# Machine Learning - 1

# Project Business Report

Name: Aishwariya Hariharan
PGP-DSBA Online September’ 23
Date: 21/01/2024

# Contents

# Part 1

## Problem:

We need to analyse the data collected by the Ads24x7 digital marketing company, use clustering methods, and provide actionable recommendations on how to optimize their digital marketing efforts, allocate budgets efficiently, and tailor ad content to specific audience segments.

## About the data set:

- Number of rows = 23066
- Columns and the datatypes

```
 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
```

- CTR
    - CTR stands for "Click through rate". CTR is the number of clicks that an 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.

- CPM
    - CPM stands for "cost per 1000 impressions."
    - Formula used here is CPM = (Total Campaign Spend / Number of Impressions) * 1000.
- 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.

- Statistical summary of data

| | 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 |

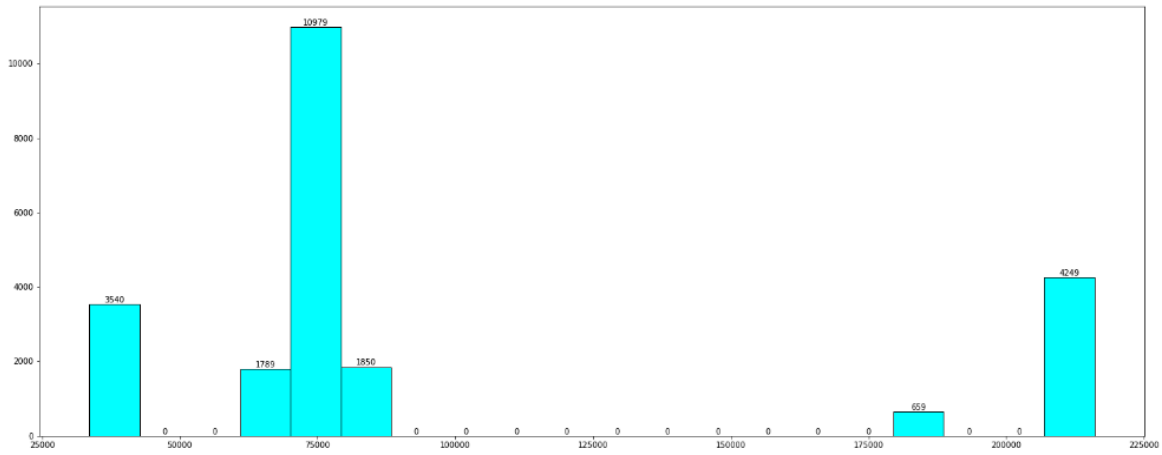
## Observations:

- There are 7 Inventory types: Format 1, Format 2, Format 3, Format 4, Format 5, Format 6, and Format 7.

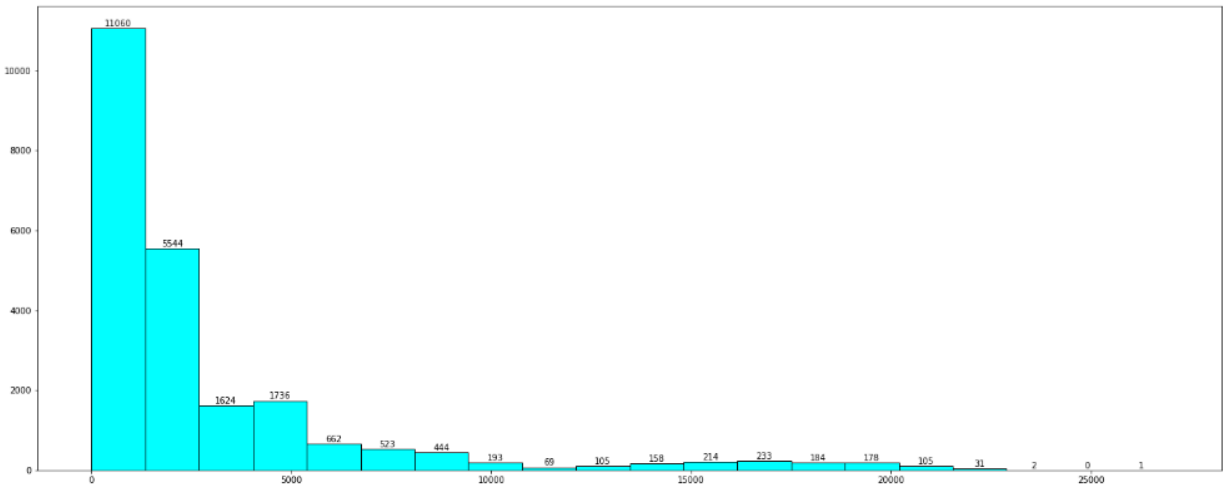
- There are 3 types of platforms: Video, Web, App

- There are 2 Formats: Video and Display

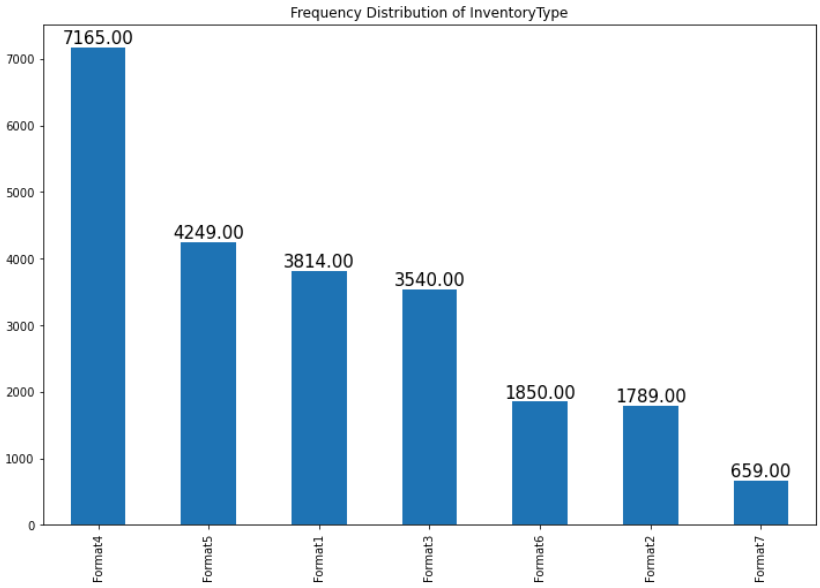
- There are no duplicate values
- About key variables:
    - The areas of the maximum number of Ad-sizes are about 75000 sq units



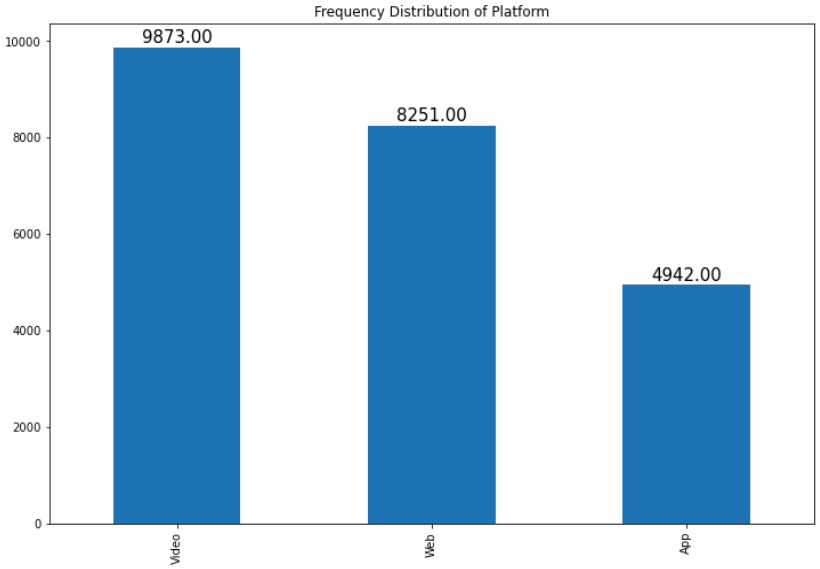

    - The maximum spend and revenue ranges are below 5000.



