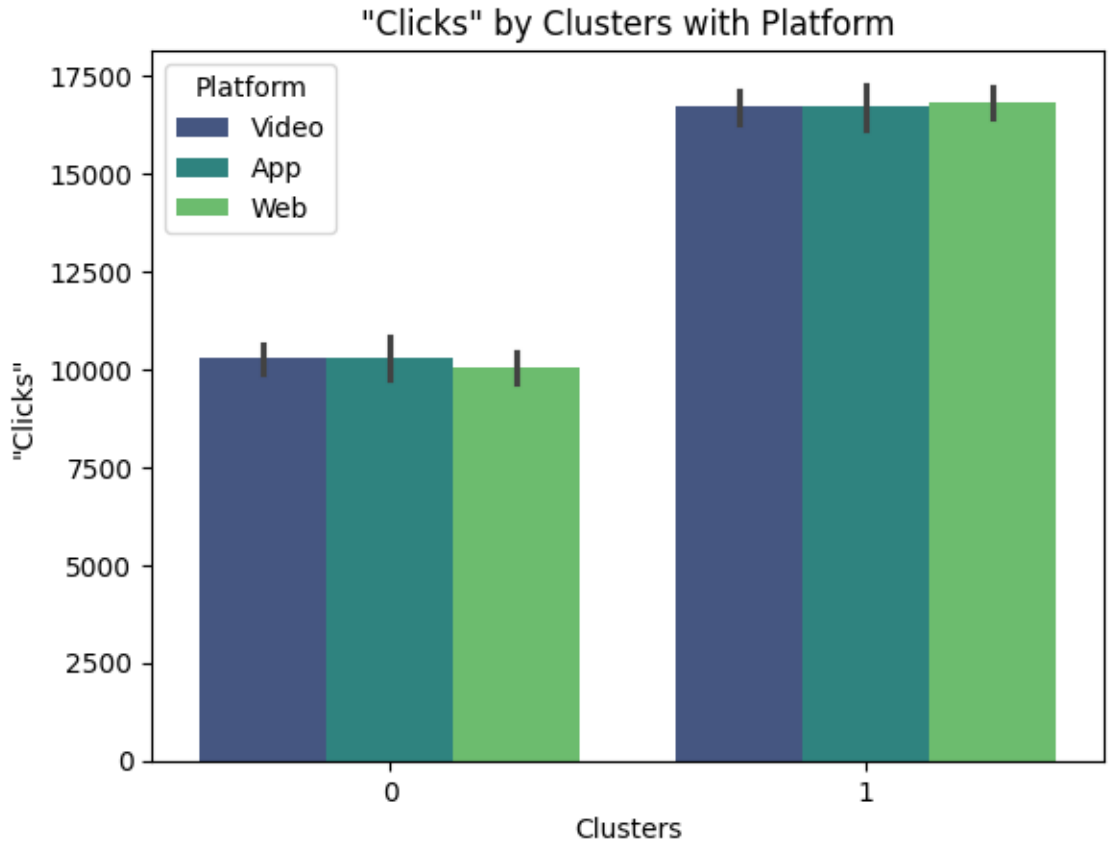


Figure 40: Click by clusters with platform

- For Cluster 0, the average count of clicks is slightly higher for App than video and web.

- For cluster 1, the average count of clicks is higher for web than video and app.

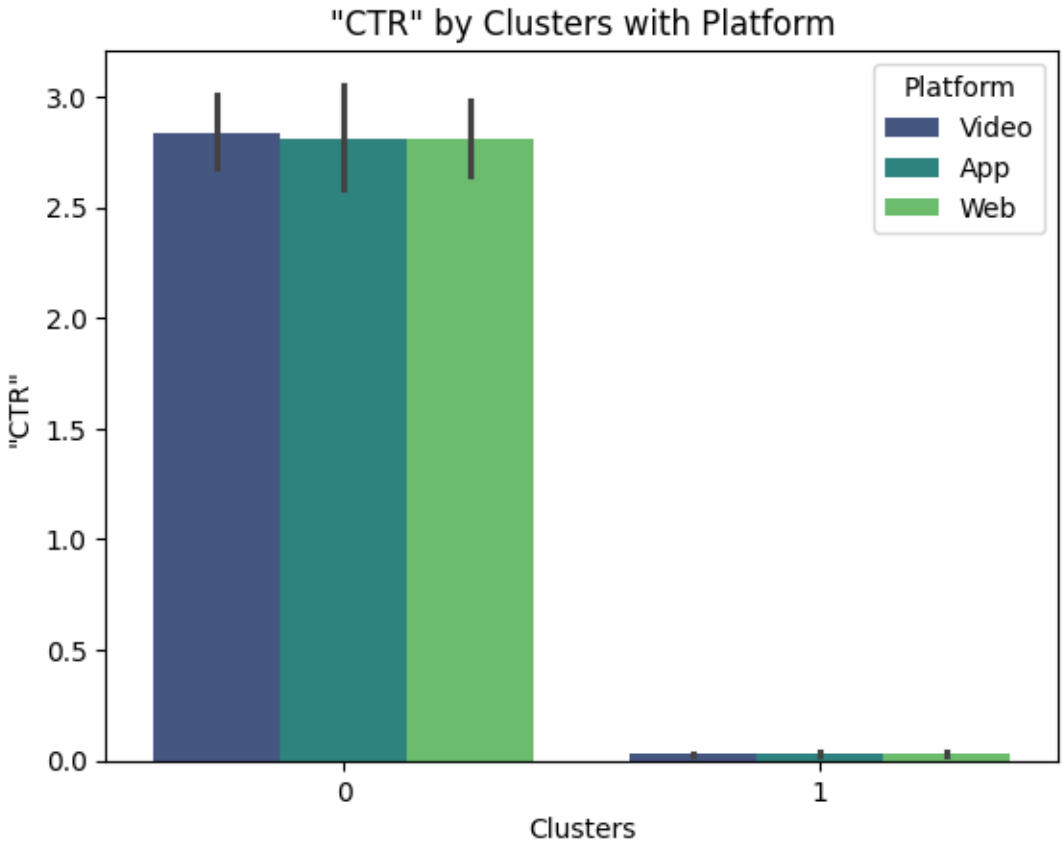


Figure 41: CTR by clusters with platform

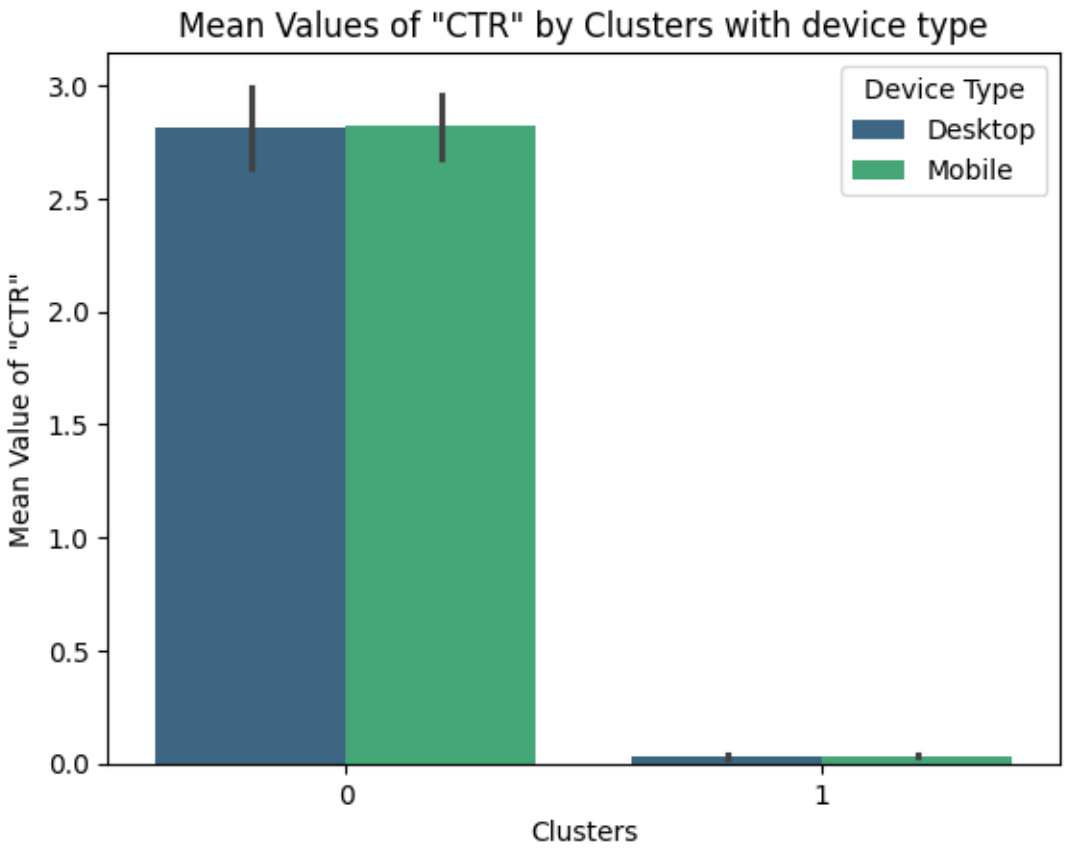


Figure 42: CTR by clusters with device

- The people from cluster 0 who see an advertisement and click on it more often are from the video, apps, and web. The videos are watched most on Mobile.
- For cluster 1, the overall CTR is low. However, if we compare between the platforms so web is slightly higher than app and web.

