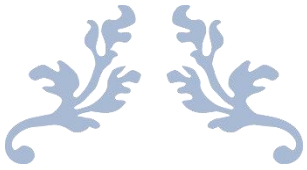


MACHINE LEARNING-1-REPORT-CODED

FASNA-PP

TABLE OF CONTENTS: Problem 1:

TABLE OF FIGURES

Tables:

Problem 1:

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.

DATA DICTIONARY

1-**Timestamp**-The Timestamp of the particular Advertisement.

2-**InventoryType**-The Inventory Type of the particular Advertisement. Format 1 to 7. This is a Categorical Variable.

3-**Ad – Length-**The Length Dimension of the particular Adverstisement.

4-**Ad- Width**-The Width Dimension of the particular Advertisement.

5-**Ad Size**-The Overall Size of the particular Advertisement. Length*Width.

6-**Ad Type**-The type of the particular Advertisement. This is a Categorical Variable.

7-**Platform-**The platform in which the particular Advertisement is displayed. Web, Video or App. This is a Categorical Variable.

8-**Device Type**-The type of the device which supports the partciular Advertisement. This is a Categorical Variable.

9-**Format**-The Format in which the Advertisement is displayed. This is a Categorical Variable.

10-**Available_Impressions**-How often the particular Advertisement is shown. An impression is counted each time an Advertisement is shown on a search result page or other site on a Network.

11**-Matched_Queries**-Matched search queries data is pulled from Advertising Platform and consists of the exact searches typed into the search Engine that generated clicks for the particular Advertisement.

12-**Impressions-**The impression count of the particular Advertisement out of the total available impressions.

13**-Clicks**-It is a marketing metric that counts the number of times users have clicked on the particular advertisement to reach an online property.

14-**Spend-**It is the amount of money spent on specific ad variations within a specific campaign or ad set. This metric helps regulate ad performance.

15**-Fee**-The percentage of the Advertising Fees payable by Franchise Entities.

16**-Revenue**-It is the income that has been earned from the particular advertisement.

17-**CTR-**CTR stands for "Click through rate". CTR is the number of clicks that your ad receives divided by the number of times your ad is shown. Formula used here is CTR = Total Measured Clicks / Total Measured Ad Impressions x 100. Note that the Total Measured Clicks refers to the 'Clicks' Column and the Total Measured Ad Impressions refers to the 'Impressions' Column.

18**-CPM-**CPM stands for "cost per 1000 impressions." Formula used here is CPM = (Total Campaign Spend / Number of Impressions) * 1,000. Note that the Total Campaign Spend refers to the 'Spend' Column and the Number of Impressions refers to the 'Impressions' Column.

19**-CPC**-CPC stands for "Cost-per-click". Cost-per-click (CPC) bidding means that you pay for each click on your ads. The Formula used here is CPC = Total Cost (spend) / Number of Clicks. Note that the Total Cost (spend) refers to the 'Spend' Column and the Number of Clicks refers to the 'Clicks' Column.

Part 1:--1)Clustering: Define the problem and perform Exploratory Data Analysis¶

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

ANSWRS:Import all the necessary libraries

read the dataset, print first 5 rows

	Timestamp	Inventory Type	Ad-Length	Ad-Width	Ad Size	Ad Type	Platform	Device Type	Format	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
0	2020-9-2-17	Format1	300	250	750000	Inter222	Video	Desktop	Display	1806	325	323	1	0.0	0.35	0.0	0.00031	0.0	0.0
1	2020-9-2-10	Format1	300	250	750000	Inter227	App	Mobile	Video	1780	285	285	1	0.0	0.35	0.0	0.00035	0.0	0.0
2	2020-9-1-22	Format1	300	250	750000	Inter222	Video	Desktop	Display	2727	356	355	1	0.0	0.35	0.0	0.00028	0.0	0.0
3	2020-9-3-20	Format1	300	250	750000	Inter228	Video	Mobile	Video	2430	497	495	1	0.0	0.35	0.0	0.00020	0.0	0.0
4	2020-9-4-15	Format1	300	250	750000	Inter217	Web	Desktop	Video	1218	242	242	1	0.0	0.35	0.0	0.00041	0.0	0.0

Table -1-first 5 rows

1---1)-a)- Check shape

#shape of the data set: `(23066, 19)`

1---1)-b)-Data types

```
<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
```

1---1)-c)-statistical summary

	count	mean	std	min	25%	50%	75%	max
Ad - Length	23066.0	385.16	233.65	120.00	120.00	300.00	720.00	728.00
Ad- Width	23066.0	337.90	203.09	70.00	250.00	300.00	600.00	600.00
Ad Size	23066.0	96674.47	61538.33	33600.00	72000.00	72000.00	84000.00	216000.00
Available_Impressions	23066.0	2432043.67	4742887.76	1.00	33672.25	483771.00	2527711.75	27592861.00
Matched_Queries	23066.0	1295099.14	2512969.86	1.00	18282.50	258087.50	1180700.00	14702025.00
Impressions	23066.	124151	2429399.9	1.00	7990.50	225290.0	1112428.50	14194774.00

	0	9.52	6			0		
Clicks	23066.0	10678.52	17353.41	1.00	710.00	4425.00	12793.75	143049.00
Spend	23066.0	2706.63	4067.93	0.00	85.18	1425.12	3121.40	26931.87
Fee	23066.0	0.34	0.03	0.21	0.33	0.35	0.35	0.35
Revenue	23066.0	1924.25	3105.24	0.00	55.37	926.34	2091.34	21276.18
CTR	18330.0	0.07	0.08	0.00	0.00	0.08	0.13	1.00
CPM	18330.0	7.67	6.48	0.00	1.71	7.66	12.51	81.56
CPC	18330.0	0.35	0.34	0.00	0.09	0.16	0.57	7.26

Table-2-describe the data set

#check the null values
null values in CTR,CPM,CPC
#check the total duplicated values in the data set
no duplicated values

#create data frame with categorical variables

#create data frame with numerical variables

#check the counts of each columns for checking data irregularities

1---1-d)Univariate analysis

for numerical columns
#plot histogram 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
```

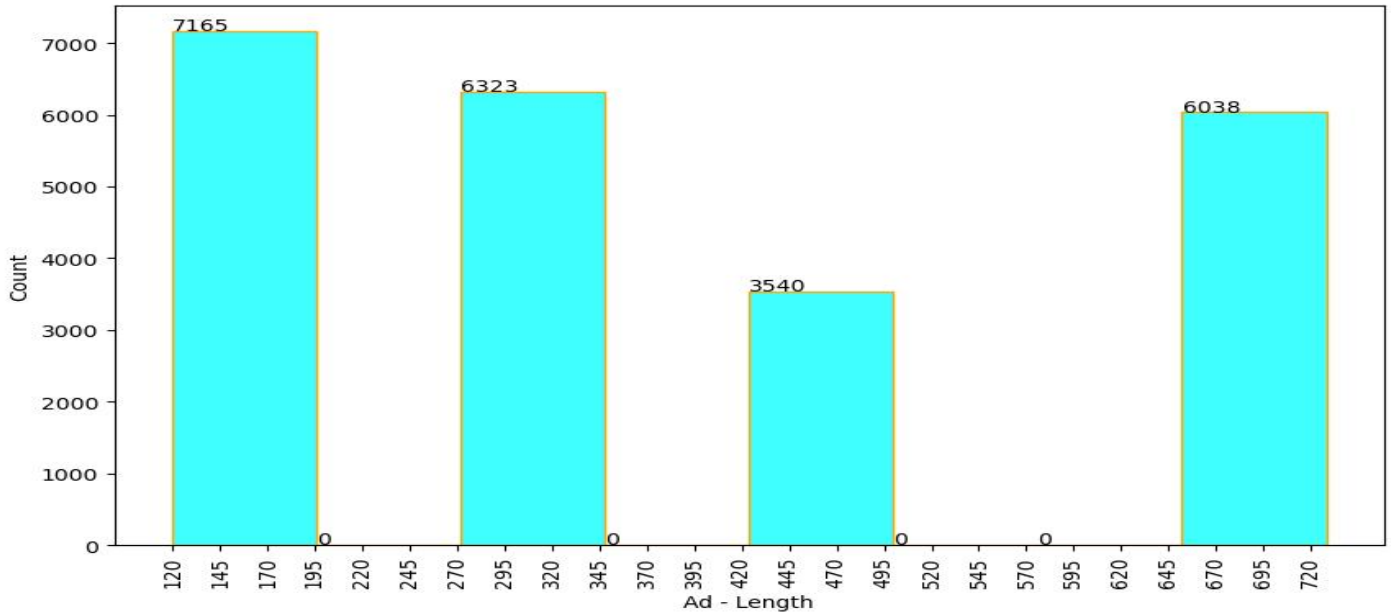


Fig-1 histogram of Ad_length

#plot histogram and boxplot of Ad-length

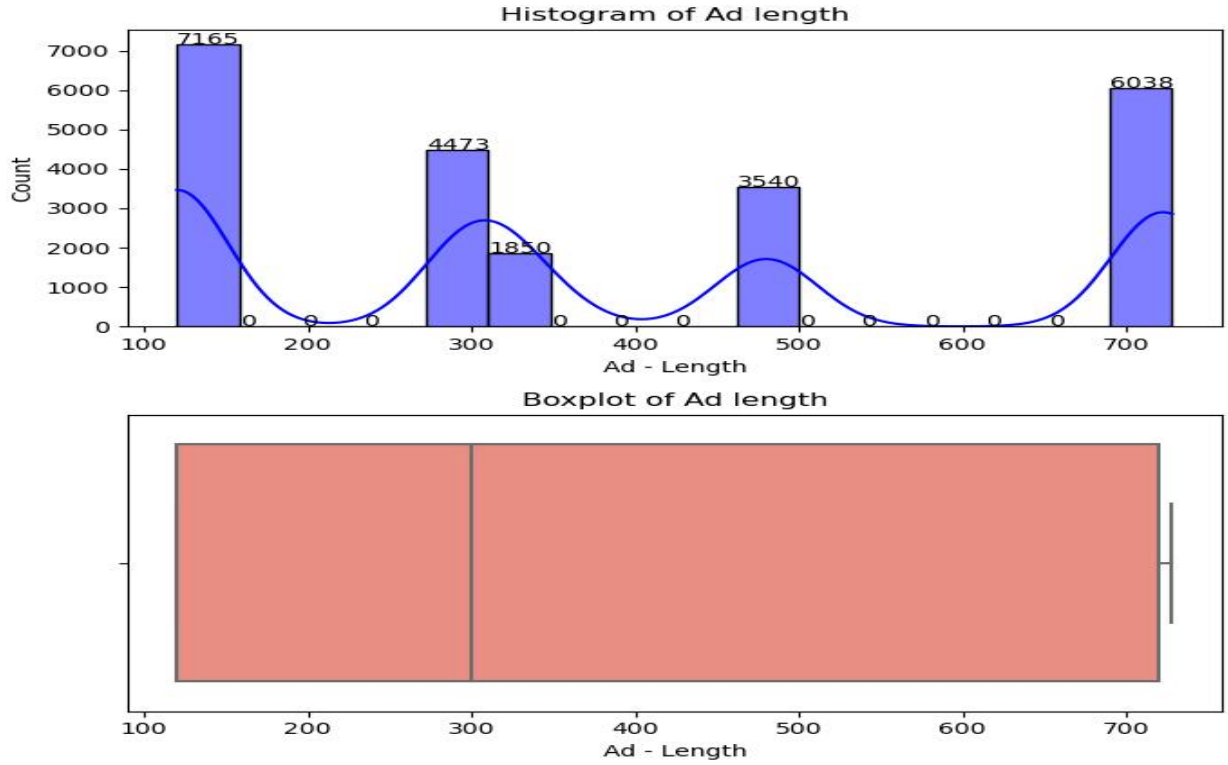


Fig-2-histogram and boxplot of Ad-length

plot histogram and boxplot 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
```