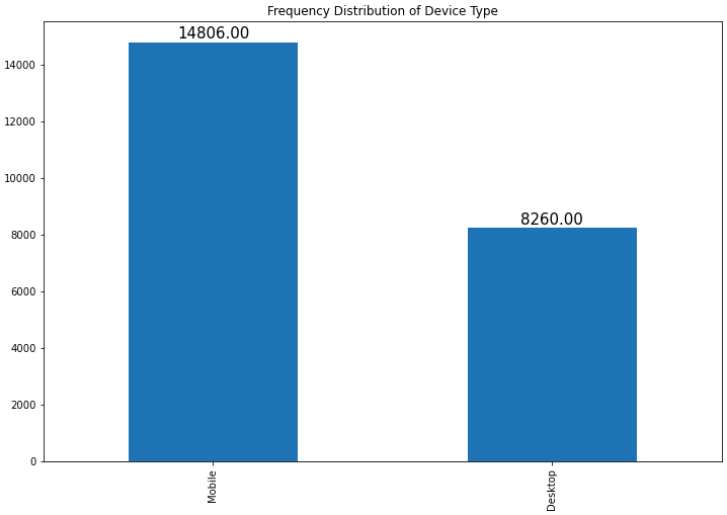
- Frequency of distribution of inventory type



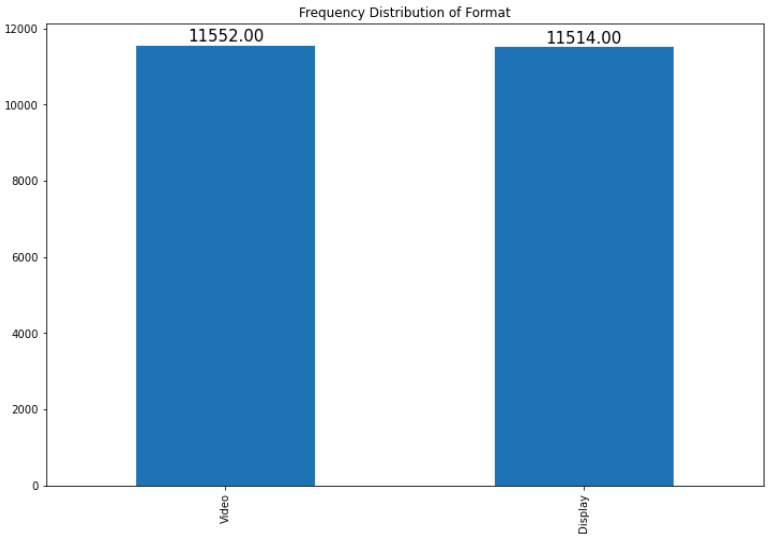

- Frequency of distribution of platforms



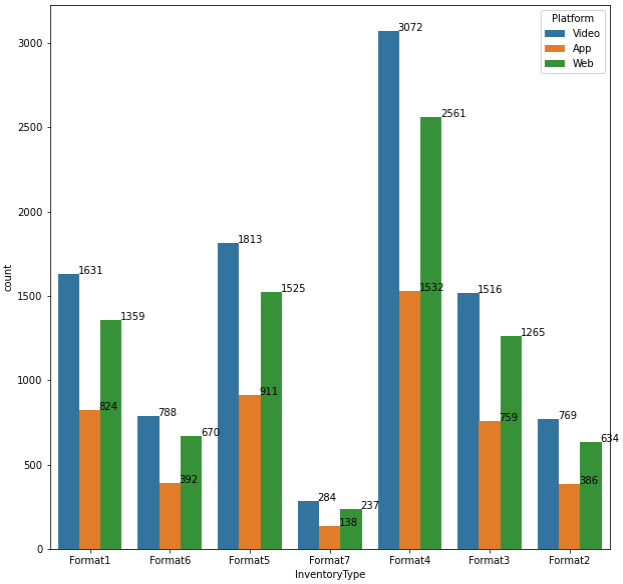
- Frequency of distribution of device type