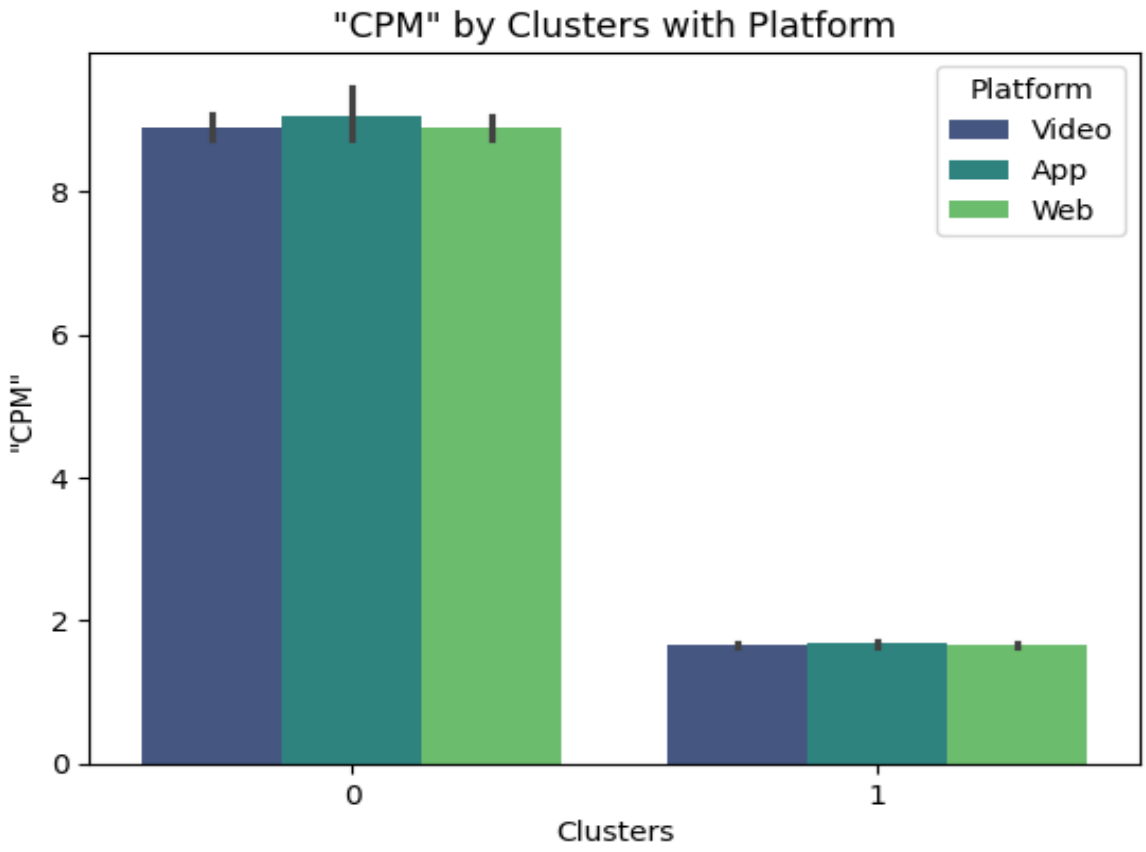


Figure 43: CPM by clusters with platform

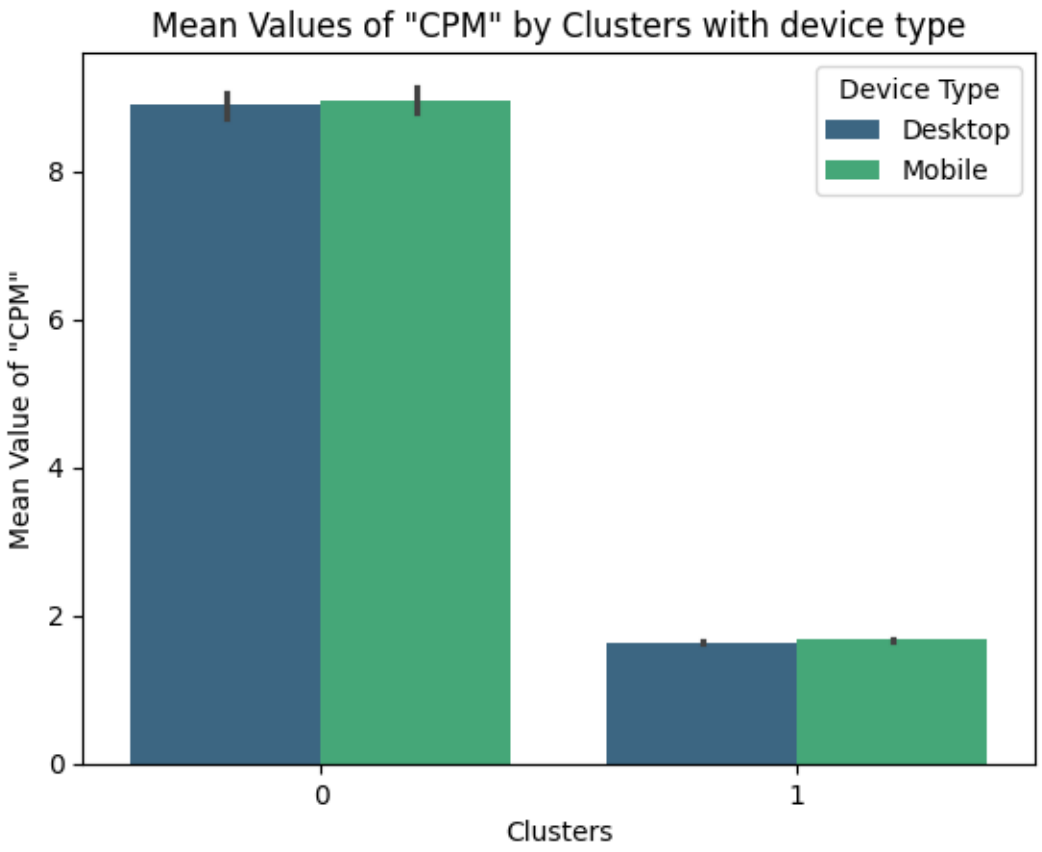


Figure 44: CPM by clusters with device type

- The CPM value of cluster 0 is higher than cluster 1 which means that cluster 0's Ad are being shown to many people than cluster 1 people. Their Ad's are shown more on mobiles. There is not much difference between mobiles and desktop.
- The average number of Ad shown to people is higher for App of cluster zero than video and web.

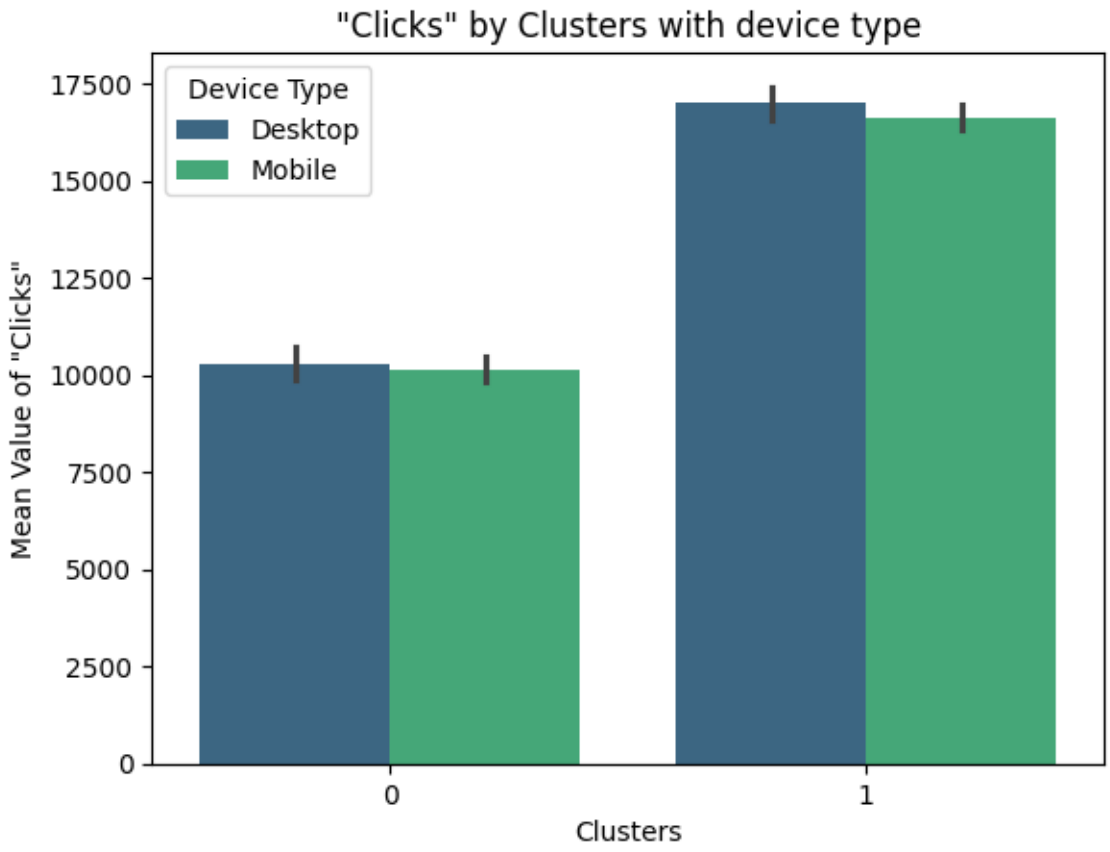


Figure 45: "Click' by clusters with Device type

- The number of times users have clicked more on device desktop for both Cluster 0 and 1.

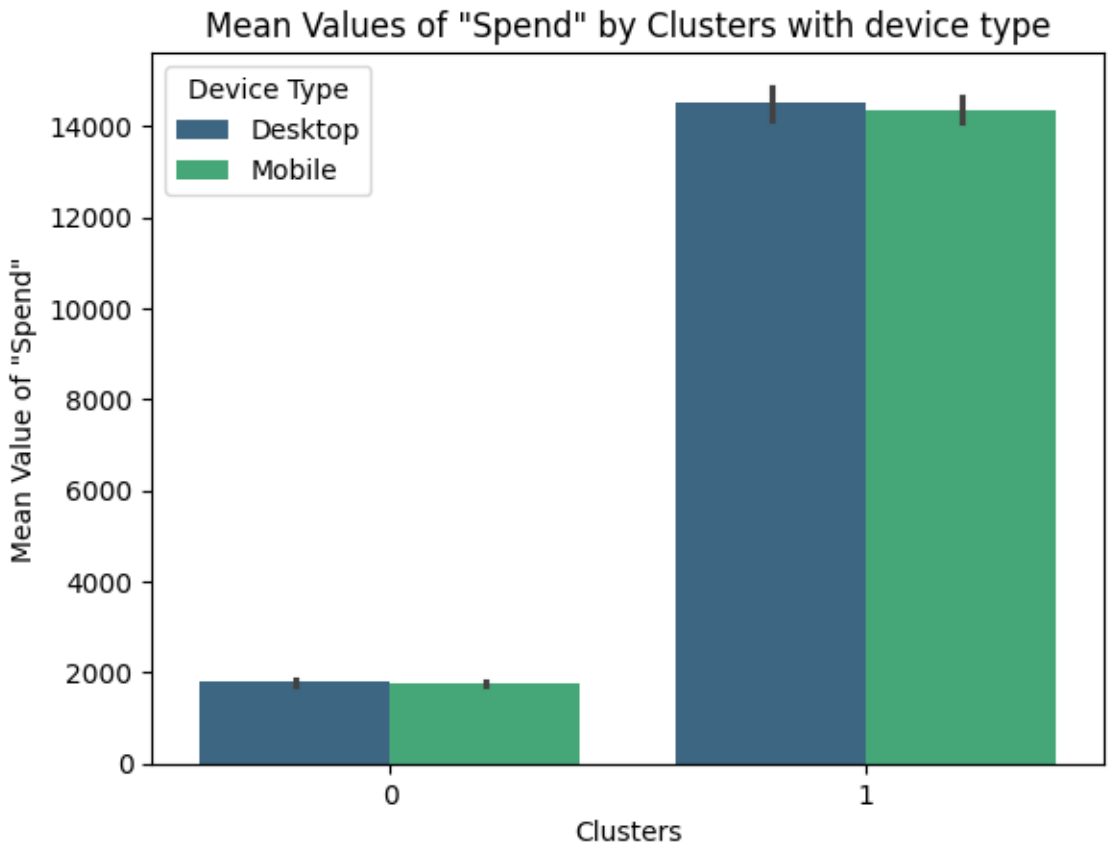


Figure 46: Spend by clusters with platform

- The average amount of money is spent less by cluster 0 than cluster 1.

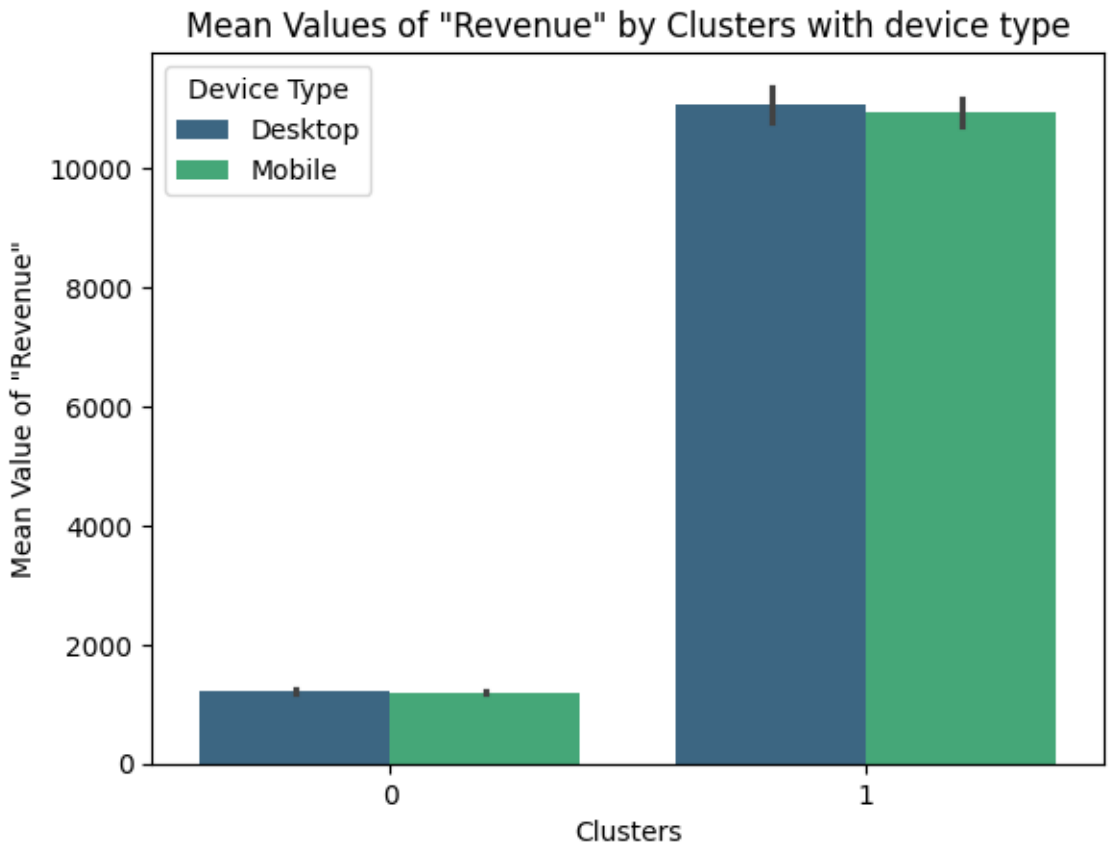


Figure 47: Revenue by clusters with platform

- The revenue is higher of cluster 1 than cluster zero. The greater amount of income has come from Desktop for cluster 1.