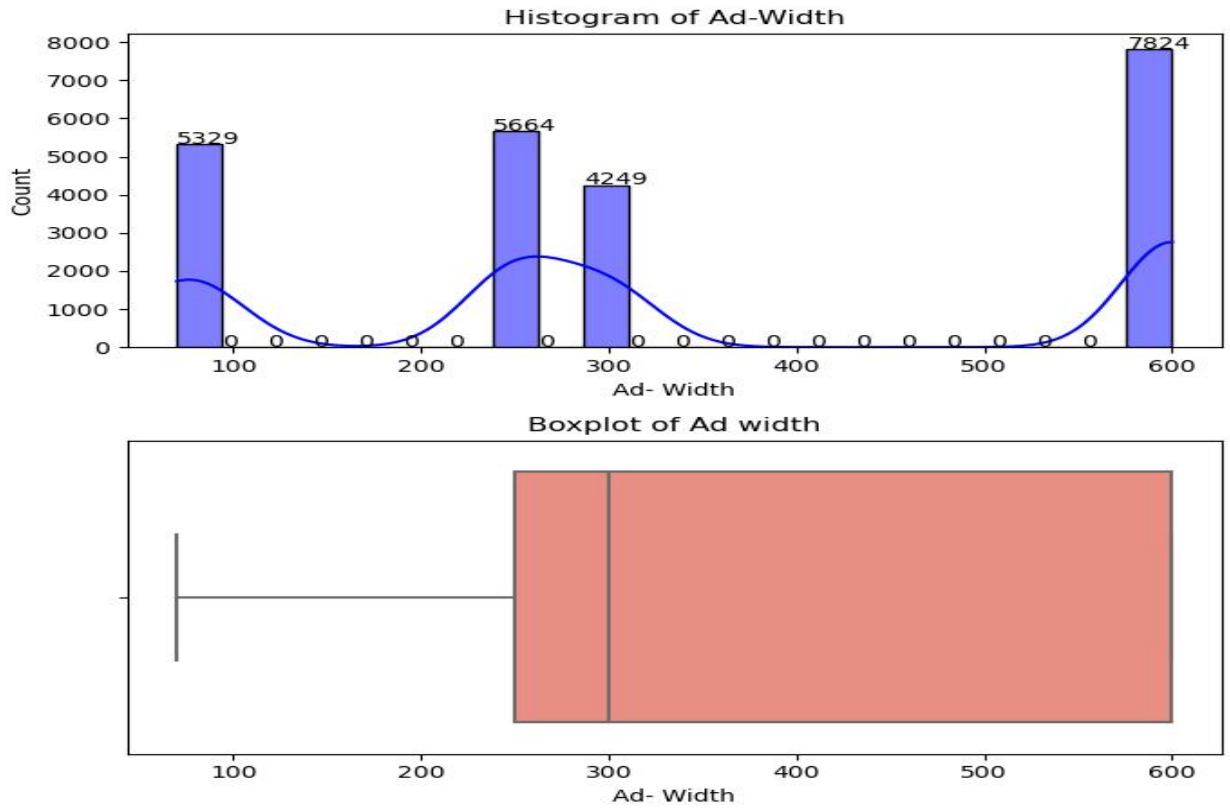


Fig-3-histogram and boxplot of Ad-width

#plot histogram and boxplot 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
```

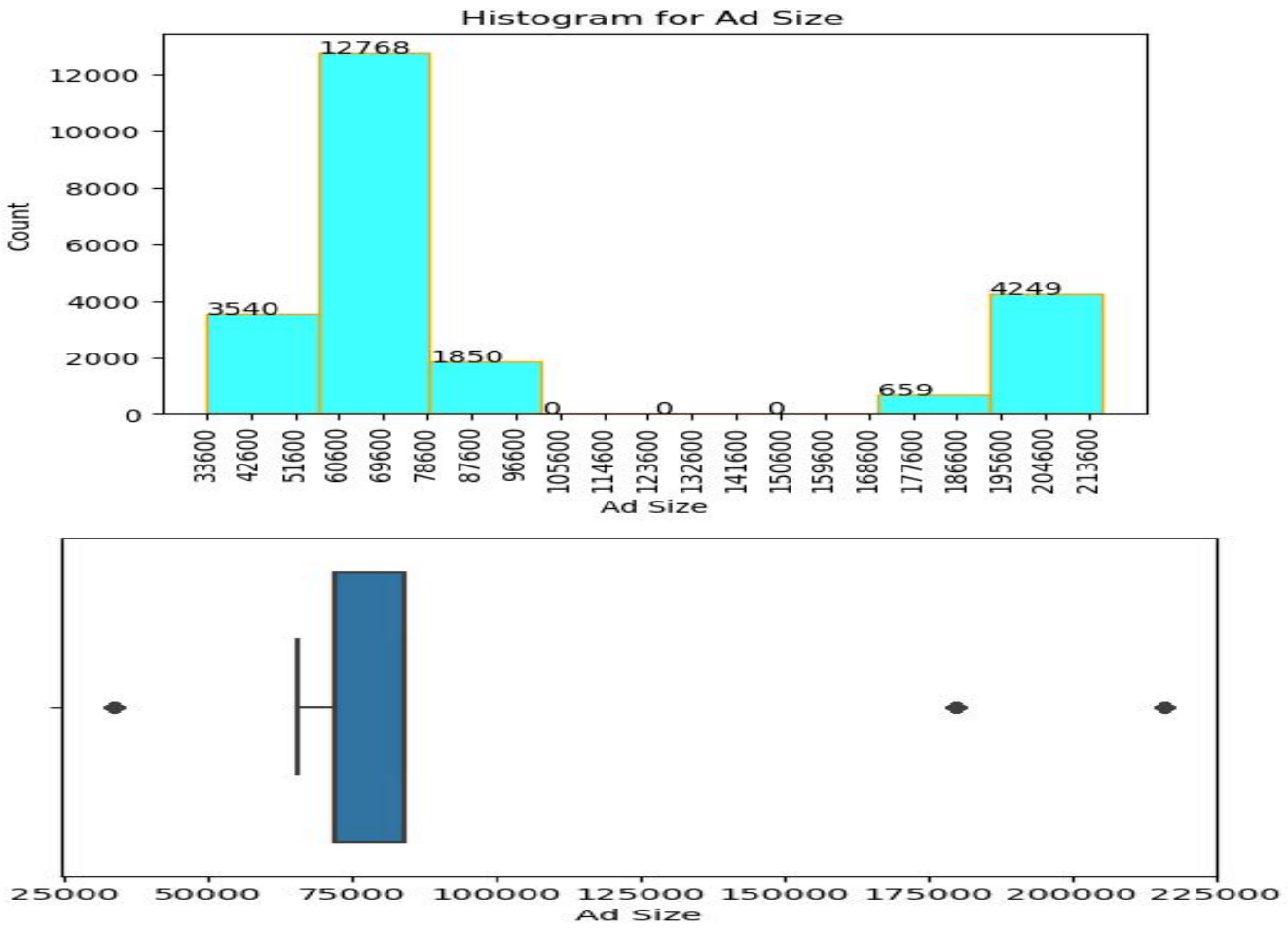


Fig-4-histogram and boxplot of Ad-size

#plot histogram and boxplot of available impression