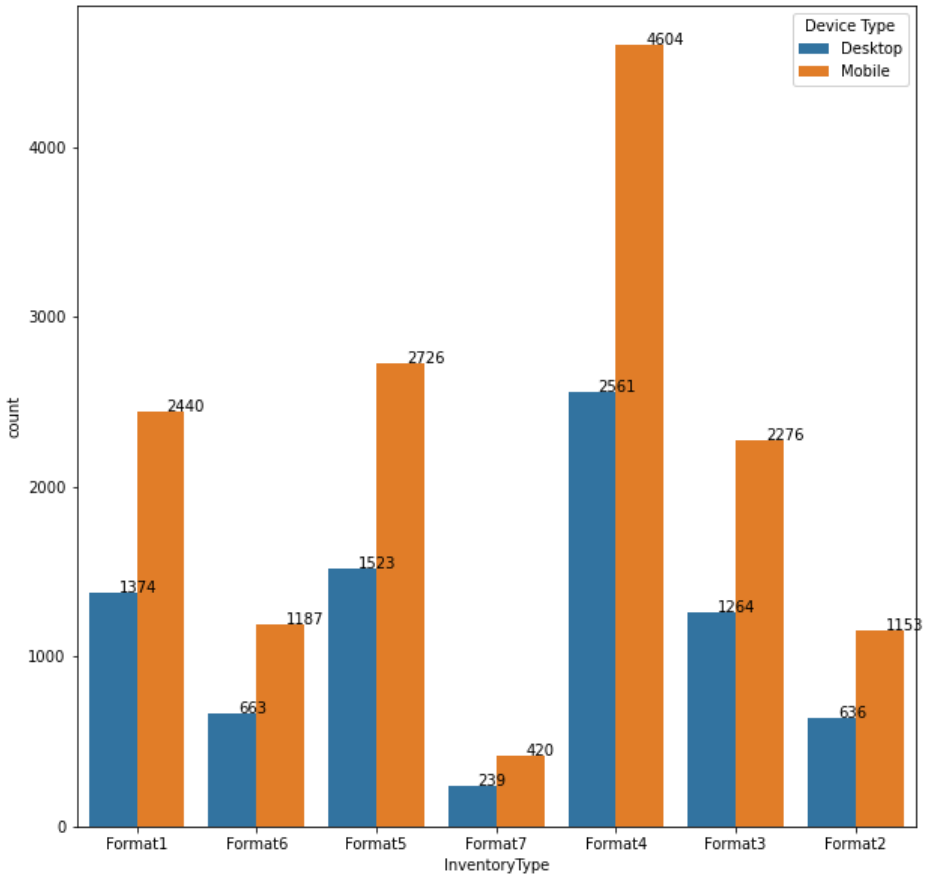

- Frequency of distribution of format



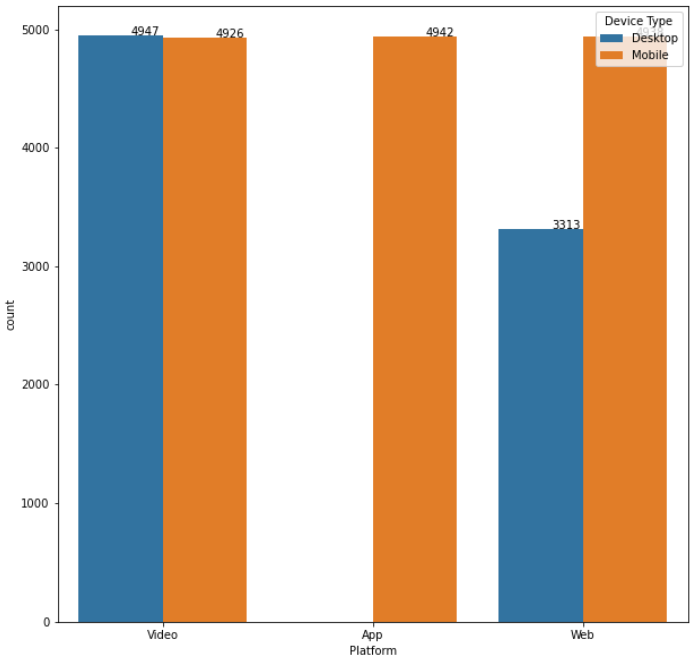
- Distribution of inventory type across platforms



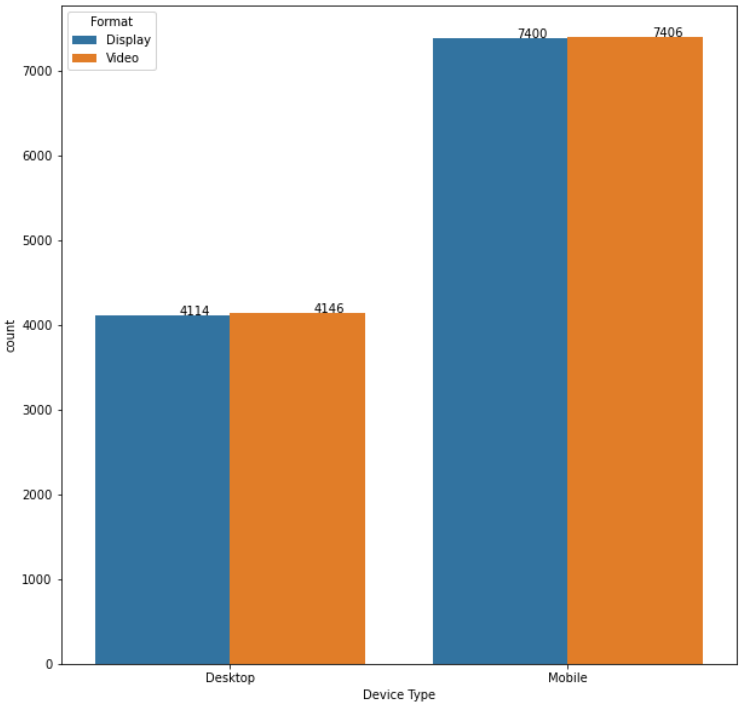

- Distribution of inventory type based on device types



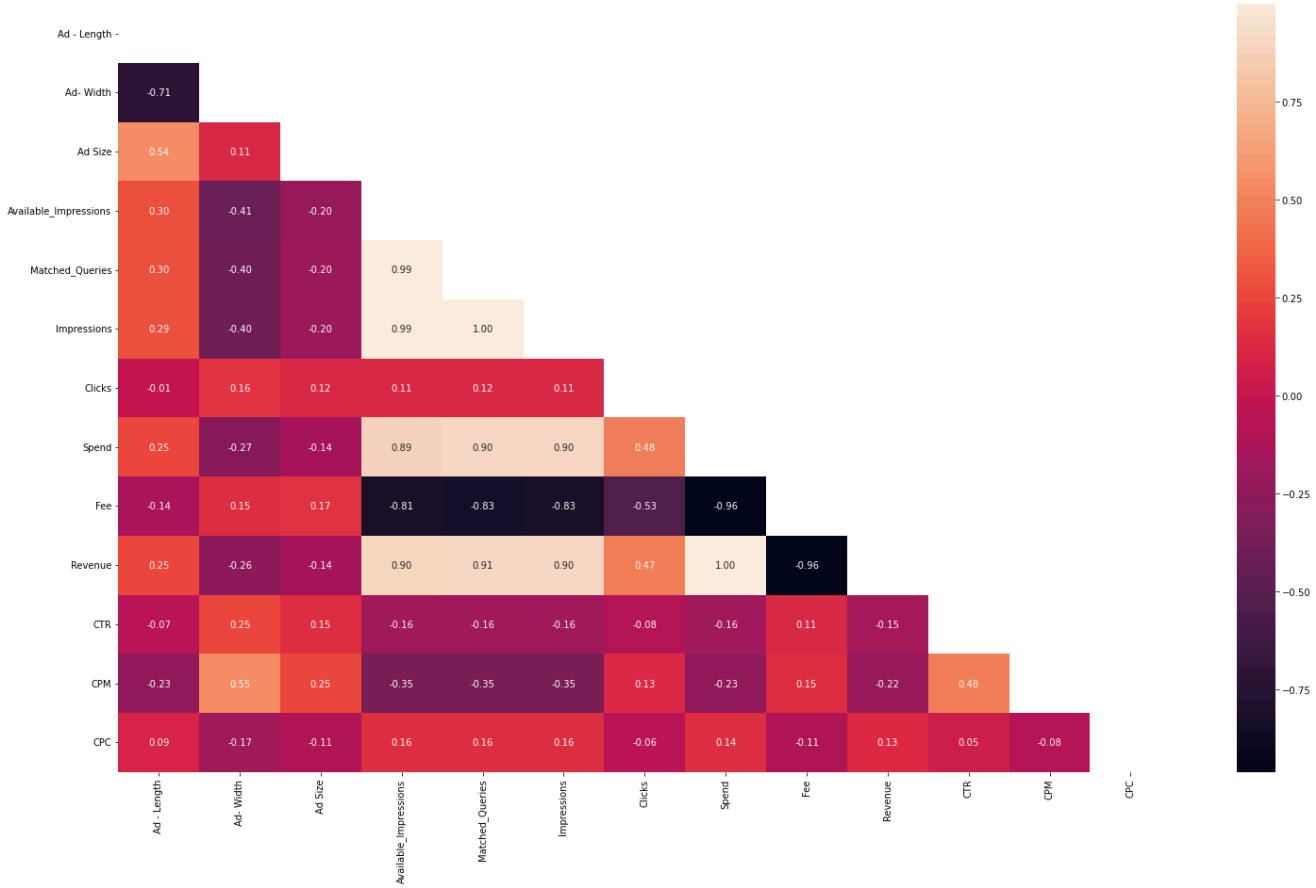
- Device types used across different platforms