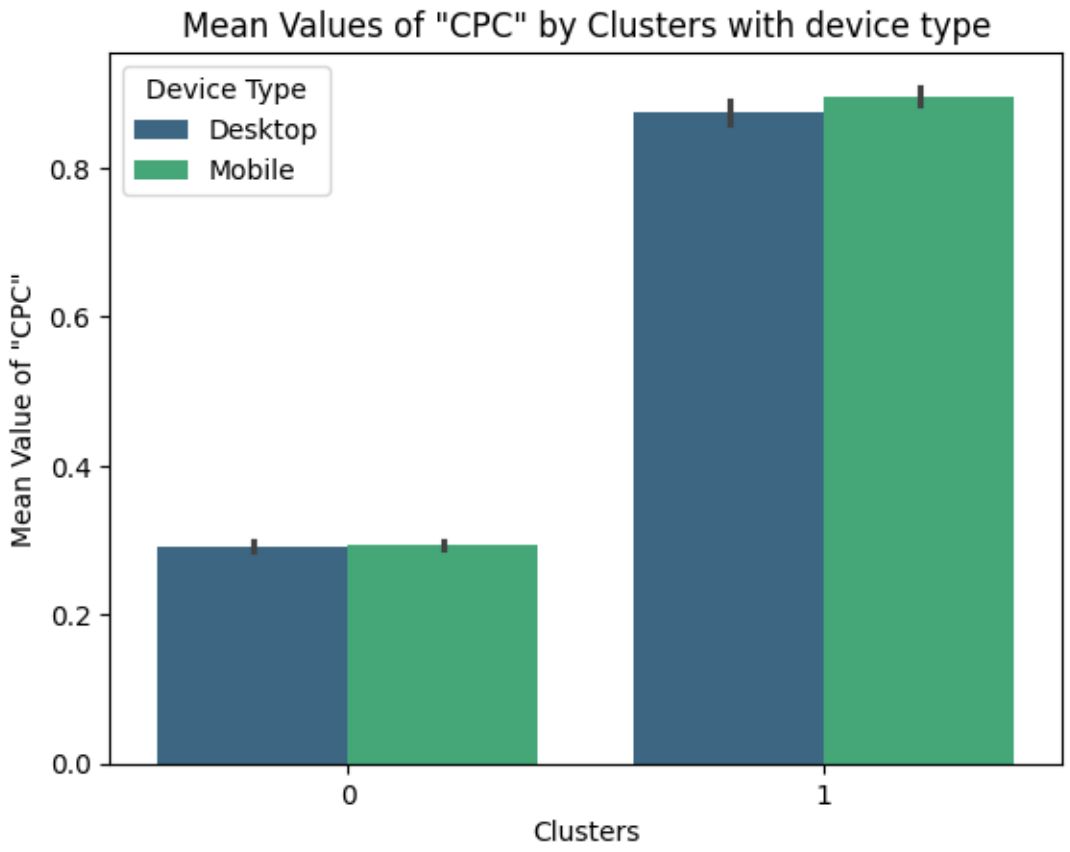


Figure 48: CPC by clusters with platform

- The amount on Ads paid more of cluster 1 than cluster 0.

Recommendations

- Cluster 0 has higher CTR and CPM compared to Cluster 1.
- Cluster 0 has lower spending compared to Cluster 1.
- Cluster 1 has a higher CPC compared to Cluster 0.
- But, Cluster 1 has a higher revenue than Cluster 0.

Cluster 0

- Cluster 0 excels in creating Ads that attract clicks and have a broader reach. Cluster 0 has a lower spend but also generate lower revenue compared to cluster 1.
- Despite the lower revenue, Cluster 0's lower spending might indicate potential efficiency in cost management.
- They should focus on those strategies which can help to increase revenue.

Cluster 1

- Despite a higher CPC, Cluster 1 has managed to generate significantly higher revenue.
- Shows effectiveness in converting clicks into revenue despite higher spending.
- The higher revenue could suggest better conversion optimization or targeting strategies compared to Cluster 0.

PCA:

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

Problem 2 - Define the problem and perform Exploratory Data Analysis

Loading and checking the dataset

Checking top 5 rows using the head function

State Code	Dist.Code	State	Area Name	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	...	MARG_CL_0_3_M	MARG_CL_0_3_F	MARG_AL_0_3_M	MARG_AL_0_3_F	MARG_HH_0_3_M	MARG_HH_0_3_F	MARG_OT_0_3_
1	1	Jammu & Kashmir	Kupwara	7707	23388	29796	5862	6196	3	...	1150	749	180	237	680	252	
1	2	Jammu & Kashmir	Badgam	6218	19585	23102	4482	3733	7	...	525	715	123	229	186	148	
1	3	Jammu & Kashmir	Leh(Ladakh)	4452	6546	10964	1082	1018	3	...	114	188	44	89	3	34	
1	4	Jammu & Kashmir	Kargil	1320	2784	4206	563	677	0	...	194	247	61	128	13	50	
1	5	Jammu & Kashmir	Punch	11654	20591	29981	5157	4587	20	...	874	1928	465	1043	205	302	

Table 15: Top 5 rows data

Checking last 5 rows using tail function

	State Code	Dist.Code	State	Area Name	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	...	MARG_CL_0_3_M	MARG_CL_0_3_F	MARG_AL_0_3_M	MARG_AL_0_3_F	MARG_HH_0_3_M	MARG_HH_0_3_F	MARG_
635	34	636	Puducherry	Mahe	3333	8154	11781	1146	1203	21	...	32	47	0	0	0	0	
636	34	637	Puducherry	Karaikal	10612	12346	21691	1544	1533	2234	...	155	337	3	14	38	130	
637	35	638	Andaman & Nicobar Island	Nicobars	1275	1549	2630	227	225	0	...	104	134	9	4	2	6	
638	35	639	Andaman & Nicobar Island	North & Middle Andaman	3762	5200	8012	723	664	0	...	136	172	24	44	11	21	
639	35	640	Andaman & Nicobar Island	South Andaman	7975	11977	18049	1470	1358	0	...	173	122	6	2	17	17	

Table 16: last 5 rows data

Check shape and information data types

```
(640, 61)
```

```
<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
```

Table 17: Data types and information

Insight

- There are 61 columns and 640 observations.
- There are all columns are integer type except two columns *State* and *Area name* which are object.
- There are no null values.

Statistical summary

Index	count	mean	std	min	25%	50%	75%	max
State Code	640	17.11	9.43	1	9	18	24	35
Dist.Code	640	320.5	184.9	1	160.75	320.5	480.25	640
No_HH	640	51222.87	48135.41	350	19484	35837	68892	310450

TOT_M	640	79940.58	73384.51	391	30228	58339	107918.5	485417
TOT_F	640	122372.1	113600.7	698	46517.75	87724.5	164251.8	750392
M_06	640	12309.1	11500.91	56	4733.75	9159	16520.25	96223
F_06	640	11942.3	11326.29	56	4672.25	8663	15902.25	95129
M_SC	640	13820.95	14426.37	0	3466.25	9591.5	19429.75	103307
F_SC	640	20778.39	21727.89	0	5603.25	13709	29180	156429
M_ST	640	6191.81	9912.67	0	293.75	2333.5	7658	96785
F_ST	640	10155.64	15875.7	0	429.5	3834.5	12480.25	130119
M_LIT	640	57967.98	55910.28	286	21298	42693.5	77989.5	403261
F_LIT	640	66359.57	75037.86	371	20932	43796.5	84799.75	571140
M_ILL	640	21972.6	19825.61	105	8590	15767.5	29512.5	105961
F_ILL	640	56012.52	47116.69	327	22367	42386	78471	254160
TOT_WORK_M	640	37992.41	36419.54	100	13753.5	27936.5	50226.75	269422
TOT_WORK_F	640	41295.76	37192.36	357	16097.75	30588.5	53234.25	257848
MAINWORK_M	640	30204.45	31480.92	65	9787	21250.5	40119	247911
MAINWORK_F	640	28198.85	29998.26	240	9502.25	18484	35063.25	226166
MAIN_CL_M	640	5424.34	4739.16	0	2023.5	4160.5	7695	29113
MAIN_CL_F	640	5486.04	5326.36	0	1920.25	3908.5	7286.25	36193
MAIN_AL_M	640	5849.11	6399.51	0	1070.25	3936.5	8067.25	40843
MAIN_AL_F	640	8926	12864.29	0	1408.75	3933.5	10617.5	87945
MAIN_HH_M	640	883.89	1278.64	0	187.5	498.5	1099.25	16429
MAIN_HH_F	640	1380.77	3179.41	0	248.75	540.5	1435.75	45979
MAIN_OT_M	640	18047.1	26068.48	36	3997.5	9598	21249.5	240855
MAIN_OT_F	640	12406.04	18972.2	153	3142.5	6380.5	14368.25	209355
MARGWORK_M	640	7787.96	7410.79	35	2937.5	5627	9800.25	47553
MARGWORK_F	640	13096.91	10996.47	117	5424.5	10175	18879.25	66915
MARG_CL_M	640	1040.74	1311.55	0	311.75	606.5	1281	13201
MARG_CL_F	640	2307.68	3564.63	0	630.25	1226	2659.25	44324
MARG_AL_M	640	3304.33	3781.56	0	873.5	2062	4300.75	23719
MARG_AL_F	640	6463.28	6773.88	0	1402.5	4020.5	9089.25	45301
MARG_HH_M	640	316.74	462.66	0	71.75	166	356.5	4298
MARG_HH_F	640	786.63	1198.72	0	171.75	429	962.5	15448

MARG_OT_M	640	3126.15	3609.39	7	935.5	2036	3985.25	24728
MARG_OT_F	640	3539.32	4115.19	19	1071.75	2349.5	4400.5	36377
MARGWORK_3_6_M	640	41948.17	39045.32	291	16208.25	30315	57218.75	300937
MARGWORK_3_6_F	640	81076.32	82970.41	341	26619.5	56793	107924	676450
MARG_CL_3_6_M	640	6394.99	6019.81	27	2372	4630	8167	39106
MARG_CL_3_6_F	640	10339.86	8467.47	85	4351.5	8295	15102	50065
MARG_AL_3_6_M	640	789.85	905.64	0	235.5	480.5	986	7426
MARG_AL_3_6_F	640	1749.58	2496.54	0	497.25	985.5	2059	27171
MARG_HH_3_6_M	640	2743.64	3059.59	0	718.75	1714.5	3702.25	19343
MARG_HH_3_6_F	640	5169.85	5335.64	0	1113.75	3294	7502.25	36253
MARG_OT_3_6_M	640	245.36	358.73	0	58	129.5	276	3535
MARG_OT_3_6_F	640	585.88	900.03	0	127.75	320.5	719.25	12094
MARGWORK_0_3_M	640	2616.14	3036.96	7	755	1681.5	3320.25	20648
MARGWORK_0_3_F	640	2834.55	3327.84	14	833.5	1834.5	3610.5	25844
MARG_CL_0_3_M	640	1392.97	1489.71	4	489.5	949	1714	9875
MARG_CL_0_3_F	640	2757.05	2788.78	30	957.25	1928	3599.75	21611
MARG_AL_0_3_M	640	250.89	453.34	0	47	114.5	270.75	5775
MARG_AL_0_3_F	640	558.1	1117.64	0	109	247.5	568.75	17153
MARG_HH_0_3_M	640	560.69	762.58	0	136.5	308	642	6116
MARG_HH_0_3_F	640	1293.43	1585.38	0	298	717	1710.75	13714
MARG_OT_0_3_M	640	71.38	107.9	0	14	35	79	895
MARG_OT_0_3_F	640	200.74	309.74	0	43	113	240	3354
NON_WORK_M	640	510.01	610.6	0	161	326	604.5	6456
NON_WORK_F	640	704.78	910.21	5	220.5	464.5	853.5	10533

Table 18: Statistical summary

Insight

- The average of female population is higher than male population.
- Literates' population is more than Illiterate population.
- The mean value is higher than 50% (median) for all variables except State Code and Dist. Code which indicates if there are outliers then those will be positive skew. Will check more when see the outliers.

There are no duplicate rows found.

Perform an EDA on the data to extract useful insights Note: 1. Pick 5 variables out of the given 24 variables below for EDA.

- Copying columns **(State Code, Dist.Code, State and Area Name)** in new Dataframe then dropping same columns from the dataframe.
- Picking 5 variables out of the given 24 variables below for EDA and storing into a new dataframe.

Data information of selected 5 variables.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 640 entries, 0 to 639
Data columns (total 5 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   No_HH   640 non-null    int64
 1   TOT_M   640 non-null    int64
 2   TOT_F   640 non-null    int64
 3   M_06    640 non-null    int64
 4   F_06    640 non-null    int64
dtypes: int64(5)
memory usage: 25.1 KB
```

Table 19: Data types and information for selected 5 variables

We have selected five variables (Number of households, total male population, total female population, Male population in the age group 0-6, Female population in the age group 0-6. Now will do the Univariate Analysis and Bivariate Analysis

Univariate Analysis