```
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
```

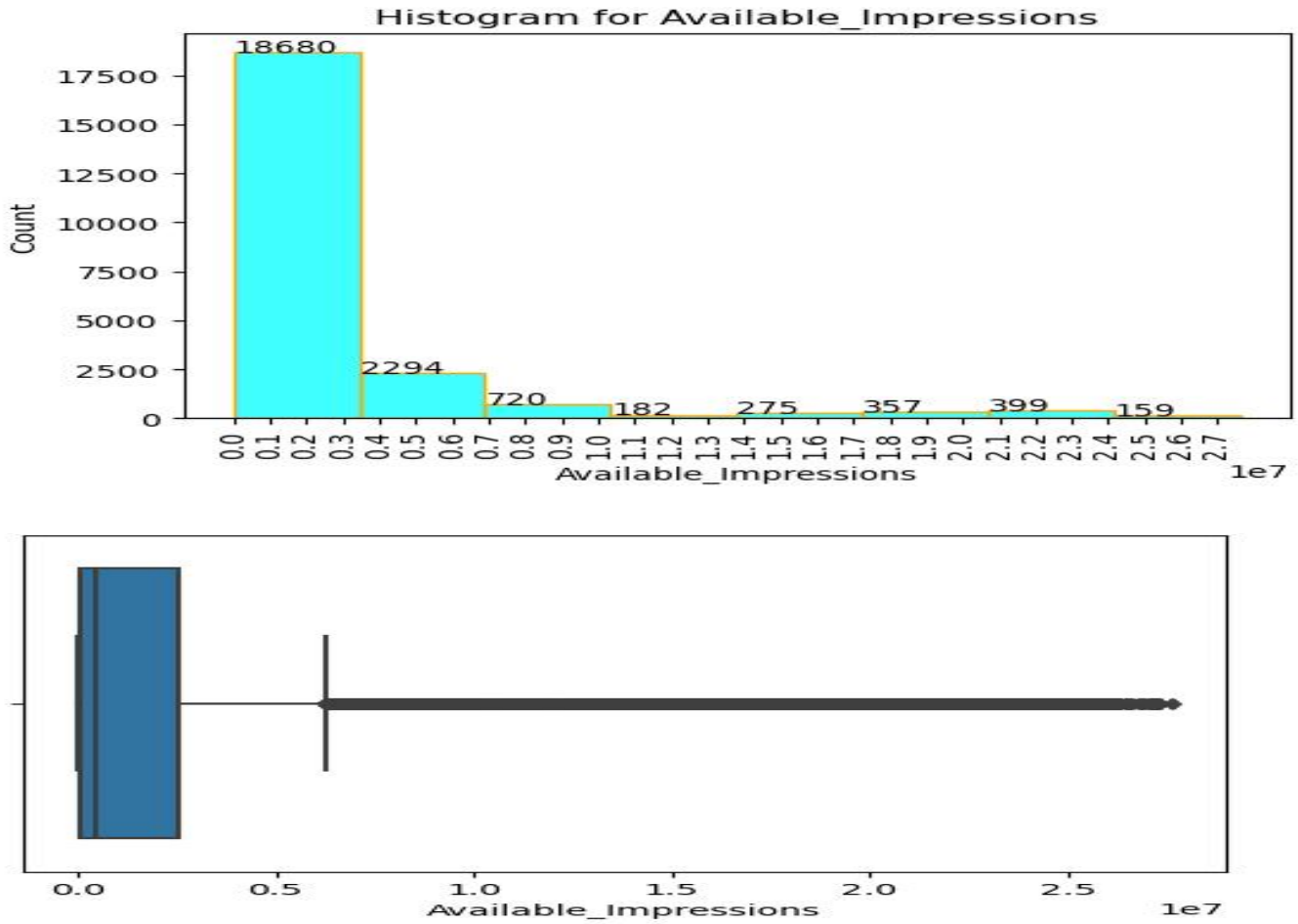


Fig-5-histogram and boxplot of available impression

#plot histogram and boxplot 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
```

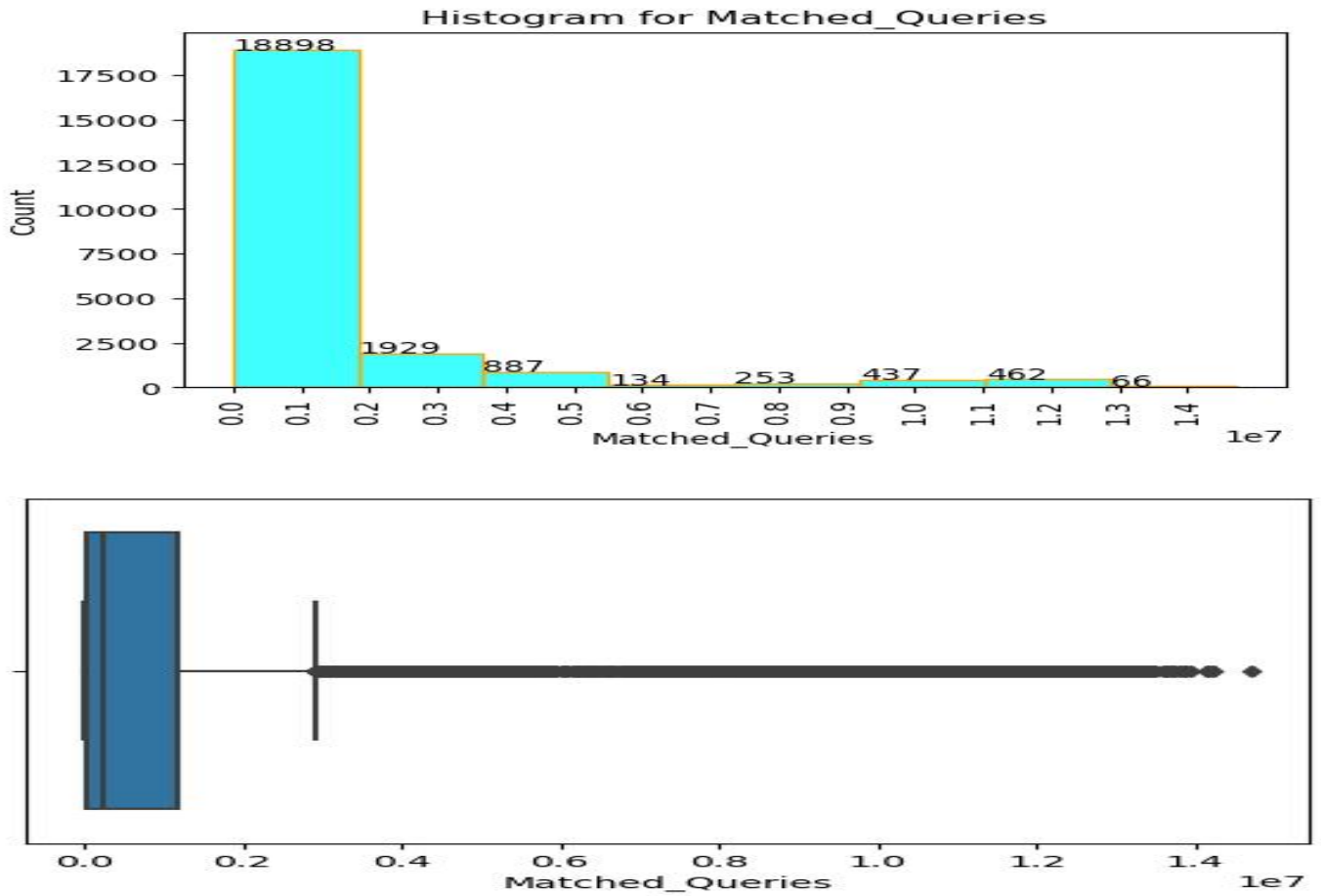


Fig-6-histogram and boxplot of matched queries

#plot histogram and boxplot of impression

```
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
```

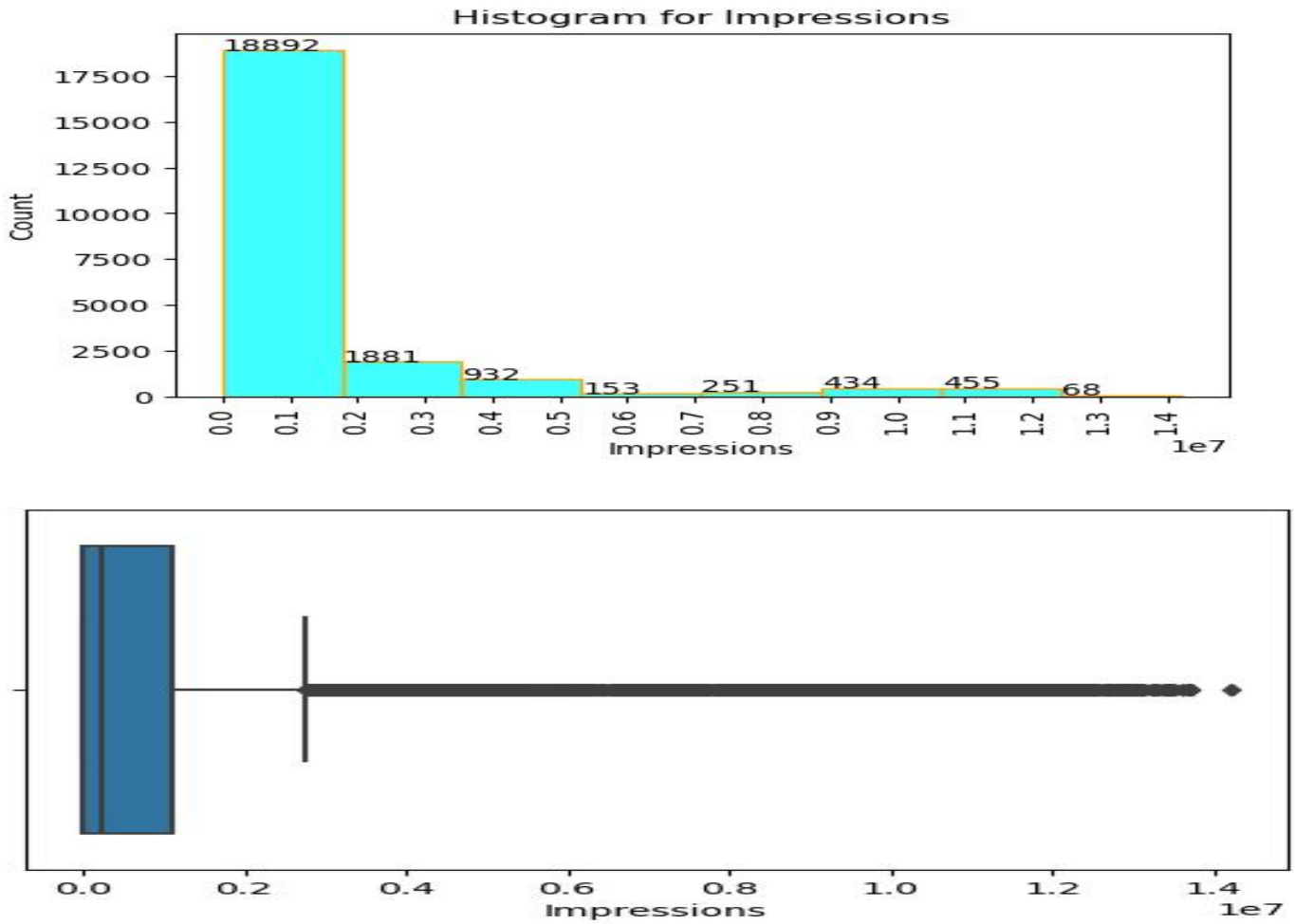
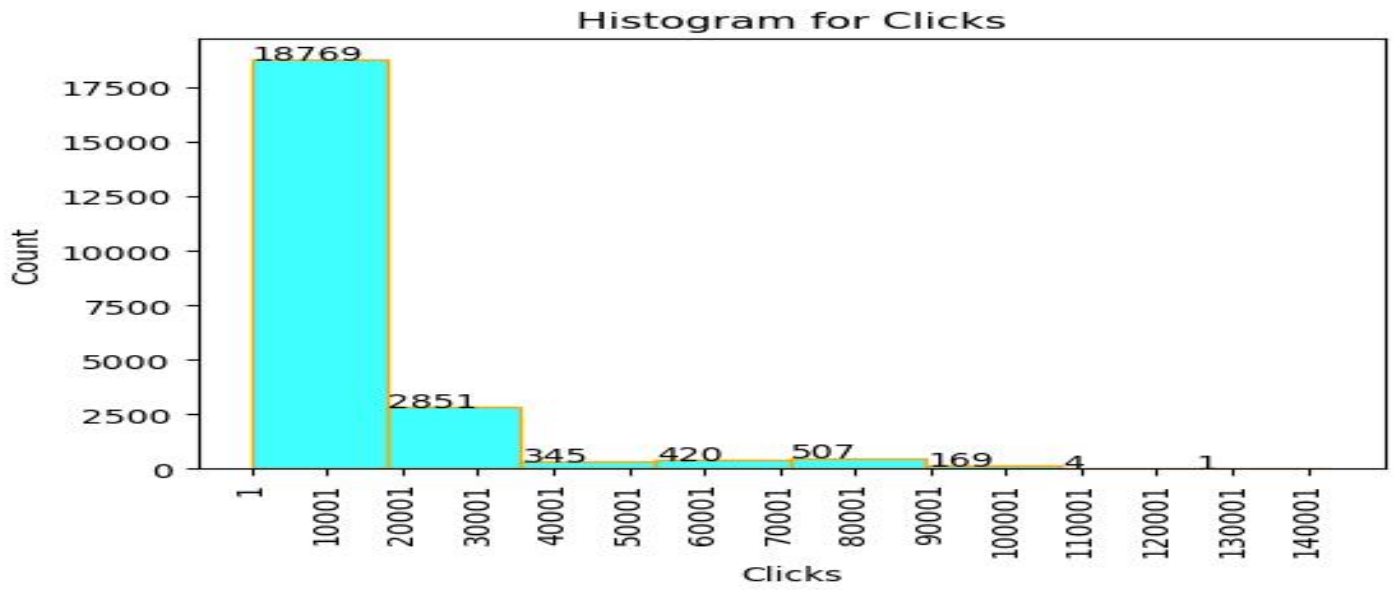