- Format used across devices



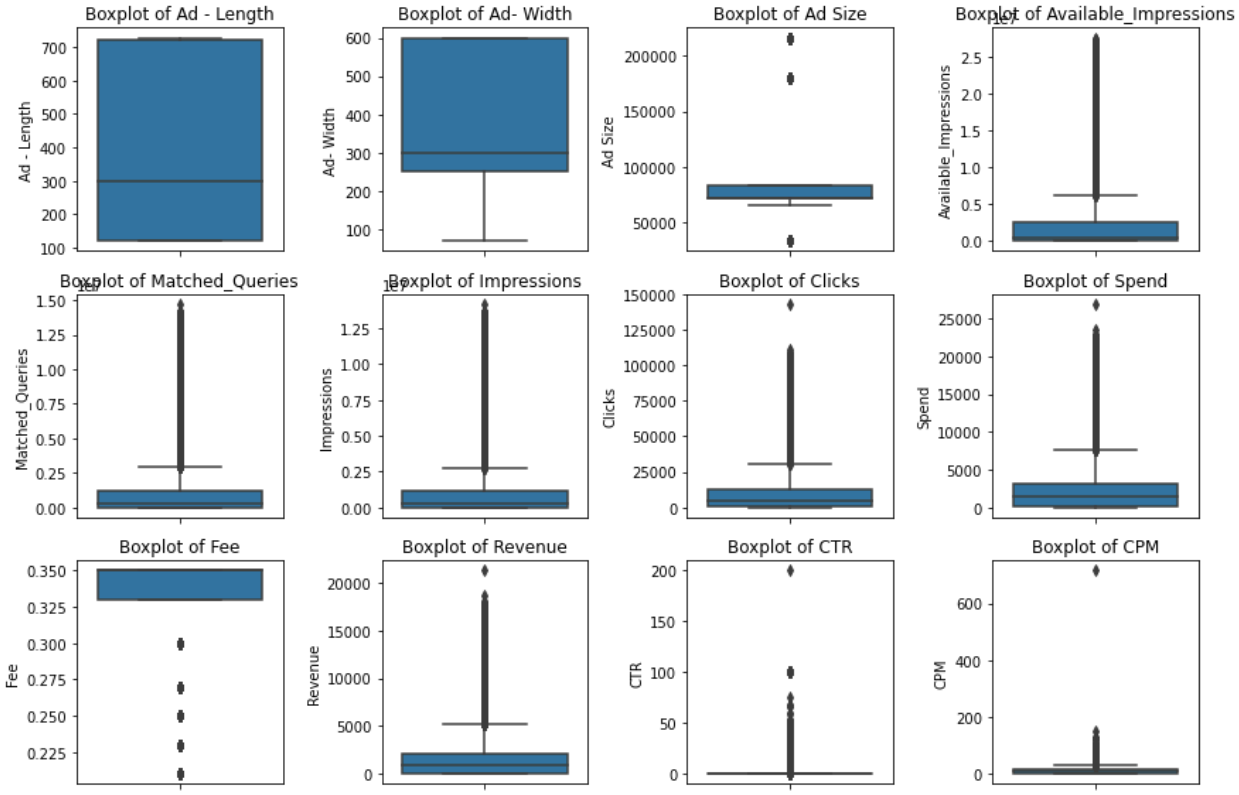
- Correlation between variables



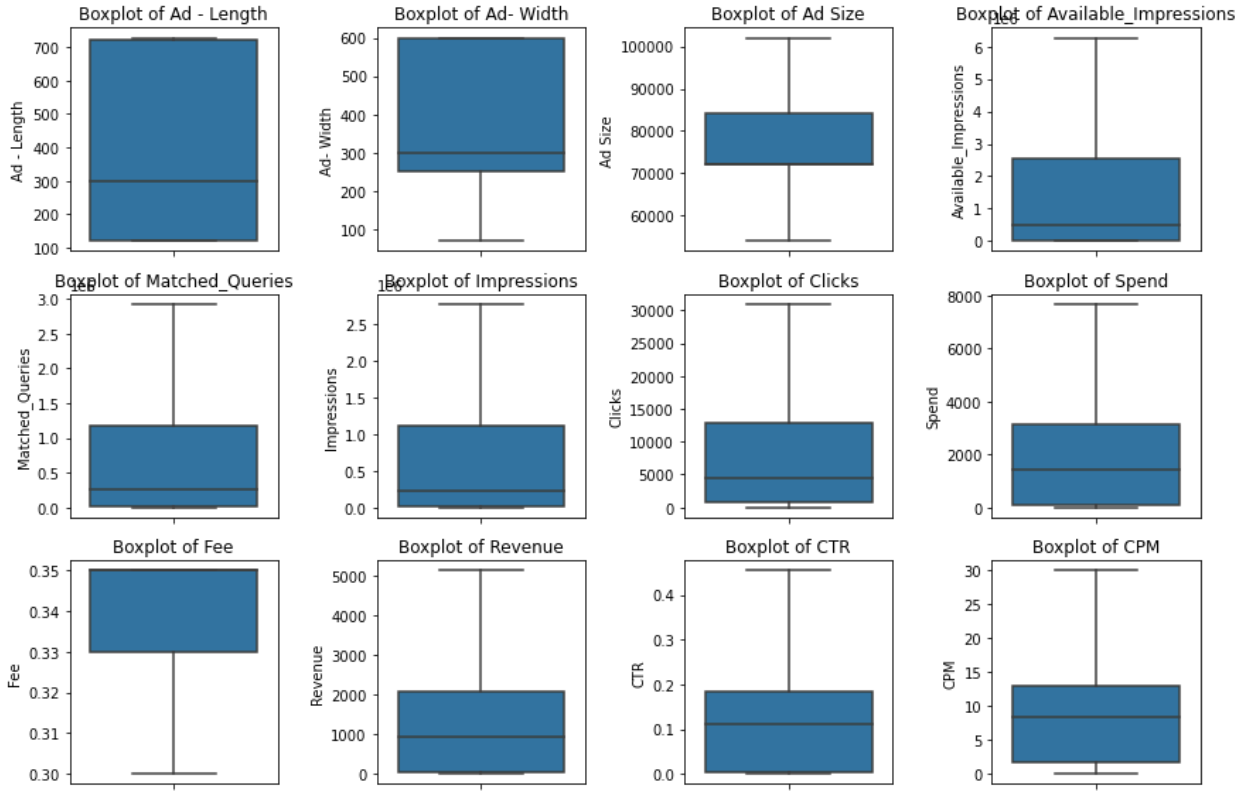

- There's a high correlation between:
  - The matched queries and available impressions.
  - Revenue and available impressions as well as matched queries and Impressions
  - Spend and Revenue

# Data preprocessing

- The data has outliers.
    - Only the Ad-length and Ad-width columns do not have outliers.