```
Description of No_HH
-------------------------------------------------------------------------
--
count       640.000000
mean      51222.871875
std       48135.405475
min         350.000000
25%       19484.000000
50%       35837.000000
75%       68892.000000
max      310450.000000
```

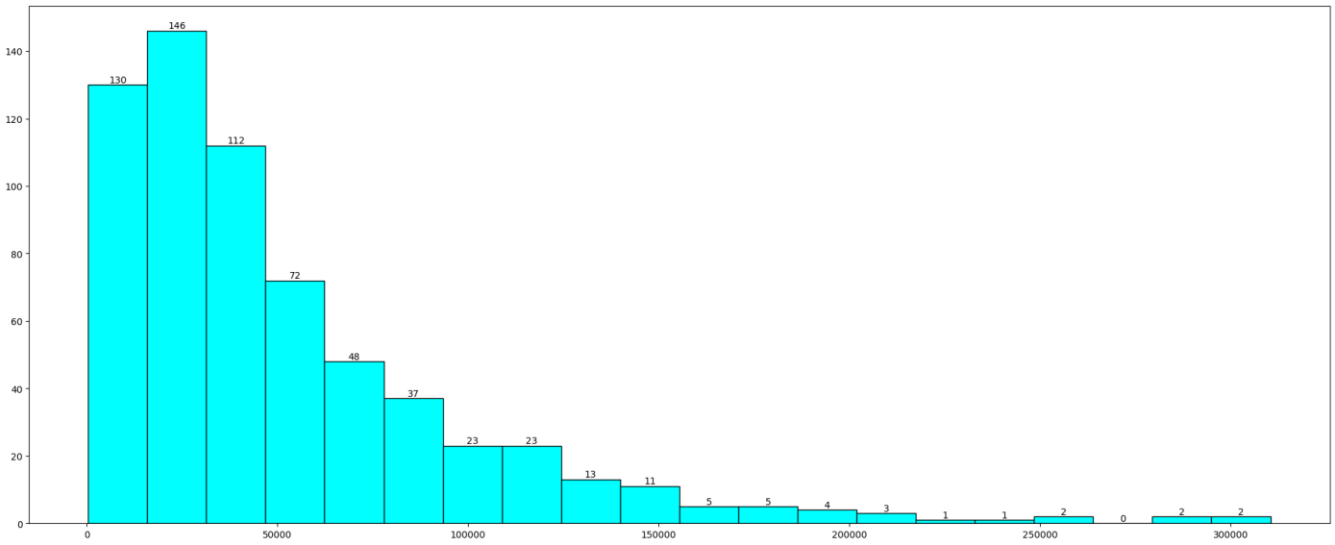


Figure 49: Plot of No_HH (No. of household)

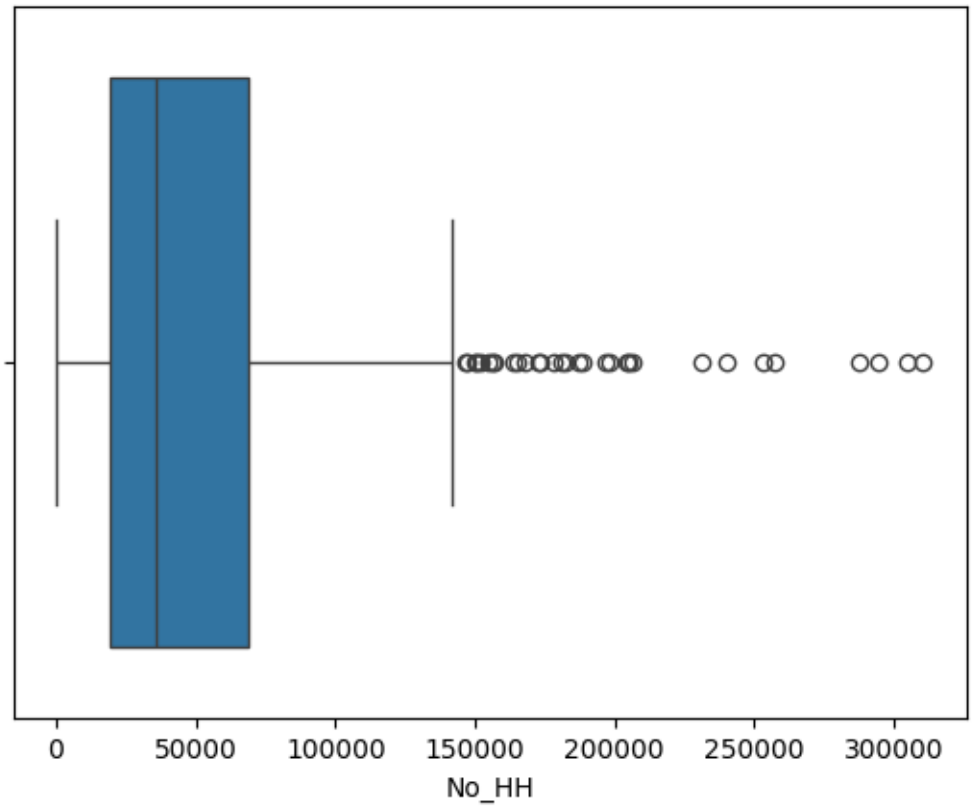


Figure 50: Boxplot of No_HH

```
Description of TOT_M
-------------------------------------------------------------------------------
--
count       640.000000
mean      79940.576563
std       73384.511114
min         391.000000
25%       30228.000000
50%       58339.000000
75%      107918.500000

max      485417.000000
```

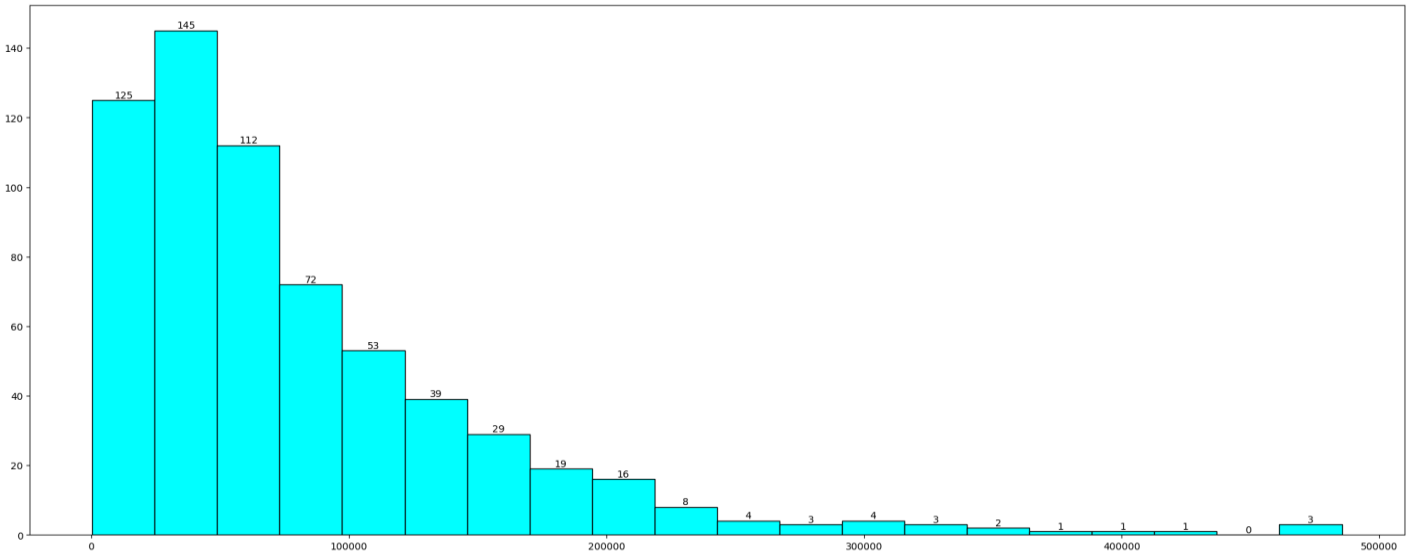


Figure 51: Plot of Tot_M (Total population of male)

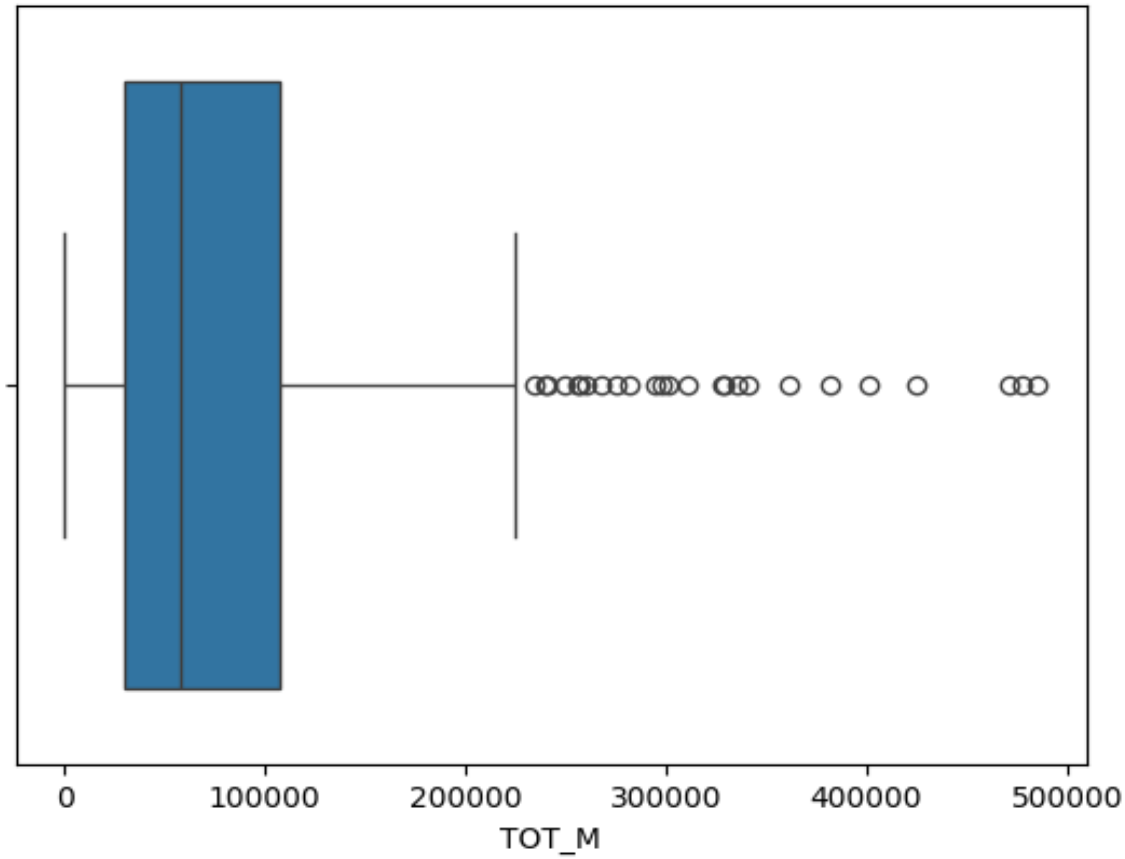


Figure 52: Boxplot of TOT_M (Total population of male)

```
count       640.000000
mean     122372.084375
std      113600.717282
min         698.000000
25%       46517.750000
50%       87724.500000
75%      164251.750000
max      750392.000000
```

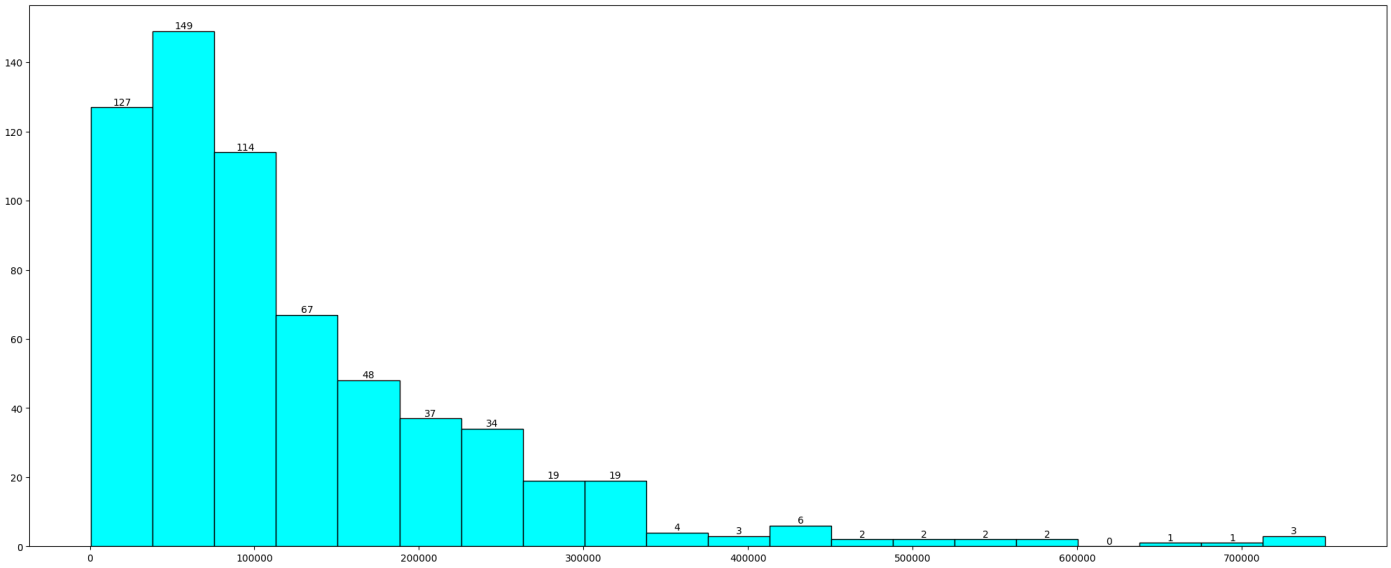


Figure 53: Plot of Tot_F (Total population of Female)

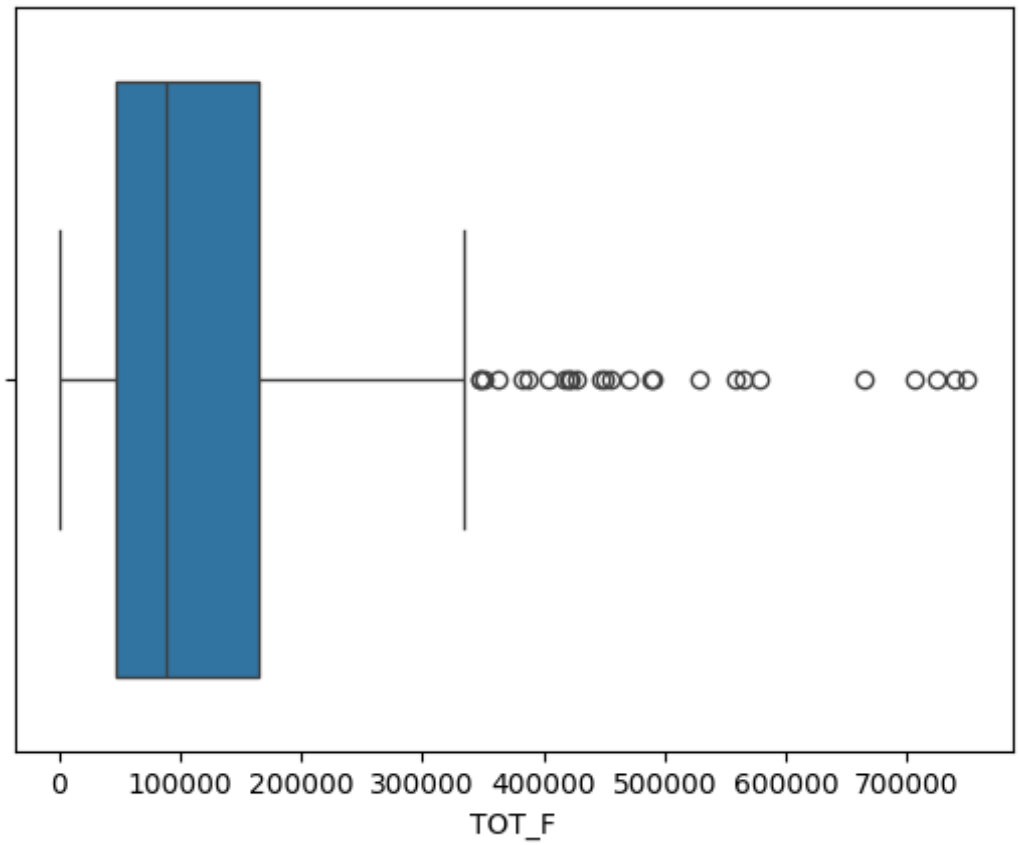


Figure 54: Boxplot of Tot_F (Total population of Female)