Fig-7-histogram and boxplot of impression

#plot histogram and boxplot 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
```



BoxPlot of clicks

<Figure size 640x480 with 0 Axes>

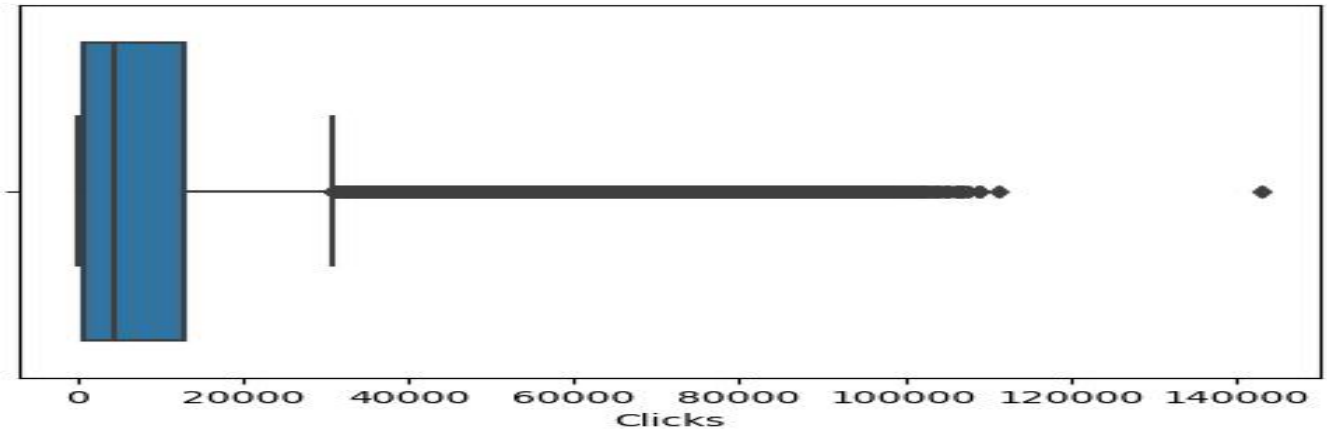
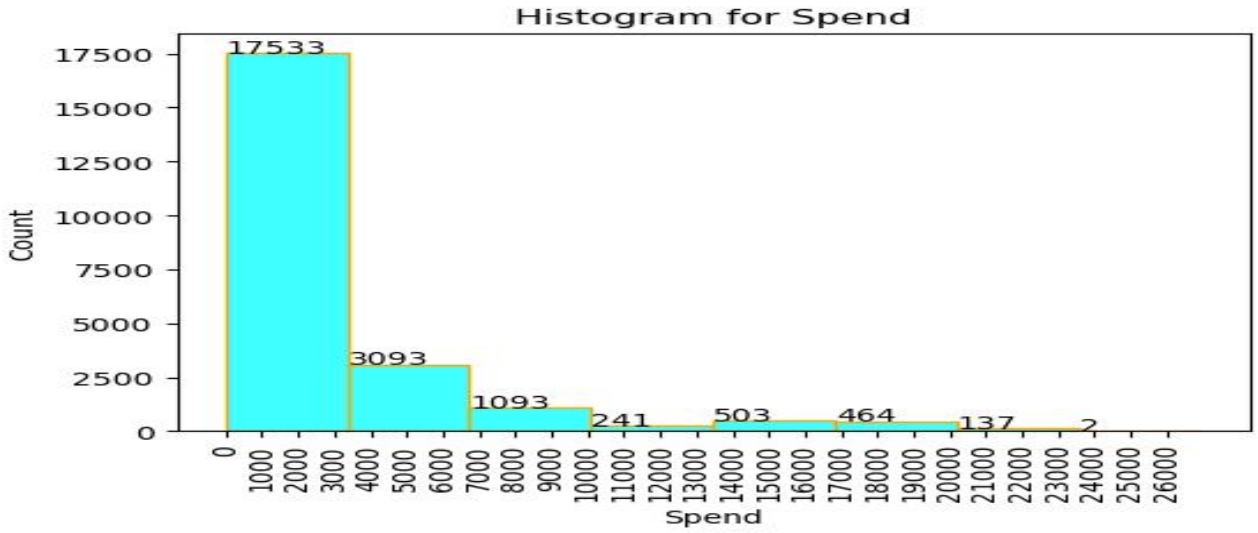


Fig-8-histogram and boxplot of clicks

#plot histogram and boxplot 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
```