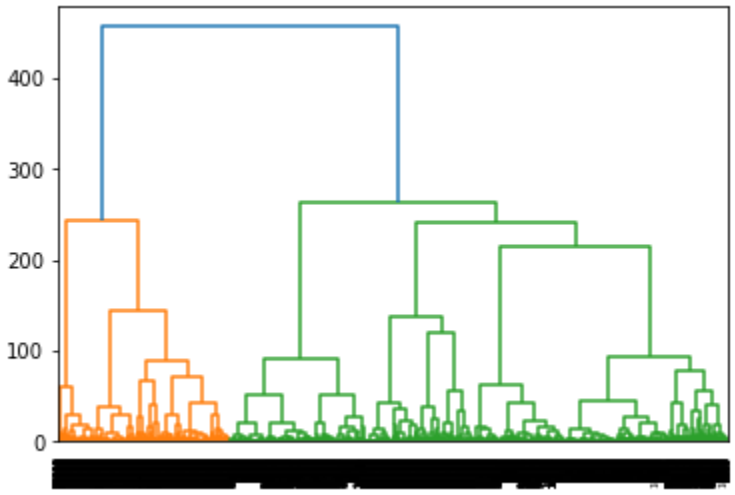
- The outliers are treated using the quantile method



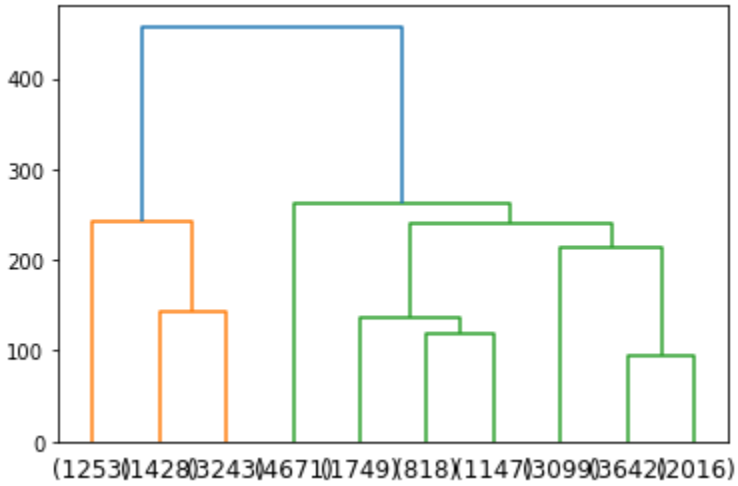

- The variables are scaled using the z-score method.
    - For k-means clustering and PCA, it's generally recommended to use Z-score scaling (standardization).
    - Z-score scaling standardizes the features by centering them around their mean and scaling them by their standard deviation. This ensures that all features contribute equally to the distance calculations, making k-means clustering more robust and effective.
    - Z-score scaling is often applied before PCA to ensure that all features have comparable scales. Standardizing the features helps PCA identify the directions (principal components) of maximum variance more accurately, leading to a more meaningful representation of the data.
    - Additionally, Z-score scaling helps in interpreting the importance of each principal component since the scaling ensures that the variance captured by each component is not influenced by the scale of the original features.
- Null values exist in the original dataset and are treated.
    - Missing values in CPC, CTR, and CPM are found using their respective formulas.

## Clustering the data

- Using the Hierarchical Clustering method, we construct a dendrogram using Ward linkage and Euclidean distance.

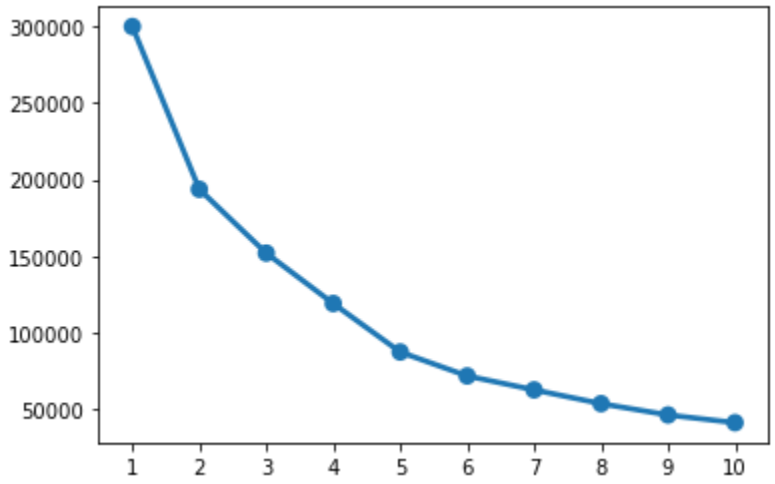



The truncated cluster:




  - The optimum number of clusters is 2.

- Applying the K-means clustering, we get:
    - An elbow curve such as this:




      From the elbow curve we can observe that there is no significant drop after the 6th point. Hence we can cluster the data in 6 groups - 0, 1, 2, 3, 4, 5.
- The Silhouette scores verify that 6 is an ideal number of clusters

```
For n_clusters=2, the silhouette score is 0.3932561565198712
For n_clusters=3, the silhouette score is 0.30716308427464184
For n_clusters=4, the silhouette score is 0.38928044688821073
For n_clusters=5, the silhouette score is 0.4456101369724424
For n_clusters=6, the silhouette score is 0.46578764296561953
For n_clusters=7, the silhouette score is 0.4612993260631106
For n_clusters=8, the silhouette score is 0.49024324848326967
For n_clusters=9, the silhouette score is 0.5098023825328091
For n_clusters=10, the silhouette score is 0.5077423641698394
```

- After profiling the clusters, we can observe the grouping as follows:

| Clus_kmeans | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Ad - Length | 154.602855 | 140.233577 | 4.263269e+02 | 4.656149e+02 | 671.426699 | 347.042553 |
| Ad- Width | 559.601850 | 573.988056 | 1.402837e+02 | 1.993366e+02 | 308.490308 | 482.322695 |
| Ad Size | 79248.743213 | 75400.132714 | 5.185801e+04 | 7.521959e+04 | 203829.647830 | 126677.659574 |
| Available_Impressions | 58570.012869 | 814312.321168 | 1.842541e+06 | 1.041504e+07 | 318841.528529 | 57833.497872 |
| Matched_Queries | 33171.650111 | 572667.611812 | 8.816976e+05 | 5.639272e+06 | 169279.338247 | 34156.909220 |
| Impressions | 25580.851599 | 483060.311214 | 8.430617e+05 | 5.460381e+06 | 145054.111111 | 27995.592553 |
| Clicks | 2177.394933 | 66006.355674 | 3.305285e+03 | 1.127562e+04 | 16488.094458 | 3553.767021 |
| Spend | 252.944812 | 7060.567034 | 1.530991e+03 | 8.663623e+03 | 1436.591518 | 350.066195 |
| Fee | 0.349984 | 0.287445 | 3.491786e-01 | 2.903267e-01 | 0.349339 | 0.349887 |
| Revenue | 164.464013 | 5071.300161 | 9.978016e+02 | 6.386974e+03 | 936.021468 | 227.936613 |
| CTR | 0.144604 | 2.132338 | 6.264290e-02 | 3.458832e-02 | 0.522129 | 19.131128 |
| CPM | 13.504273 | 15.364200 | 1.709338e+00 | 1.571005e+00 | 11.005087 | 16.124231 |

  - Cluster 4 has ads that have a higher mean length than other clusters.
  - Clusters 0 and 1 have ads that have a higher mean width than other clusters.
  - Cluster 2 has a minimum ad size.
  - Available impressions are highest for cluster 2.
  - There is not much difference in the fee charged.
  - Cluster 3  has very high mean spend and mean revenue compared to the others.