```
count      640.000000
mean     12309.098438
std      11500.906881
min         56.000000
25%       4733.750000
50%       9159.000000
75%      16520.250000
max      96223.000000
```

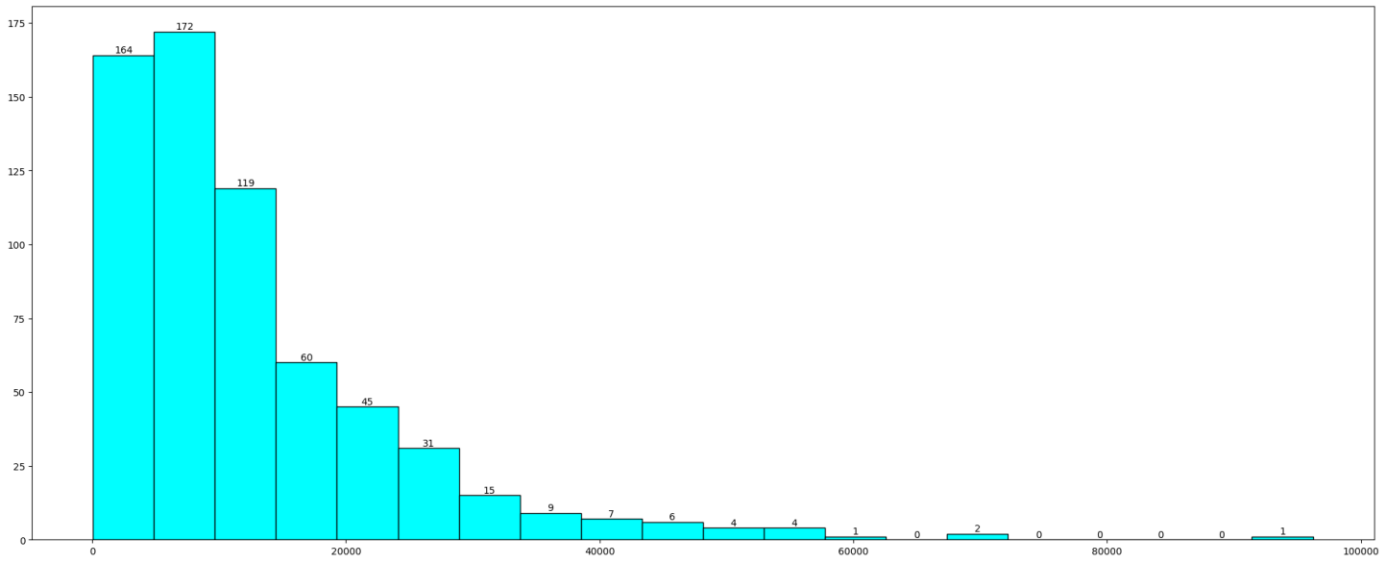


Figure 55: Plot of M_06 (Population of male in age group 0-6)

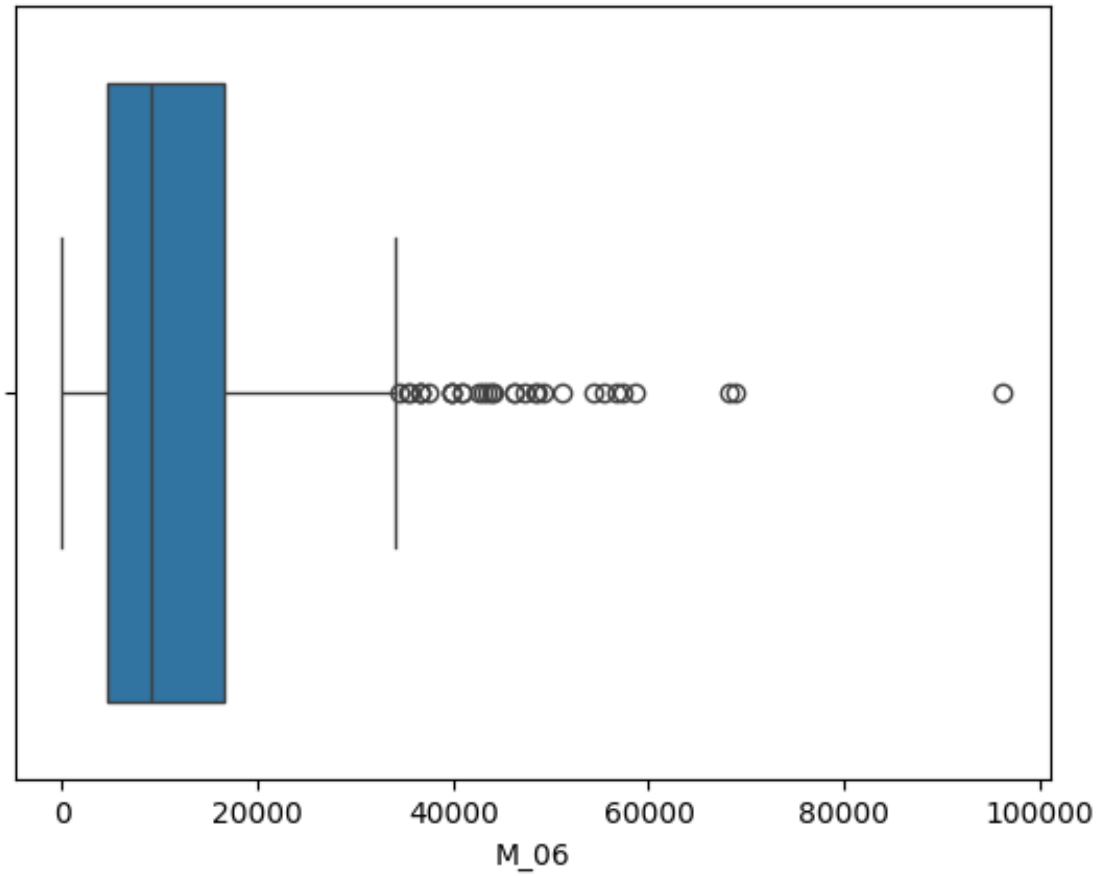


Figure 56: Boxplot of M_06 (Population of male in age group 0-6)