BoxPlot of Spend
<Figure size 640x480 with 0 Axes>

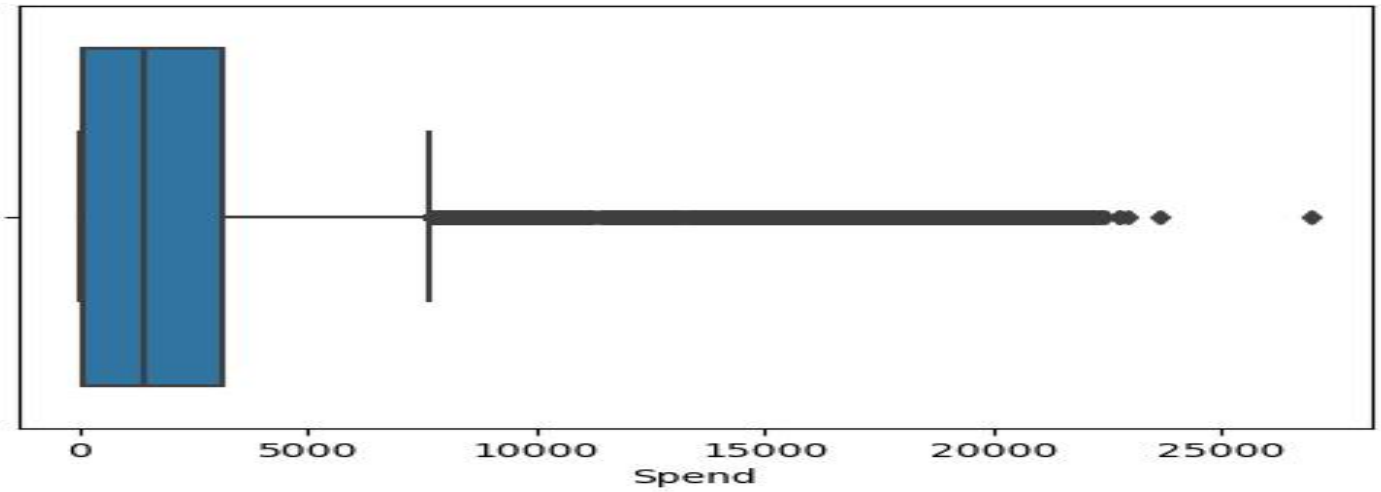


Fig-9-histogram and boxplot of spend

plot histogram and boxplot 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
```

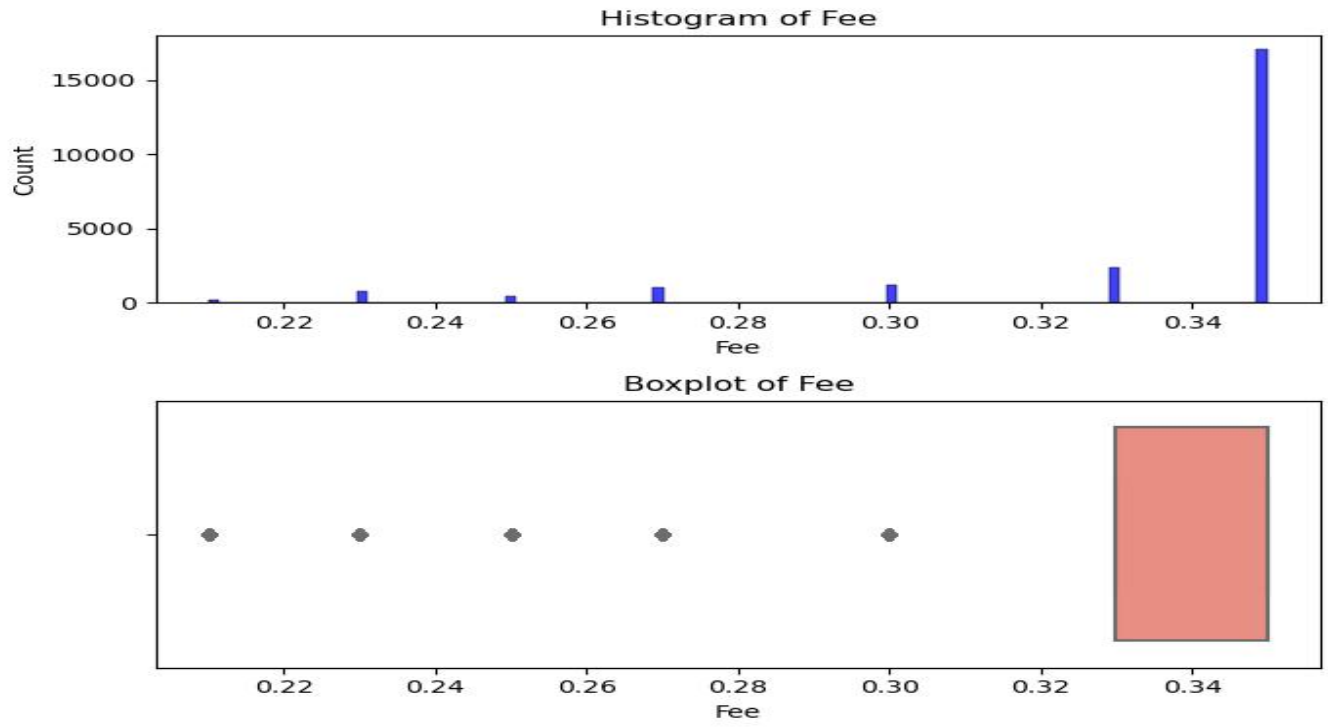
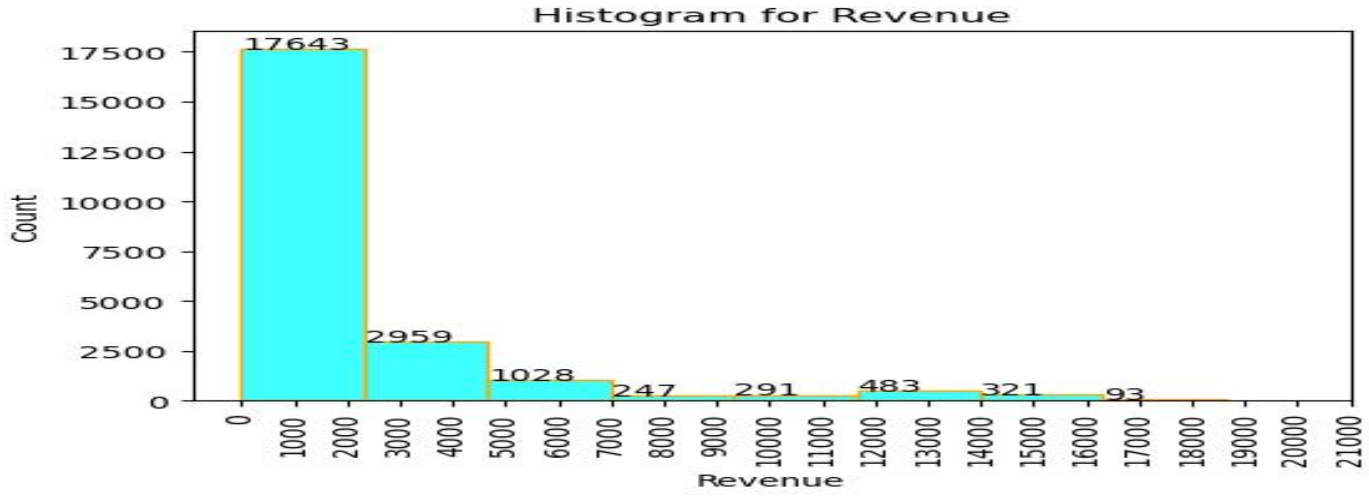


Fig-10- histogram and boxplot of Fee

#plot histogram and boxplot 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
```



```
BoxPlot of Revenue
<Figure size 640x480 with 0 Axes>
```

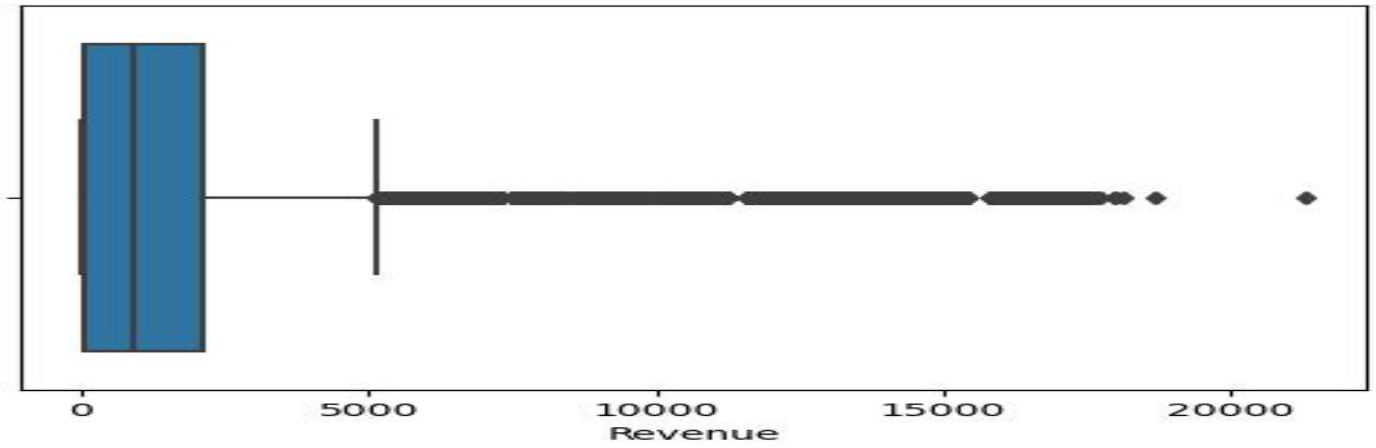


Fig-11-histogram and boxplot of revenue

plot histogram and boxplot of CTR

```
count    18330.000000
mean         0.073661
std          0.075160
min          0.000100
25%          0.002600
50%          0.082550
75%          0.130000
max          1.000000
Name: CTR, dtype: float64
```

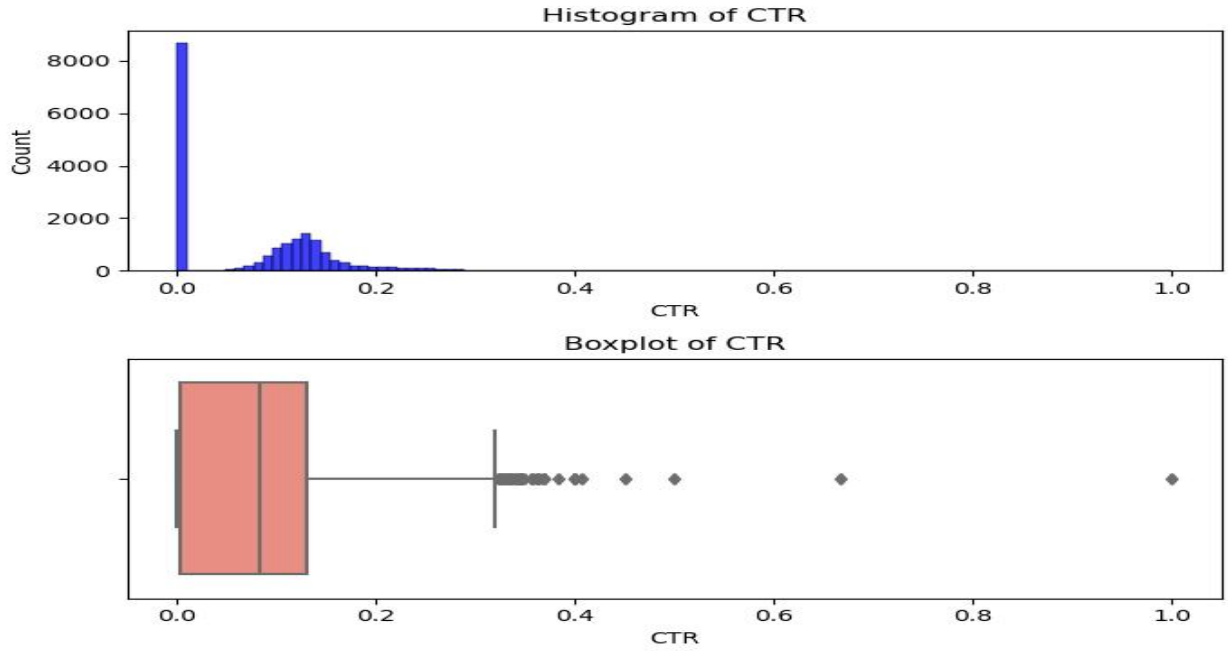


Fig-12- histogram and boxplot of CTR

#plot histogram and boxplot of CPM

```
count    18330.000000
mean         7.672045
std          6.481391
min          0.000000
25%          1.710000
50%          7.660000
75%         12.510000
max         81.560000
Name: CPM, dtype: float64
```