## Actionable Insights & Recommendations

- Selling ads according to CPM puts a ceiling on revenue.
- If you want to increase your revenue, you have to spend money on increasing your reach to create more ad opportunities or pumping out more ads to the same users before seeing a return.
- But if you sell on CTR, revenue is not capped.
- You can increase engagement on the same number of impressions per person, or DAU (daily active user).
- Whereas with CPM, you stretch to reach more and more people or degrade your user experience with more ads per user.

# Part 2

# Problem:

Analyse the Indian census data and perform exploratory data analysis for key variables and principle component analysis to explain the variance.

# About the data set:

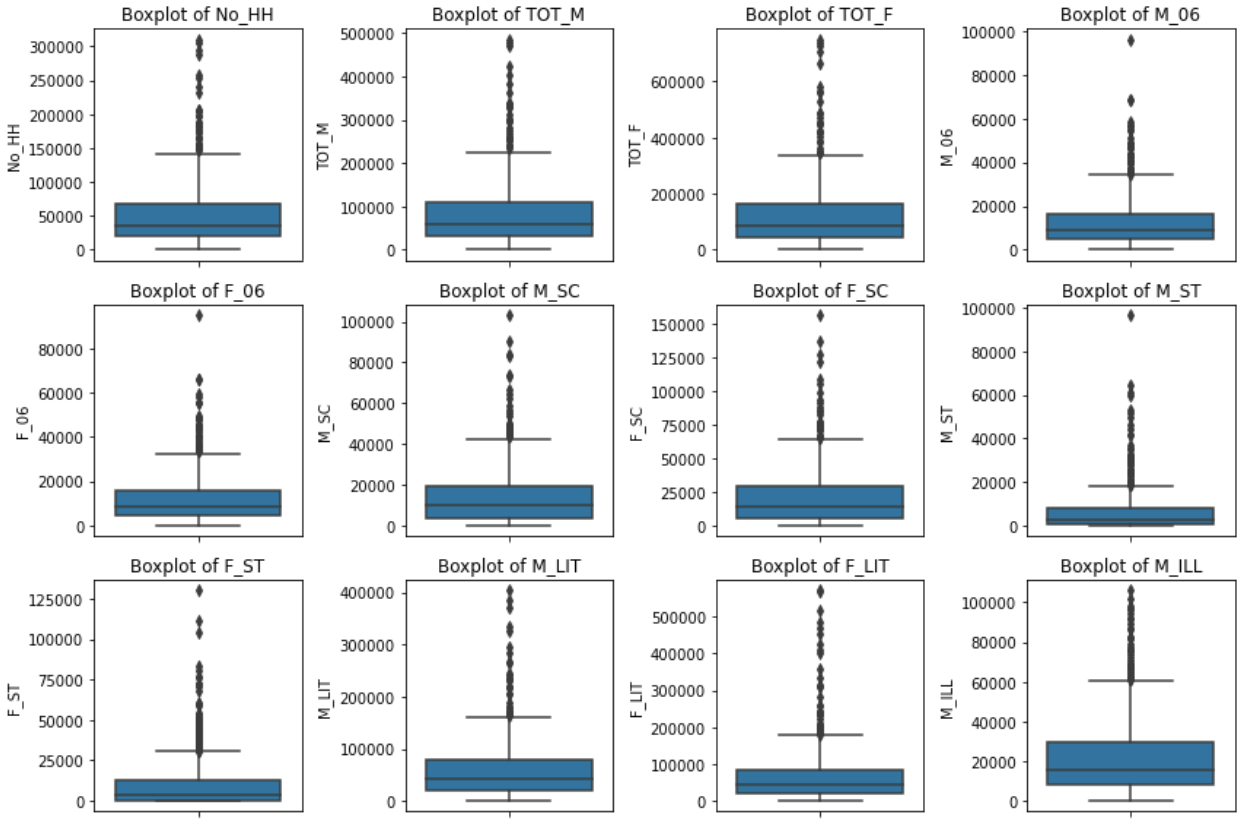
| | State Code | Dist.Code | State | Area Name | No_HH | TOT_M | TOT_F | M_06 | F_06 | M_SC | ... | MARG_CL_0_3_M | MARG_CL_0_3_F | MARG_AL_0_3_M | MARG_AL_0_ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 1 | Jammu & Kashmir | Kupwara | 7707 | 23388 | 29796 | 5862 | 6196 | 3 | ... | 1150 | 749 | 180 | |
| 1 | 1 | 2 | Jammu & Kashmir | Badgam | 6218 | 19585 | 23102 | 4482 | 3733 | 7 | ... | 525 | 715 | 123 | |
| 2 | 1 | 3 | Jammu & Kashmir | Leh(Ladakh) | 4452 | 6546 | 10964 | 1082 | 1018 | 3 | ... | 114 | 188 | 44 | |
| 3 | 1 | 4 | Jammu & Kashmir | Kargil | 1320 | 2784 | 4206 | 563 | 677 | 0 | ... | 194 | 247 | 61 | |
| 4 | 1 | 5 | Jammu & Kashmir | Punch | 11654 | 20591 | 29981 | 5157 | 4587 | 20 | ... | 874 | 1928 | 465 | |