```
Description of F_06
-------------------------------------------------------------------------
--
count      640.000000
mean     11942.300000
std      11326.294567
min         56.000000
25%       4672.250000
50%       8663.000000
75%      15902.250000
max      95129.000000
```

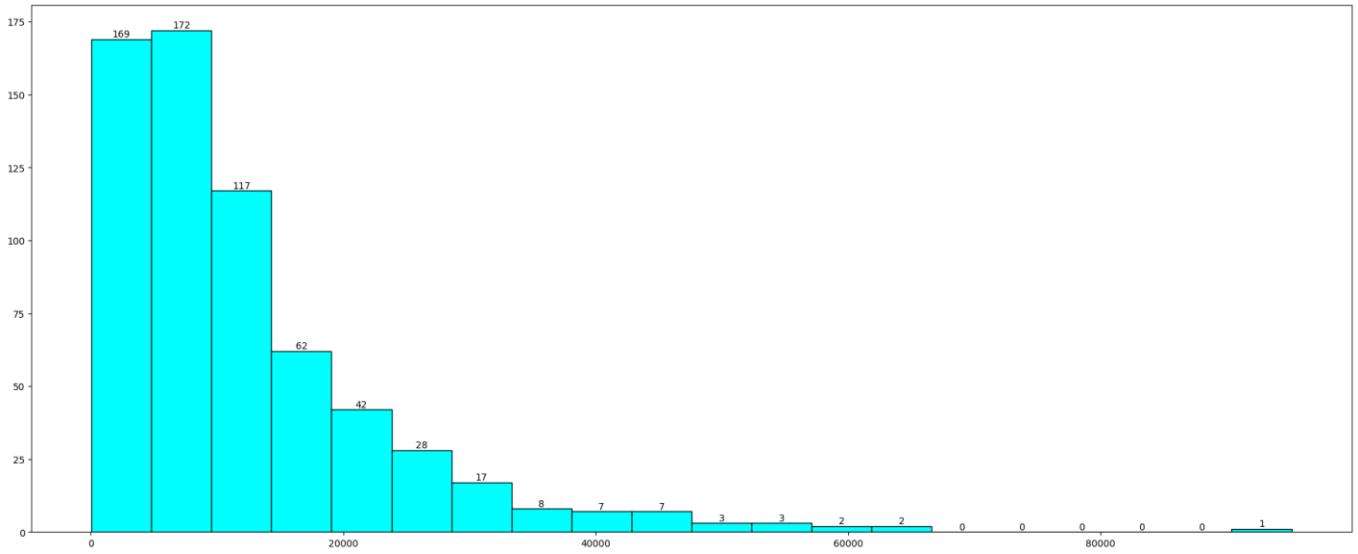


Figure 57: Plot of F_06 (Population of Female in age group 0-6)

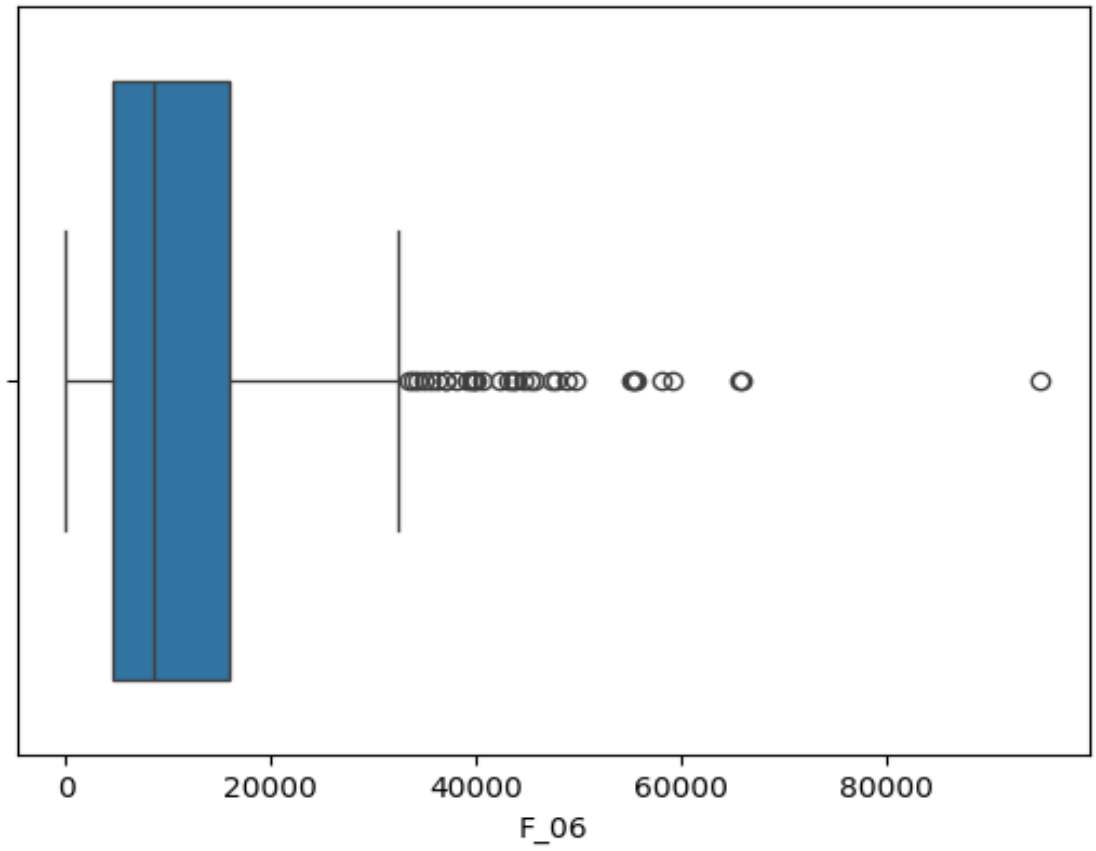


Figure 58: Boxplot of F_06 (Population of Female in age group 0-6)

Observations

- We have taken 5 variables as instructed as No_HH, TOT_M, TOT_F, M_06 and F_06
- The mean of No_HH is 51222.871875, the least number of No_HH is 350 and maximum No_HH is more than 3L.
- The average of male population is lesser than Female population.
- The minimum male population is 391 and maximum number is 485417, This is because of the geography of state.
- The minimum female population is 698 and maximum number is 750392, This is because of the geography of state and their area.
- The minimum number of male and female in the range of 0-6 is same i.e. 56. Their averages are also closed, 12309 for male and 11942 for female.

Bivariate Analysis

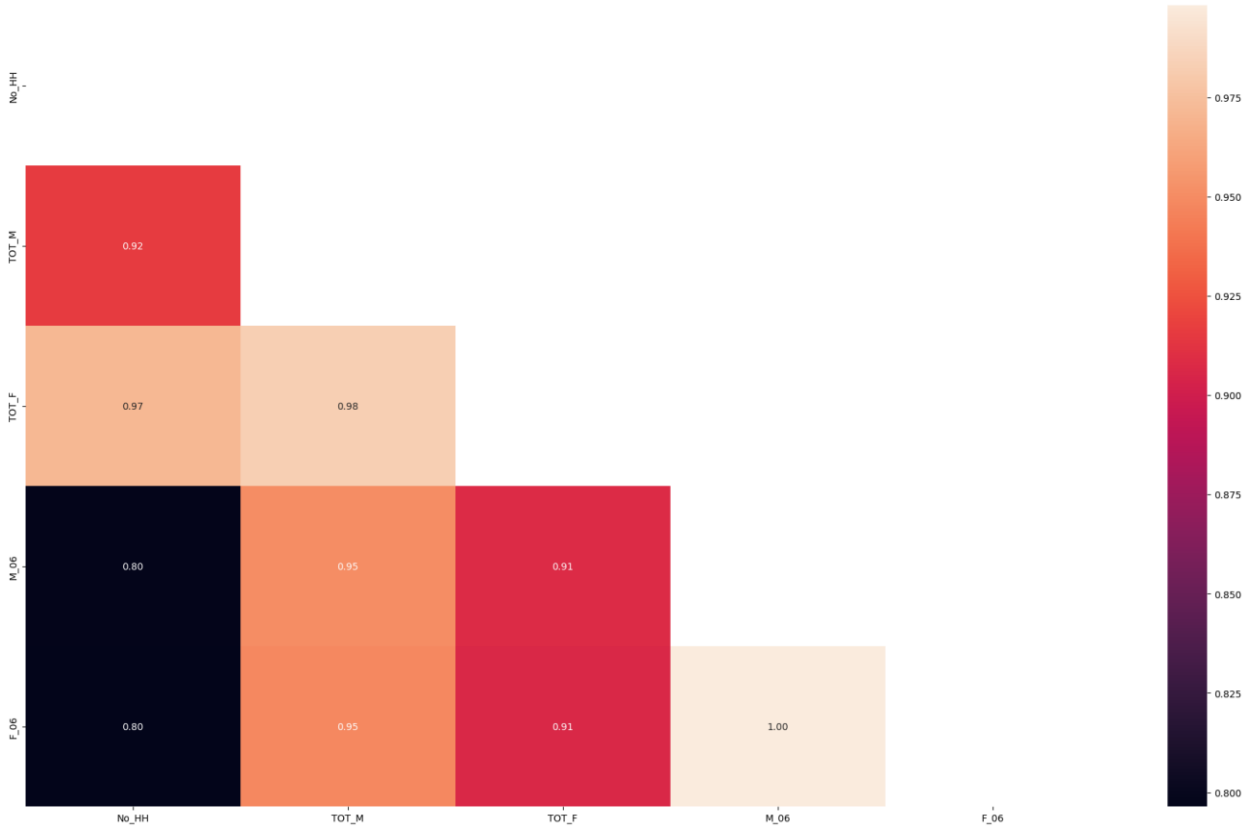


Figure 59: Correlation matrix between selected variables

Observations

- There is a strong correlation of total number of females with total number of male and no. of household.
- There is a correlation between total number of male and number of male in the range of age 0-6.

(i) Which state has highest gender ratio and which has the lowest?

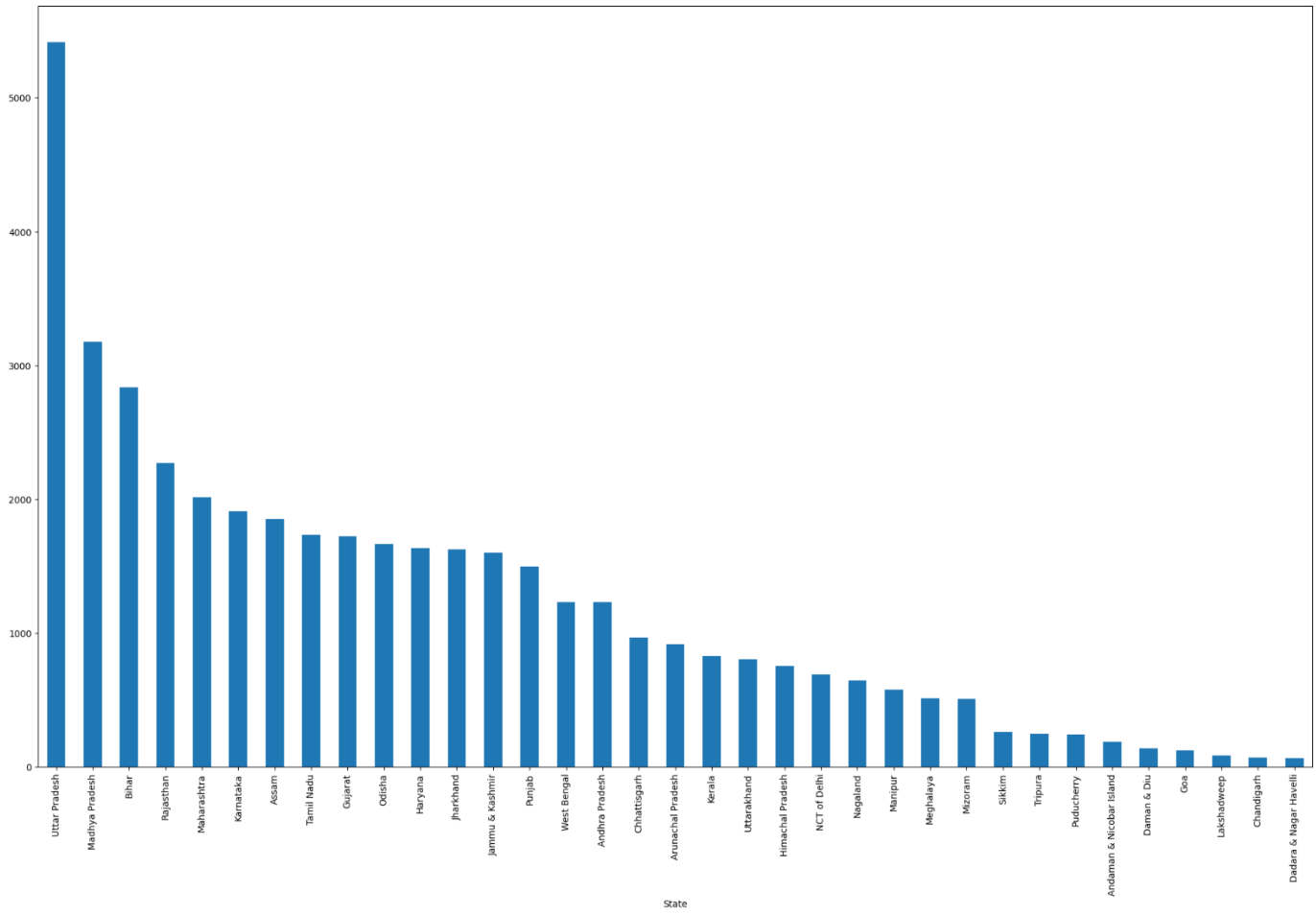


Figure 60: Bar plot of ratio gender state wise

The above graph clearly shows that state Uttar Pradesh has the highest gender ratio and Dadara & Nagar Havelli has the lowest gender ratio.

(ii) Which district has the highest & lowest gender ratio?

State	Area Name	
Lakshadweep	Lakshadweep	86.806120
Jammu & Kashmir	Badgam	84.776210
Uttar Pradesh	Mahamaya Nagar	84.731286
Rajasthan	Dhaulpur	84.691142
Uttar Pradesh	Baghpat	84.400265
		...
Odisha	Baudh	45.145505
Andhra Pradesh	West Godavari	45.007568
Tamil Nadu	Virudhunagar	44.935161
Odisha	Koraput	44.076873
Andhra Pradesh	Krishna	43.797226

Figure 61: Higher and lower gender ratio

Area Lakshadweep has the highest number of gender ratio and Krishna of Andhra Pradesh has the lowest number of gender ratio.

Data Preprocessing

There are no missing values present in data.

Before performing further operations, note that we dropped four columns from the dataframe (State Code, Dist.Code, State and Area Name.

Check outliers before scale data

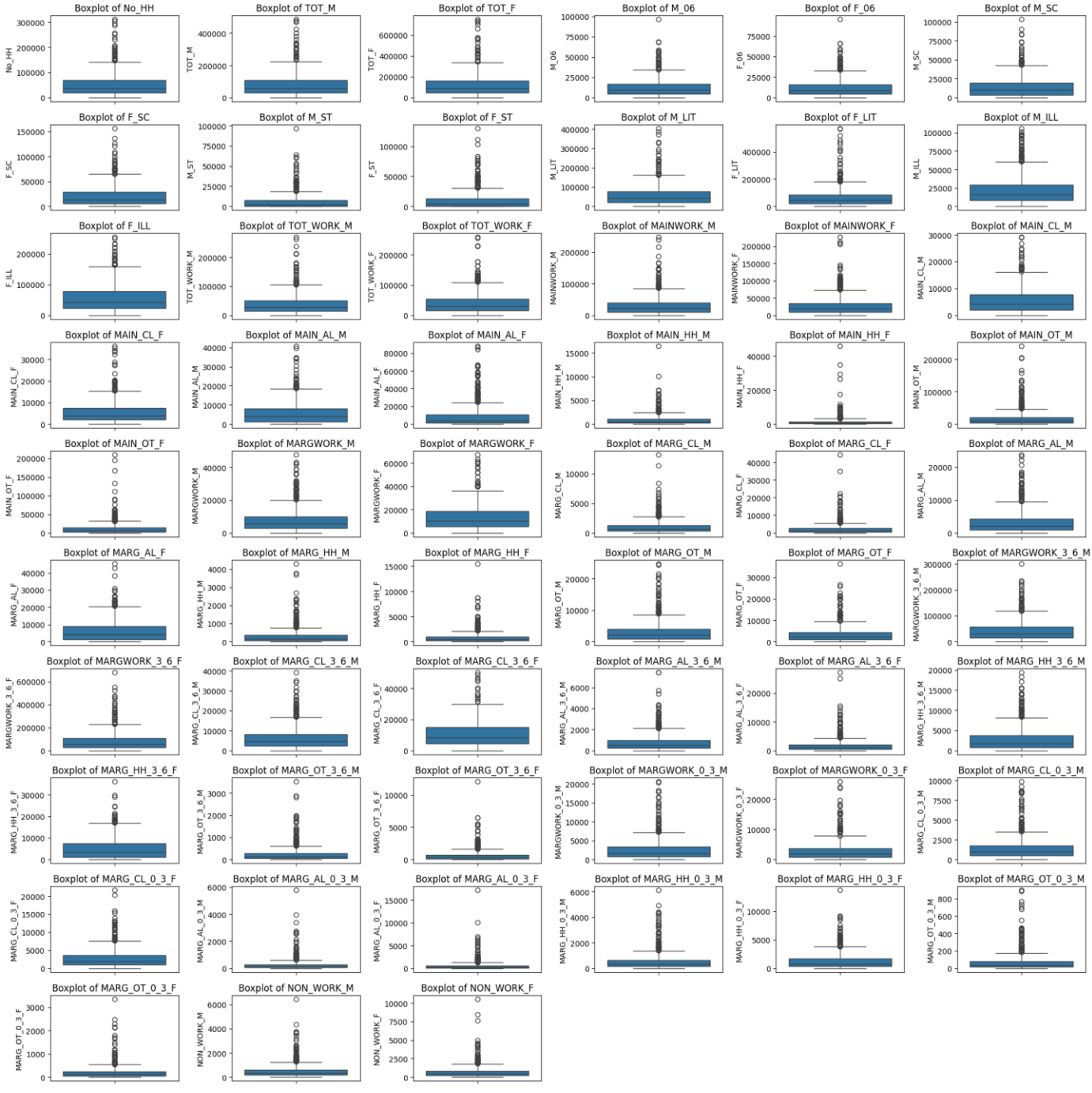


Figure 62: Checking outliers

Scale the Data using the z-score method

	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	F_ST	M_LIT	...	MARG_CL_0_3_M	MARG_CL_0_3_F	MARG_AL_0_3_M	I
0	-0.904738	-0.771236	-0.815563	-0.561012	-0.507738	-0.958575	-0.957049	-0.423306	-0.476423	-0.798097	...	-0.163229	-0.720610	-0.156494	
1	-0.935695	-0.823100	-0.874534	-0.681096	-0.725367	-0.958297	-0.956772	-0.582014	-0.607607	-0.849434	...	-0.583103	-0.732811	-0.282327	
2	-0.972412	-1.000919	-0.981466	-0.976956	-0.965262	-0.958575	-0.956772	-0.038951	-0.027273	-0.956457	...	-0.859212	-0.921931	-0.456727	
3	-1.037530	-1.052224	-1.041001	-1.022118	-0.995393	-0.958783	-0.957049	-0.355965	-0.390060	-1.004643	...	-0.805468	-0.900758	-0.419198	
4	-0.822676	-0.809381	-0.813933	-0.622359	-0.649908	-0.957395	-0.955529	0.149238	0.043330	-0.800568	...	-0.348645	-0.297513	0.472670	

MARG_AL_0_3_F	MARG_HH_0_3_M	MARG_HH_0_3_F	MARG_OT_0_3_M	MARG_OT_0_3_F	NON_WORK_M	NON_WORK_F
-0.287524	0.156577	-0.657412	-0.365258	-0.499977	-0.413053	-0.539614
-0.294688	-0.491731	-0.723062	0.042855	-0.073481	-0.606455	-0.598988
-0.420050	-0.731894	-0.795026	-0.662068	-0.635680	-0.726103	-0.707839
-0.385127	-0.718770	-0.784926	-0.624966	-0.616294	-0.645791	-0.710038
0.434200	-0.466796	-0.625849	-0.439461	-0.309346	-0.540895	-0.249344

Table 20: Top 5 rows of the scaled dataset

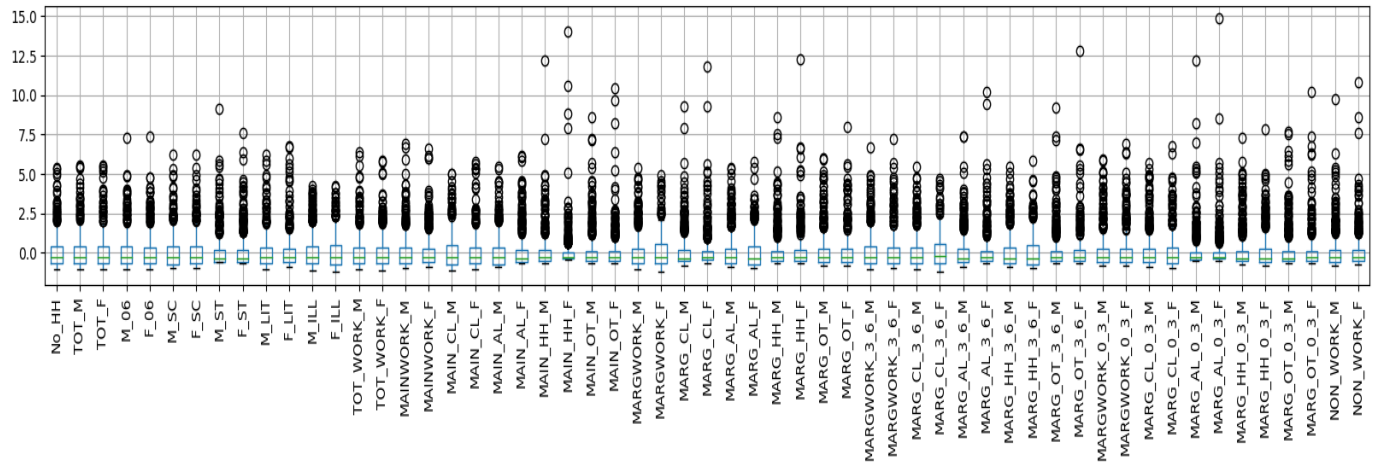


Figure 63: Scaled dataset

Check outliers after scale the data

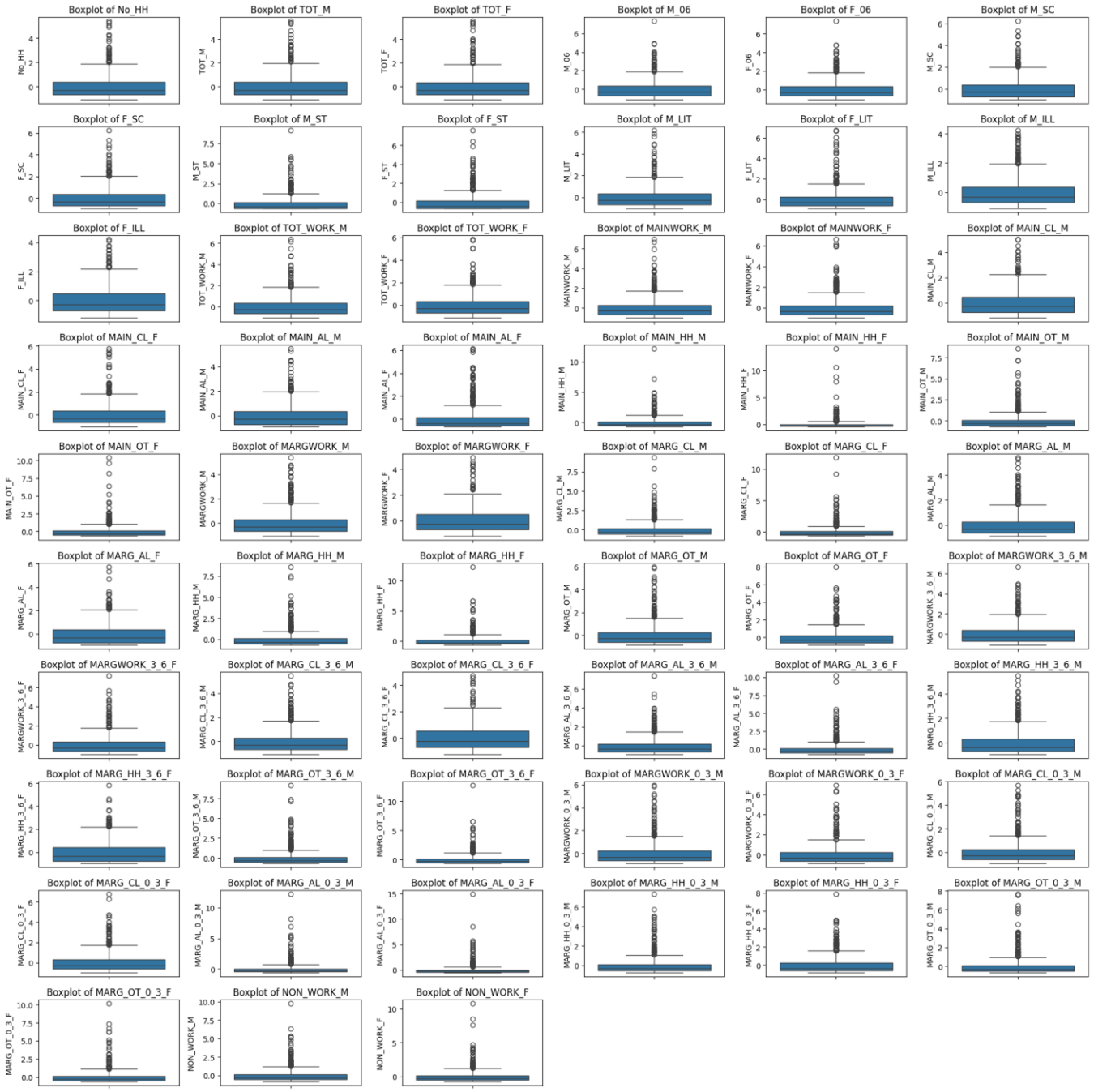


Figure 64: Box plots of all variables post scaled data

Insight

As per the graph, it shows that we have outliers in each variable. Since the data have been taken of state and their district area of India. So, some state's geographical area is bigger, so their population is higher, and some state's geographical area are smaller, so their population is smaller. Therefore, we should not treat the outliers since the data will be manipulated and it will impact on the analysis.

Principal Component Analysis

Statistical tests to be done before PCA

Bartletts Test of Sphericity Bartlett's test of sphericity tests the hypothesis that the variables are uncorrelated in the population.

Let's define the null hypothesis and alternate hypothesis.

H0: All variables in the data are uncorrelated
Ha: At least one pair of variables in the data are correlated

If the null hypothesis cannot be rejected, then PCA is not advisable. If the p-value is small, then we can reject the null hypothesis and agree that there is at least one pair of variables in the data which are correlated hence PCA is recommended.

We used the factor analyzer to calculate the P value of Bartletts Test. The P value is 0.0 which means that we reject the null hypothesis and accept the alternate hypothesis.

KMO Test The Kaiser-Meyer-Olkin (KMO) - measure of sampling adequacy (MSA) is an index used to examine how appropriate PCA is.

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

The MSA value is 0.8039889932781807. Since MSA value is higher than 0.7 then proceed with PCA

Create the covariance matrix - Get eigen values and eigen vectors.

We import the PCA class from sklearn.decomposition, create an instance of the PCA class with n_components=57, since we have 57 variables in our dataframe.

After transform the PCA, we see the array as below.