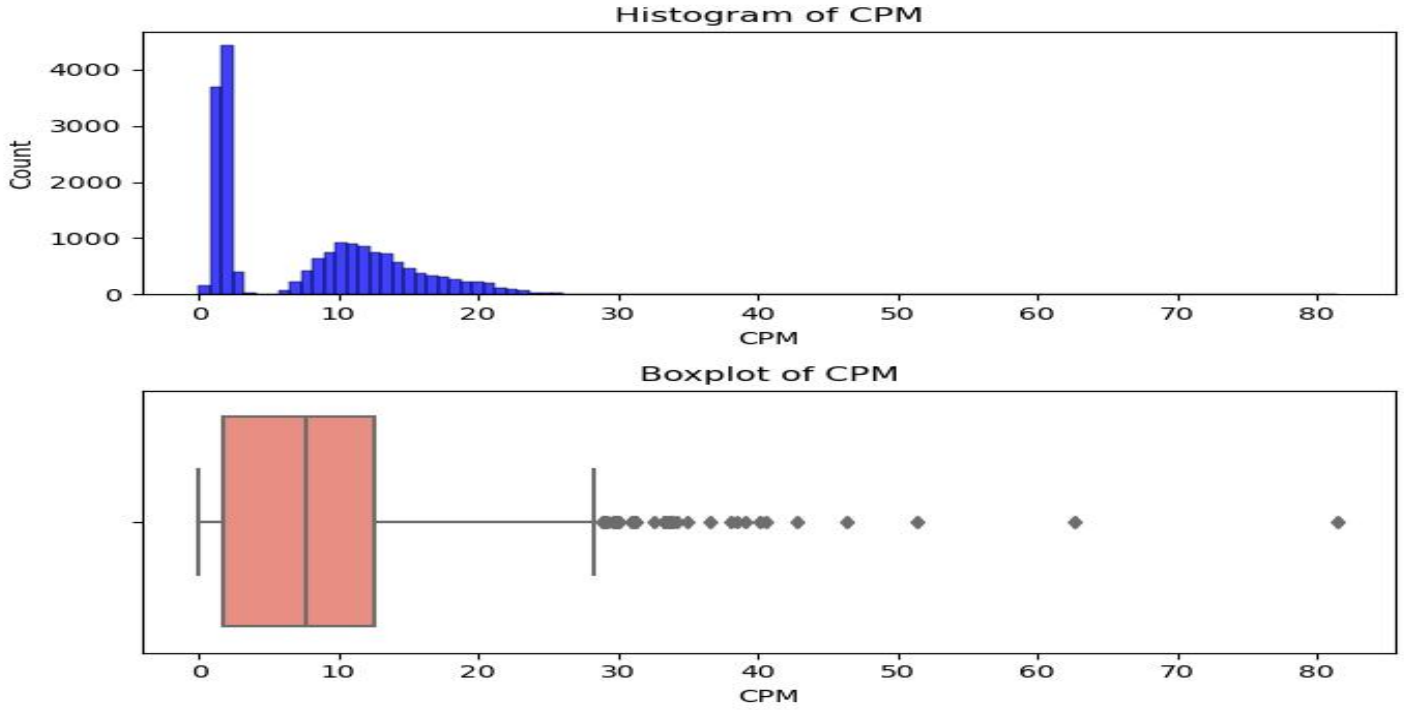


Fig-13- histogram and boxplot of CPM

#plot histogram and boxplot of CPC

```
count    18330.000000
mean         0.351061
std          0.343334
min          0.000000
25%          0.090000
50%          0.160000
75%          0.570000
max          7.260000
Name: CPC, dtype: float64
```

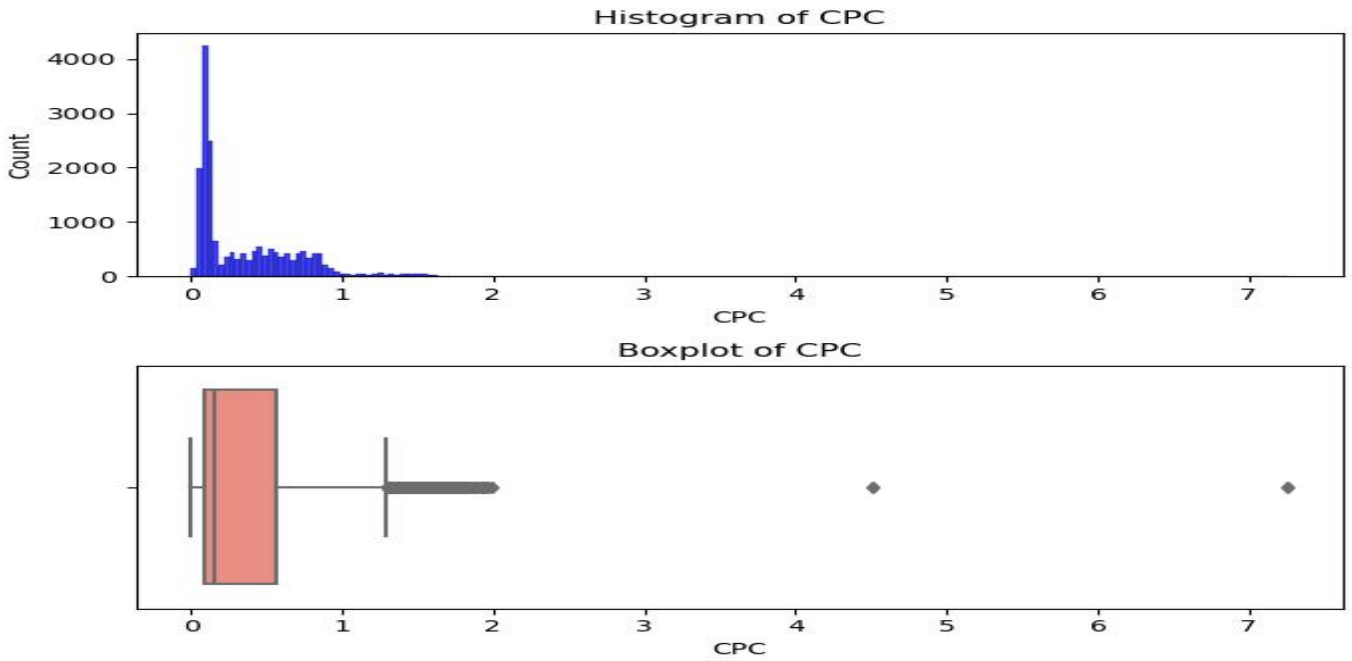


Fig-14- histogram and boxplot of CPC

#DESCRIBE THE FEE COLUMN

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

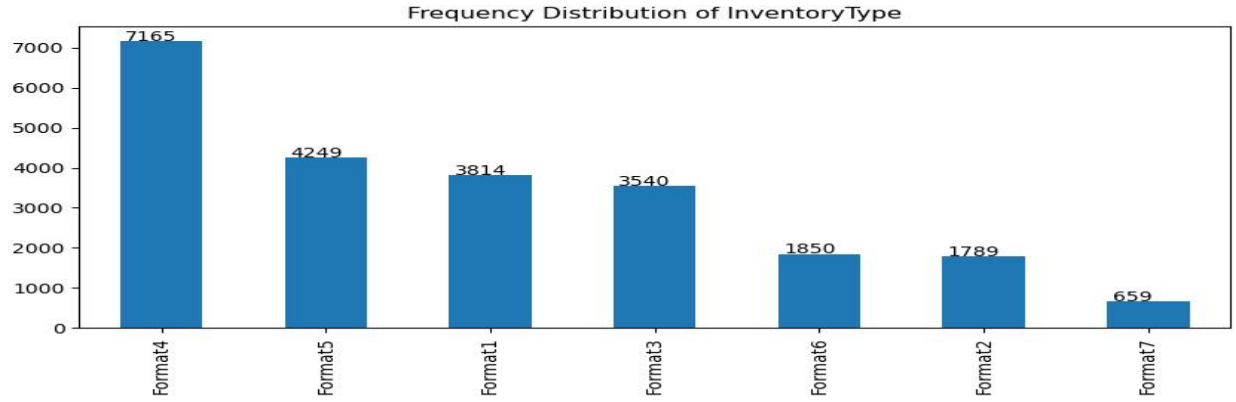
Observations

- There are 13 numeric fields in the data
- Customer ad length ranges from 120 to 728
- maximum of Ad_width is 600
- Ad size ranges from 33600 to 216000 with an average 7200
- Available_Impressions ranges from 1.000000e+00 to 2.759286e+07
- only few matched queries above 1.4
- Average impression is around 2.252900e+05
- more Ads are zero clicks very less ads are more than 80000 Clicks

- Spend range is 0.00 to 26931.870000
- 50%,75%,maximum of fee of Ads is 350000
- maximum of Revenue of Ad is 21276.180000
- CTR range in between 0.00 to 1.00
- maximum of CPM is 81.560000
- Range of CPC 0.000 to 7.260000 with average 0.160000
- Outliers to be treated

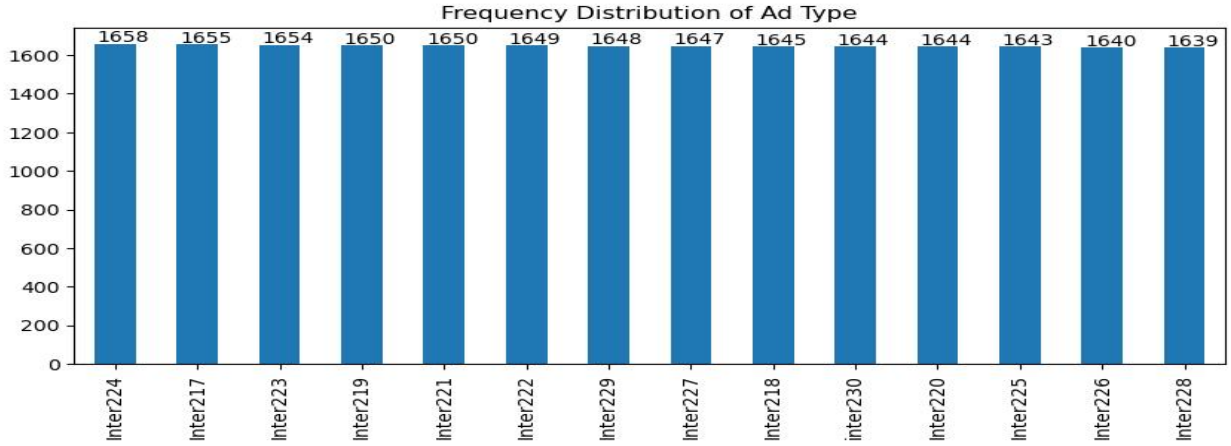
#UNIVARIATE ANALYSIS FOR CATEGORICAL COLUMNS

```
----------------------------------------------------------
Distribution of InventoryType
--------------------------------------------------------------------------------
Format4    7165
Format5    4249
Format1    3814
Format3    3540
Format6    1850
Format2    1789
Format7     659
Name: InventoryType, dtype: int64
```