5 rows × 61 columns

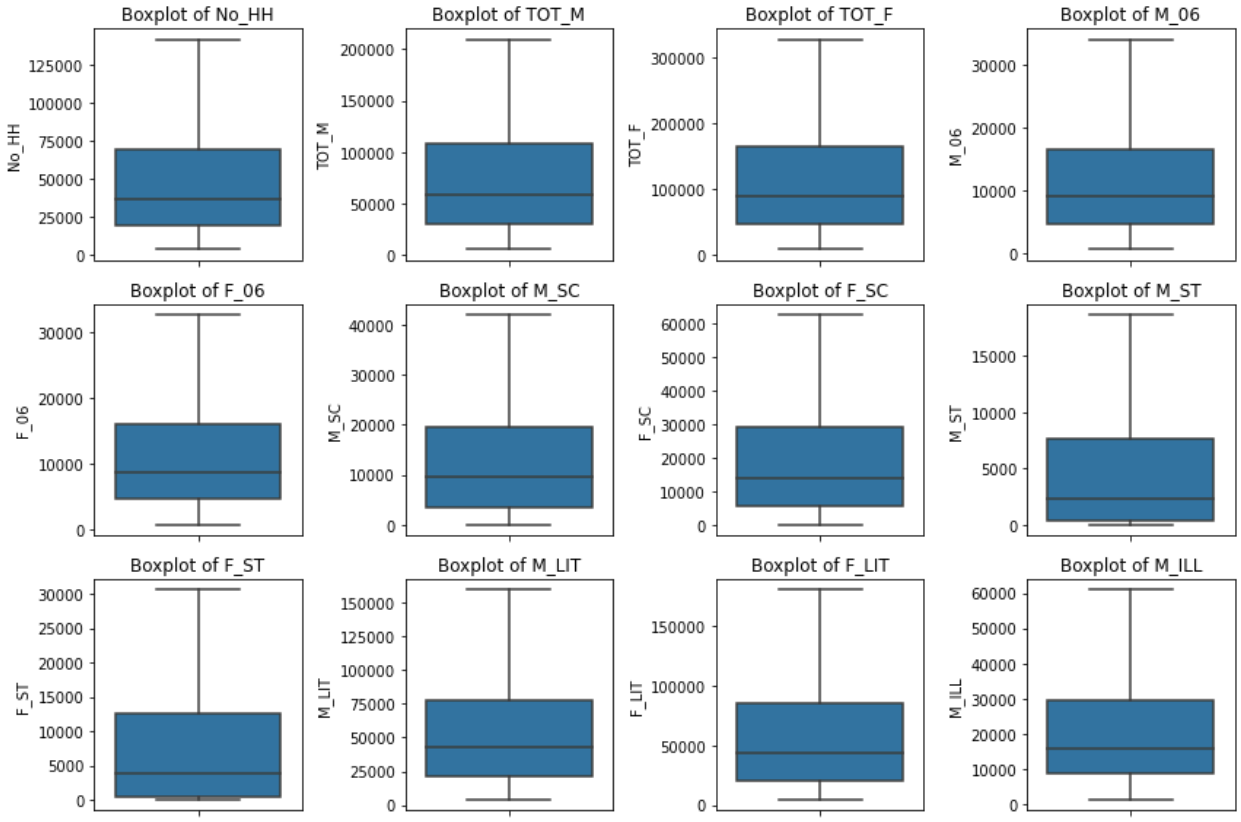
- Total number of rows = 640
- Total number of columns = 61
- Columns State Code and Dist.Code are integer values but categorical. State and Area Name are categorical variables and all other variables are numerical in nature.

# Data preprocessing

- Outliers:

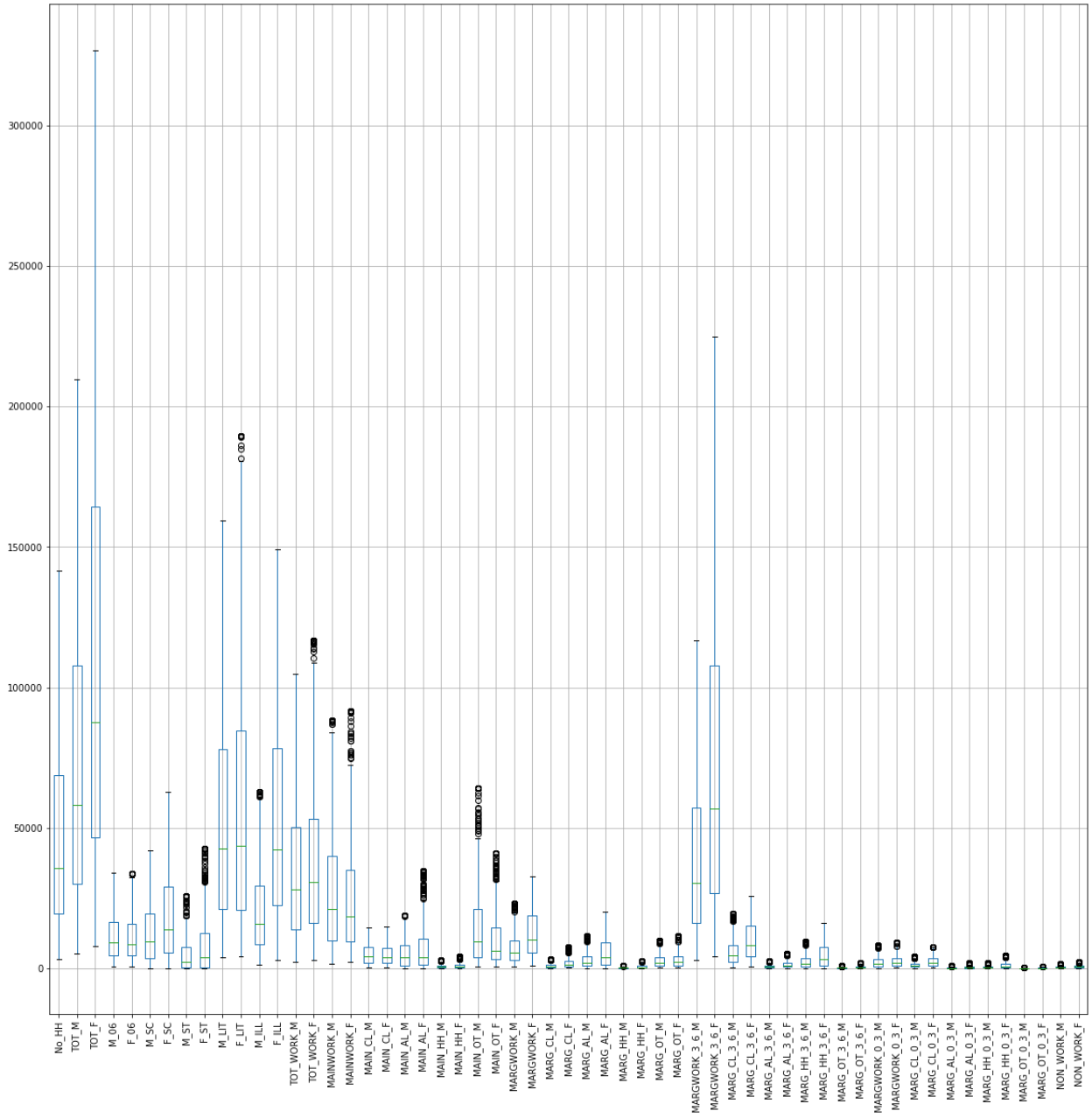



- Outliers are treated using the quantile method.