```
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: Arrays of PCA

Obtaining the Eigen Vectors when the Principal Components are kept exactly as the number of features in the scaled data.

```
Eigen Vectors
 %s [[ 0.16  0.17  0.17 ...  0.13  0.15  0.13]
 [-0.13 -0.09 -0.1  ...  0.05 -0.07 -0.07]
 [-0.    0.06  0.04 ... -0.08  0.11  0.1 ]
 ...
 [ 0.    0.38  0.15 ...  0.03 -0.08 -0.03]
 [-0.    0.24  0.09 ... -0.03 -0.02  0.04]
 [-0.   -0.09 -0.01 ...  0.01 -0.01 -0.  ]]
```

Table 22: Eigner vectors

Obtaining the Explained Variance

```
[0.56 0.14 0.07 0.06 0.04 0.03 0.02 0.01 0.01 0.01 0.01 0.01 0.01 0.
 0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.
 0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.
 0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.
 0.  ]
```

Table 23: Explained Variance

```
Cumulative Variance Explained in Percentage:

 [ 55.73  69.51  76.79  83.21  87.08  90.47  92.53  93.85  94.93  95.85
  96.61  97.23  97.75  98.24  98.57  98.81  99.01  99.2   99.37  99.51
  99.61  99.69  99.75  99.81  99.85  99.89  99.92  99.94  99.96  99.97
  99.98  99.99 100.   100.   100.   100.   100.   100.   100.   100.
 100.   100.   100.   100.   100.   100.   100.   100.   100.   100.
 100.   100.   100.   100.   100.   100.   100.  ]
```

Table 24: Cumulative explained Variance

Insight:

- We can see above that more than 90% of the variance is explained by 6 Principal Components.
- Around 93% of the variance is explained by 7 Principal Components.
- Around 97% of the variance is explained by 11 Principal Components.

Scree plot

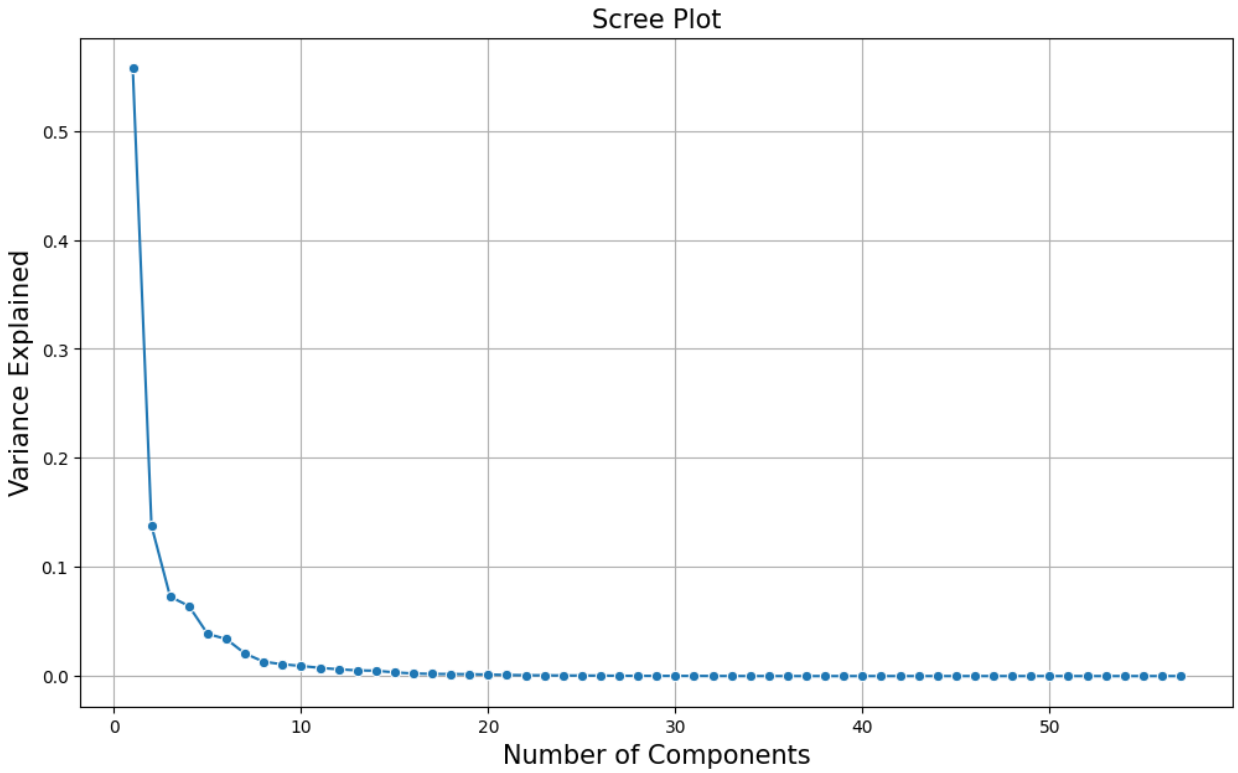


Figure 65: Scree plot

The number of components can be decided based upon the explained variance. Here, it is decided to keep the number of components as 6 as the cumulative explained variance is around 90%

Apply PCA for the number of decided components to get the loadings and component output

Component output

```
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 25: Component output

Loading of each feature on the components, Eigen Vectors when PC's are kept as 6.