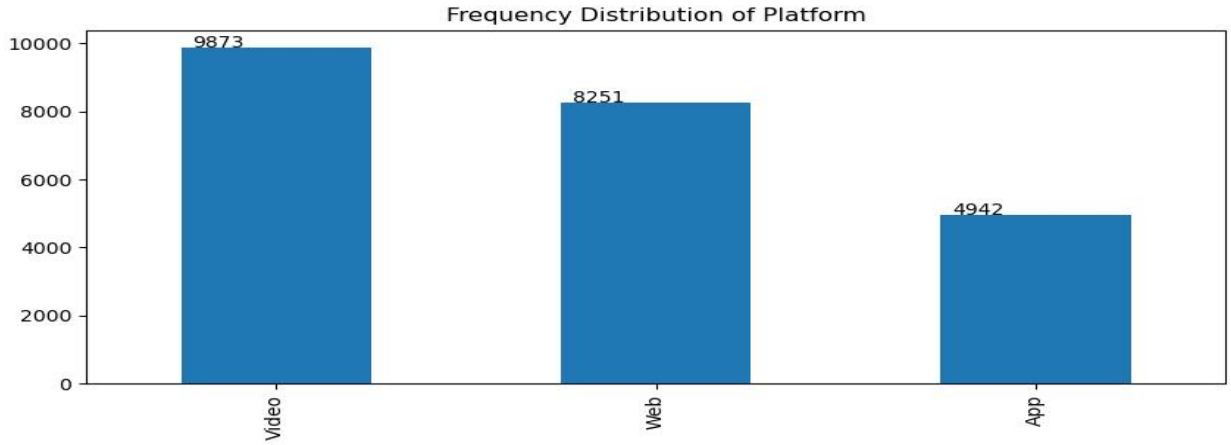



```
----------------------------------------------------------
Distribution of Ad Type
--------------------------------------------------------------------------------
Inter224    1658
Inter217    1655
Inter223    1654
Inter219    1650
Inter221    1650
```

```
Inter222    1649
Inter229    1648
Inter227    1647
Inter218    1645
inter230    1644
Inter220    1644
Inter225    1643
Inter226    1640
Inter228    1639
Name: Ad Type, dtype: int64
```

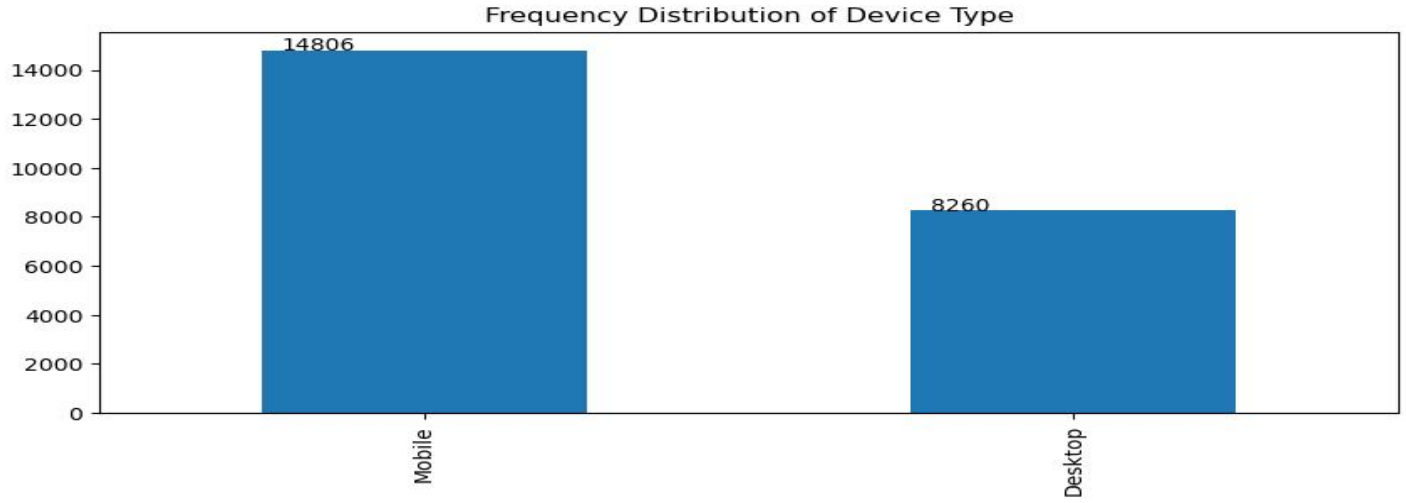



```
------------------------------------------------------------
Distribution of Platform
--------------------------------------------------------------------------------
Video     9873
Web       8251
App       4942
Name: Platform, dtype: int64
```




```
------------------------------------------------------------
Distribution of Device Type
--------------------------------------------------------------------------------
Mobile     14806
```

```
Desktop     8260
Name: Device Type, dtype: int64
```




```
----------------------------------------------------------
Distribution of Format
----------------------------------------------------------------------------
Video       11552
Display     11514
Name: Format, dtype: int64
```

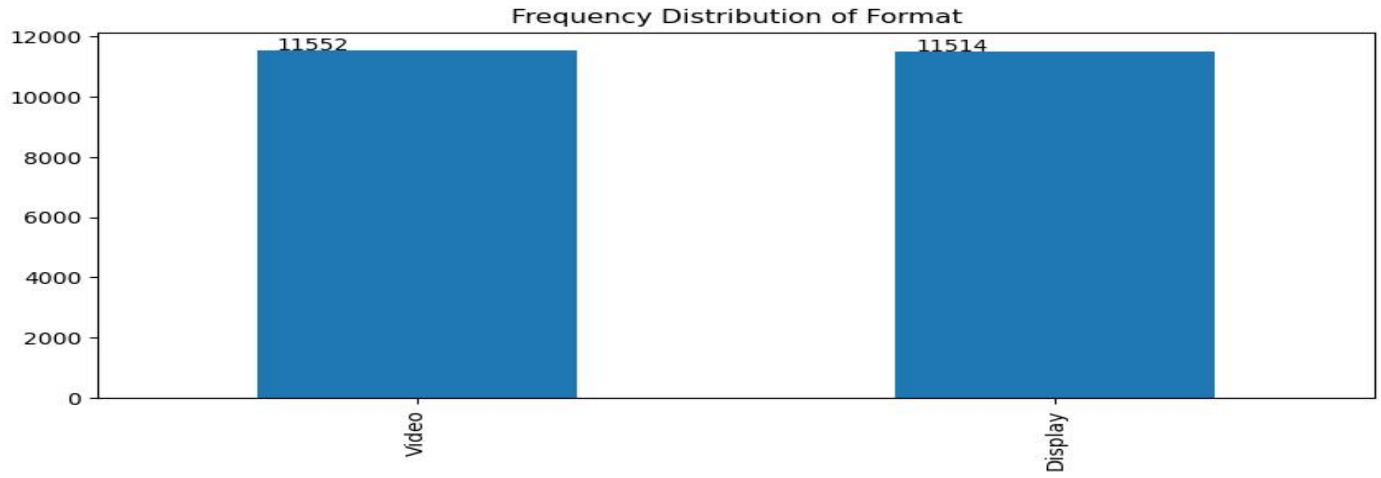


Fig-15- UNIVARIATE ANALYSIS FOR CATEGORICAL COLUMNS

- more inventory type is Format1
- counts of All Ad Type is approximately same
- most count Ad are in video platform,less counts in App platform
- mobile device have high Ad than desktop
- Format of Ads video type and display type are approximately same