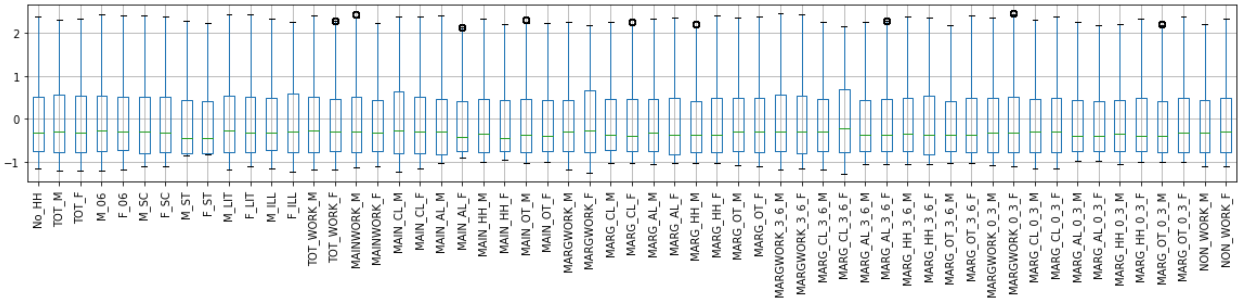



- Variables are scaled using the z-score method.

- Before scaling



- After scaling




# PCA

- Bartletts Test of Sphericity is done to tests the hypothesis that the variables are uncorrelated in the population.
    - H0: All variables in the data are uncorrelated
    - Ha: At least one pair of variables in the data are correlated
    - We get p-value = 0, which implies 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.
- The Kaiser-Meyer-Olkin (KMO) - measure of sampling adequacy (MSA) is an index used to examine how appropriate PCA is.
    - We get MSA = 0.9324595054325264
    - If MSA is less than 0.5, PCA is not recommended, since no reduction is expected.
    - Since MSA > 0.7, it is expected that a considerable reduction in the dimension and extraction of meaningful components is possible.
- Eigen vectors;

```
Eigen Vectors
 %s [[ 0.15  0.16  0.16 ...  0.14  0.15  0.14]
 [-0.12 -0.08 -0.09 ...  0.04 -0.05 -0.05]
 [ 0.1  -0.03  0.03 ... -0.1  -0.14 -0.04]
 ...
 [-0.2   0.03 -0.04 ... -0.14 -0.02  0.09]
 [-0.27  0.19 -0.04 ...  0.18 -0.1   0.21]
 [-0.2   0.26  0.05 ... -0.08 -0.19  0.01]]
```

- Explained variance ratio

```
[0.62 0.13 0.07 0.05 0.03 0.02 0.02 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.  ]
```

- Cumulative variance explained in percentages

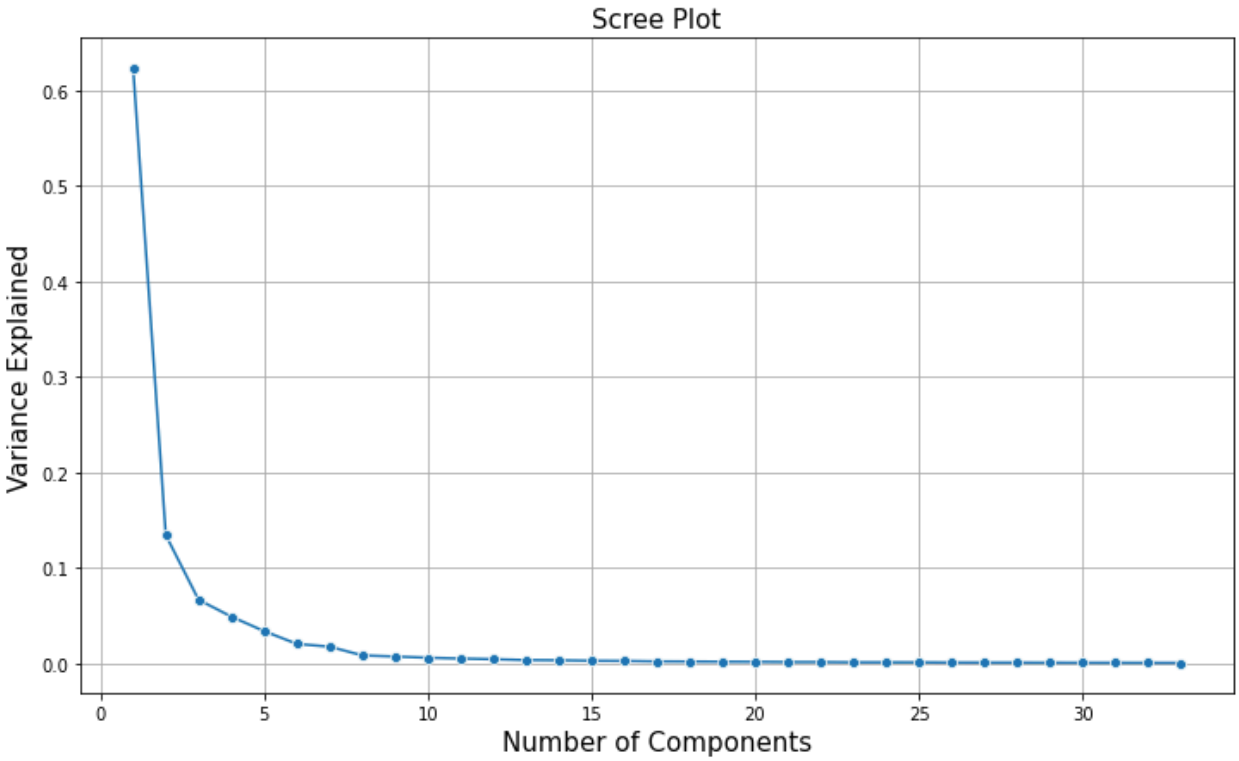
```
Cumulative Variance Explained in Percentage: [62.4  75.8  82.44 87.31 90.66 92.68 94.4  95.21 95.9  96.47 96.95 97.36
 97.68 97.97 98.22 98.45 98.63 98.8  98.95 99.09 99.2  99.32 99.4  99.48
 99.56 99.61 99.66 99.71 99.75 99.79 99.82 99.85 99.87]
```

    - The 1st Principal component is positively correlated with Number of Household, Total Male & Female population, Literacy & Illiteracy Numbers among M & F, Number of SC in Males & Females, Working population, etc. These variables explain the most variance in the data i.e. 56%
    - The 2nd Principal component is correlated with the Marginal Cultivator Male/Female population and Marginal Agriculture (Male & Female) population etc. The Second PC explains about 14% of the variation in the data.
    - The 3rd Principal Component explains about 7% variation in the data. It positively correlates with Marginal Agriculture 0-3 Female, and 3-6 M&F Population.
    - The 4th Principal Component correlated positively with Marginal Households Male, Marginal Other (0- 3,3-6) Workers Male population. It explains about 6% of the variation in the data.
    - The 5th Principal Component explains about 4% variation in data. It is positively correlated with Scheduled Tribes Population Male& Female, Non-working Male& Female population.
    - The 6th Principal Component explains about 3% variation in data. It is positively correlated with Female Marginal Other workers (0-3,3-6), Main & Marginal Households Female population.

- We can see that about 82% of the variance is explained by 3 Principal Components.
- We can see that about 90% of the variance is explained by 5 Principal Components.
- We can see that about 92% of the variance is explained by 6 Principal Components.

Each PC correlates with a different set of variables explaining how different aspects of the population contribute to the variation in data

- Scree plot to identify the number of components to be built




Here, it is decided to keep the number of components as 6 as the cumulative explained variance is around 92%.

Heatmap showing features have maximum loading across the components.