```
array([[ 0.16,  0.17,  0.17,  0.16,  0.16,  0.15,  0.15,  0.03,  0.03,
         0.16,  0.15,  0.16,  0.17,  0.16,  0.15,  0.15,  0.12,  0.1 ,
         0.07,  0.11,  0.07,  0.13,  0.08,  0.12,  0.11,  0.16,  0.16,
         0.08,  0.05,  0.13,  0.11,  0.14,  0.13,  0.16,  0.15,  0.16,
         0.16,  0.17,  0.16,  0.09,  0.05,  0.13,  0.11,  0.14,  0.12,
         0.15,  0.15,  0.15,  0.14,  0.05,  0.04,  0.12,  0.12,  0.14,
         0.13,  0.15,  0.13],
       [-0.13, -0.09, -0.1 , -0.02, -0.02, -0.05, -0.05,  0.03,  0.03,
        -0.12, -0.15, -0.01, -0.01, -0.13, -0.09, -0.18, -0.15,  0.06,
         0.09, -0.03, -0.06, -0.08, -0.08, -0.21, -0.21,  0.09,  0.13,
         0.27,  0.25,  0.17,  0.14,  0.07,  0.02, -0.09, -0.12, -0.04,
        -0.11,  0.08,  0.1 ,  0.26,  0.24,  0.16,  0.13,  0.06,  0.01,
        -0.09, -0.13,  0.15,  0.18,  0.25,  0.24,  0.19,  0.18,  0.08,
         0.05, -0.07, -0.07],
       [-0.  ,  0.06,  0.04,  0.06,  0.05,  0.  , -0.03, -0.12, -0.14,
         0.08,  0.12, -0.02, -0.09,  0.05, -0.06,  0.05, -0.06, -0.07,
        -0.01, -0.25, -0.25,  0.03, -0.06,  0.14,  0.1 , -0.01, -0.05,
         0.2 ,  0.27, -0.19, -0.27, -0.02, -0.08,  0.11,  0.1 ,  0.06,
         0.08, -0.02, -0.07,  0.15,  0.26, -0.2 , -0.28, -0.02, -0.08,
         0.11,  0.1 ,  0.05,  0.02,  0.27,  0.28, -0.14, -0.2 , -0.02,
        -0.08,  0.11,  0.1 ],
       [-0.13, -0.02, -0.07,  0.01,  0.01,  0.01, -0.03, -0.22, -0.23,
        -0.04, -0.06,  0.03, -0.08, -0.04, -0.23, -0.07, -0.25, -0.09,
        -0.29, -0.14, -0.29,  0.15,  0.05, -0.04, -0.12,  0.09, -0.09,
        -0.06, -0.17,  0.09, -0.11,  0.24,  0.2 ,  0.09,  0.03, -0.  ,
         0.  ,  0.09, -0.11, -0.04, -0.18,  0.08, -0.14,  0.24,  0.19,
         0.09,  0.03,  0.09, -0.02, -0.1 , -0.14,  0.13,  0.  ,  0.23,
         0.21,  0.08,  0.02],
       [-0.01, -0.03, -0.01, -0.05, -0.04, -0.17, -0.16,  0.43,  0.44,
        -0.01,  0.06, -0.1 , -0.12, -0.02, -0.04, -0.04, -0.08, -0.29,
        -0.24, -0.21, -0.18, -0.13, -0.14,  0.06,  0.08,  0.06,  0.09,
```

Table 26: Eigen Vectors when PC's are kept as 6

Final experienced variances are array ([0.56, 0.14, 0.07, 0.06, 0.04, 0.03])

Let's create a dataframe of component loading against each field and identify the pattern.

Let's identify which features have maximum loading across the components. We will first plot the component loading on a heatmap. For each feature, we find the maximum loading value across the components and mark the same with help of rectangular box. Features marked with rectangular red box are the one having maximum loading on the respective component. We consider these marked features to decide the context that the component represents.

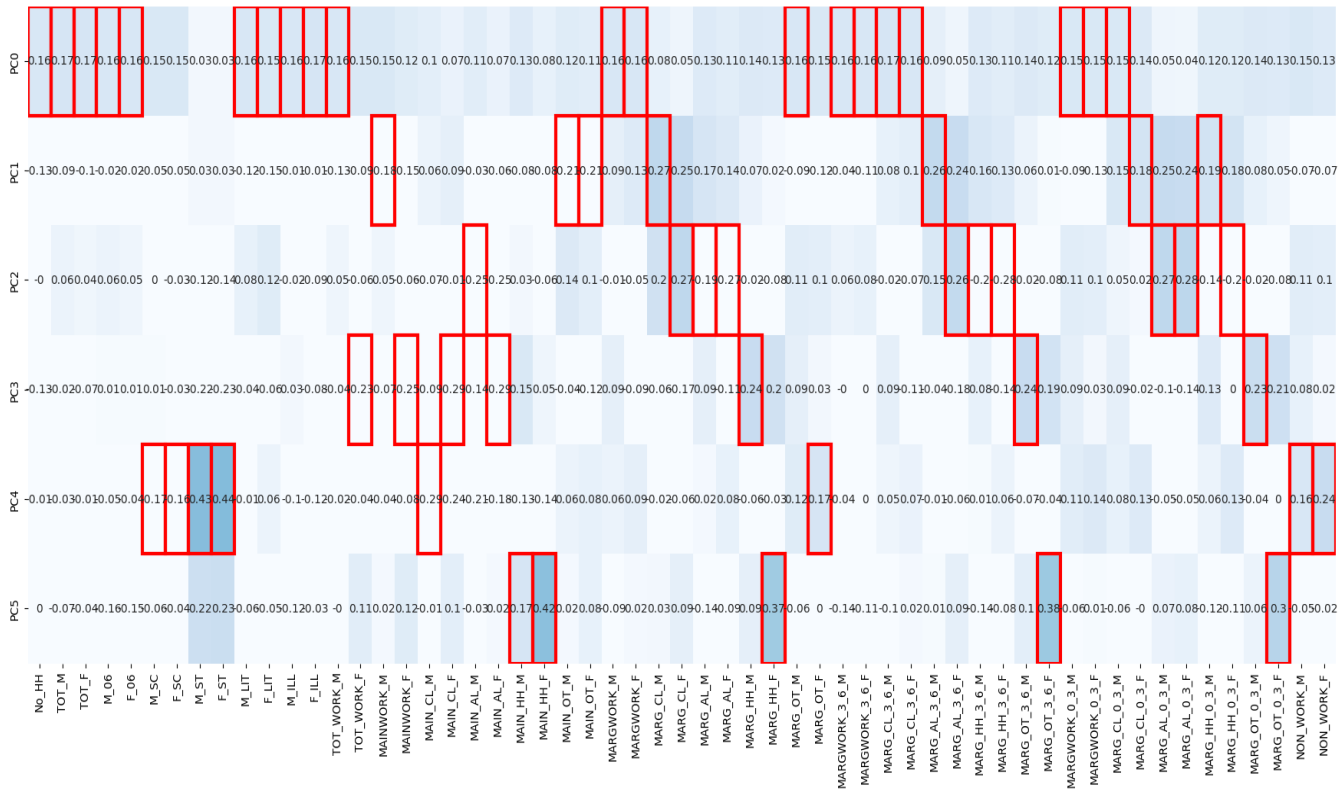


Figure 66: Component loading on heatmap

For every attribute (column, the corresponding PC's cell with the maximum magnitude has been highlighted using Rectangles. For Example, for the No_HH attribute, the PC0 has the maximum magnitude of 0.16.

Insight

We could see in the heatmap that each PCA has highlighted some high value in the boxes.

- PCA0 picks up certain essence of the information related around the number of households, the population of male and female,, Their Literates population and illiterates' population, total working male, marginal working male and female, marginal working age range between 0-3, 3-6 of male and female, marginal cultivators age between 3-6 of male and female and marginal cultivators age between 0-3 of male only. PCA0 contains 20 features.

- PCA1 has picked up in terms of the population of main work of male and main other workers population of male and female, marginal population of Cultivator of male. Marginal agriculture age range 3-6 of male, Marginal population of household of male age between 0-3 and marginal population of Cultivator of female with age group 0-3. PCA0 contains 07 features.

- PCA2 contains 10 features, it contains most of the information about marginal workers. Population of the marginal Cultivator and agricultural laborers of female. For agricultural laborers, it also contain their age range between 0-3, 3-6. Population of marginal female workers in household industries with age range 0-3, 3-6 . for male, it contains the population of marginal Cultivator, marginal household workers with age range of 3-6 and the population of age range 0-3 of agricultural laborers of male.

- PCA3 contains 7 variables, and they are in the direction of total working female and main working female as cultivator and agricultural labor and population of marginal household male, marginal other works with range of 0-3, 0-6.

- PCA4 contains 8 features, it contains the information about the population of schedule caste of male and female, schedule tribe of male and female. Population of main Cultivator of male and population of marginal other of female. And Population of non-working male and female.

- PCA 5 contains 5 features, it contains the information about Main Household Industries population of male and female, also contain the information for Main Household Industries Population Female and Marginal Other Workers Population Person 0-3 and 3-6 Female.

Concatenate the PCA information dataframe with four dropped columns which were copied in another dataframe earlier.

	State Code	Dist.Code	State	Area Name	PCA0	PCA1	PCA2	PCA3	PCA4	PCA5
0	1	1	Jammu & Kashmir	Kupwara	-4.62	0.14	0.33	1.54	0.35	-0.42
1	1	2	Jammu & Kashmir	Badgam	-4.77	-0.11	0.24	1.96	-0.15	0.42
2	1	3	Jammu & Kashmir	Leh(Ladakh)	-5.96	-0.29	0.37	0.62	0.48	0.28
3	1	4	Jammu & Kashmir	Kargil	-6.28	-0.50	0.21	1.07	0.30	0.05
4	1	5	Jammu & Kashmir	Punch	-4.48	0.89	1.08	0.54	0.80	0.34

Table 27: PCA data with State and area name

Write linear equation for first PC

Each principal component is a linear combination of original features. We can write the equation for PC1 in the following manner:

0.16* No_HH+ 0.17* TOT_M+ 0.17* TOT_F+ 0.16* M_06+0.16* F_06+ 0.15* M_SC+0.15* F_SC+0.03*M_ST+0.03*F_ST+0.16*M_LIT+0.15*F_LIT+0.16*M_ILL+0.17*F_ILL+0.16*TOT_WORK_M+0.17*TOT_WORK_F+0.15*MAINWORK_M+0.12*MAINWORK_F+0.10*MAIN_CL_M+0.07*MAIN_CL_F+0.11*MAIN_AL_M+0.07*MAIN_AL_F+0.13*MAIN_HH_M+0.08*MAIN_HH_F+0.12*MAIN_OT_M+0.11*MAIN

_OT_F+0.16*MARGWORK_M+0.16*MARGWORK_F+0.08*MARG_CL_M+0.05*MARG_CL_F+0.13*MARG_AL_M+0.11*MARG_AL_F+0.14*MARG_HH_M+0.16*MARG_OT_M+0.15*MARG_OT_F+0.16*MARGWORK_3_6_M+0.16*MARGWORK_3_6_F+0.17*MARG_CL_3_6_M+0.16*MARG_CL_3_6_F+0.09*MARG_AL_3_6_M+0.05*MARG_AL_3_6_F+0.13*MARG_HH_3_6_M+0.11*MARG_HH_3_6_F+0.14*MARG_OT_3_6_M+0.12*MARG_OT_3_6_F+0.15*MARGWORK_0_3_M+0.015*MARGWORK_0_3_F+0.015*MARG_CL_0_3_M+0.14*MARG_CL_0_3_F+0.05*MARG_AL_0_3_M+0.04*MARG_AL_0_3_F+0.12*MARG_HH_0_3_M+0.12* MARG_HH_0_3_F+0.14*MARG_OT_0_3_M+0.13* MARG_OT_0_3_F+0.15* NON_WORK_M+0.13* NON_WORK_F

Report the cumulative explained variance of the retained principal components. - Discuss how much of the total variance is captured by the selected principal components

Cumulative Variance Explained in Percentage:

[56. 70. 77. 83. 87. 90.]

PCA0 contains the 56% variance.

PCA1 contains 14% variance, the cumulative variance is 70%.

PCA2 contains 7% variance, the cumulative variance is 77%.

PCA3 contains 6% variance, the cumulative variance is 83%.

PCA4 contains 4% variance, the cumulative variance is 87%.

PCA5 contains 3% variance, the cumulative variance is 90%.

Selected PCA are 6 and 90% variance is captured by them.