1---1--e)-Bivariate analysis

#plot the correlation map

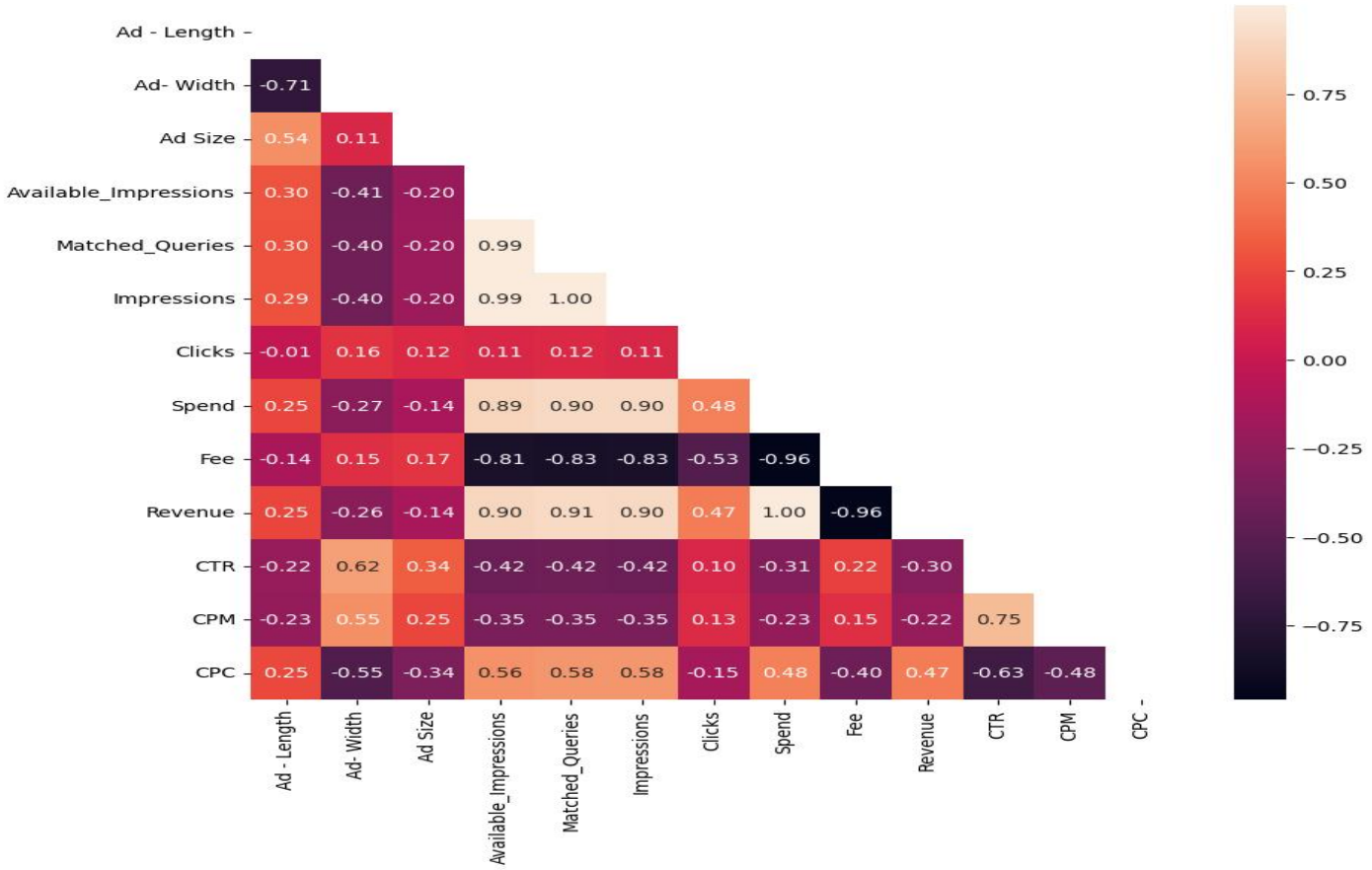


Fig-16-correlation map

Observation

- 'Spend' shows high correlation with 'Available impression','Matched Queries','impression'
- "Revenue"shows high correlation with 'Available impression','Matched Queries','impression','Spend'
- "Fee"negatively correlated with 'Available impression','Matched Queries','impression','Spend','Revenue'
- 'CPC'negatively correlated with 'CTR','CPM'

#barplot of impression with Ad Type

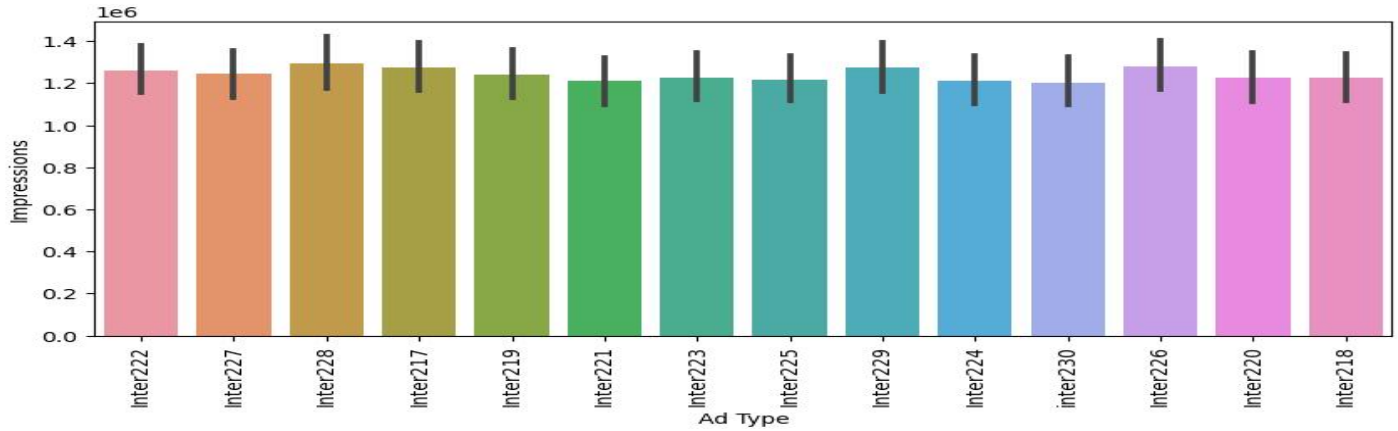


Fig-17-barplot of impression with Ad Type

impression for every Ad Types are approximately same

#barplot of CTR with Device Type

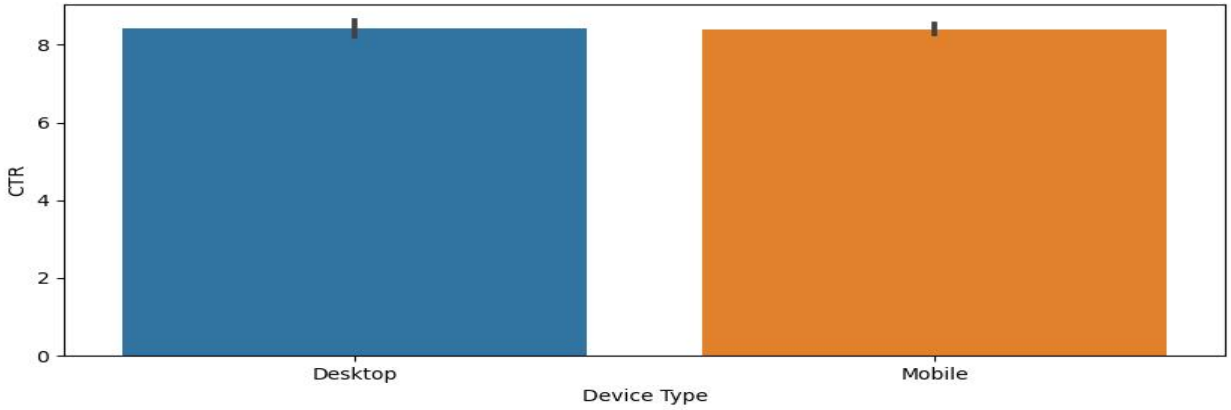


Fig-18-barplot of CTR with Device Type

CTR for both device type are same

#barplot with CPM with Device Type

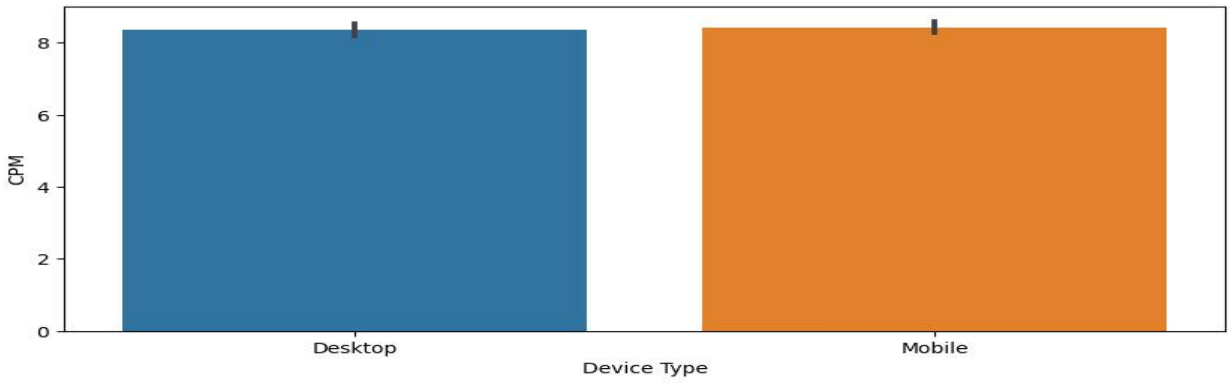


Fig-19-barplot with CPM with Device Type

CPM for both device type are same

#barplot of clicks with device type

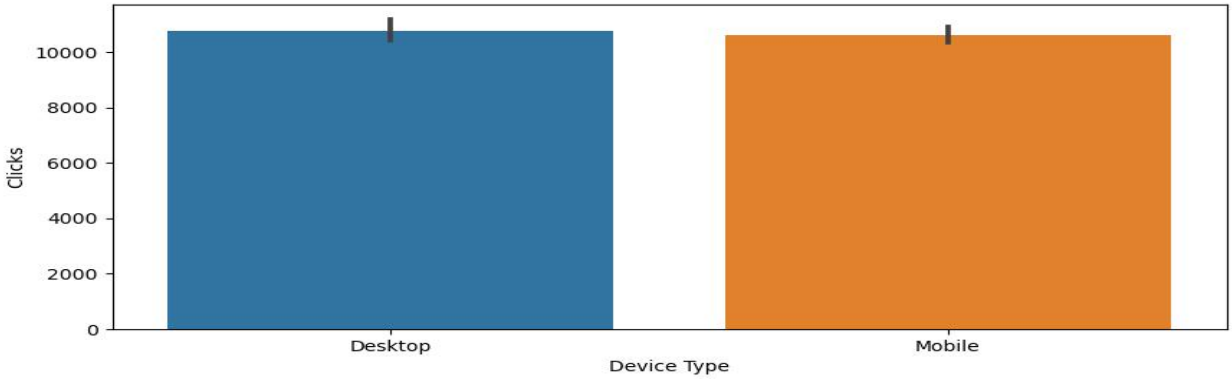


Fig-20-barplot of clicks with device type

Clicks for both device type are same.

Key meaningful observations on individual variables and the relationship between variables

- There are 13 numeric fields in the data
- Customer ad length ranges from 120 to 728
- maximum of Ad_width is 600
- Ad size ranges from 33600 to 216000 with an average 7200
- Available_Impressions ranges from 1.000000e+00 to 2.759286e+07
- only few matched queries above 1.4
- Average impression is around 2.252900e+05
- more Ads are zero clicks very less ads are more than 80000 Clicks
- Spend range is 0.00 to 26931.870000

- 50%,75%,maximum of fee of Ads is 350000
- maximum of Revenue of Ad is 21276.180000
- CTR range in between 0.00 to 1.00
- maximum of CPM is 81.560000
- Range of CPC 0.000 to 7.260000 with average 0.160000
- Outliers to be treated
- more inventory type is Format1
- counts of All Ad Type is approximately same
- most count Ad are in video platform,less counts in App platform
- mobile device have high Ad than desktop
- Format of Ads video type and display type are approximately same
- *relationship between variables*
- Spend' shows high correlation with 'Available impression','Matched Queries','impression'
- "Revenue"shows high correlation with 'Available impression','Matched Queries','impression','Spend'
- "Fee"negatively correlated with 'Available impression','Matched Queries','impression','Spend','Revenue'
- 'CPC'negatively correlated with 'CTR','CPM'
- CTR for both device type are same
- CPM for both device type are same
- Clicks for both device type are same

Part 1:--2) Clustering: Data Preprocessing

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

1---2)-a)Missing value check and treatment

#check missing values
CPM,CTR,CPC columns have `4736` *null values,this null values treat using these equations.*

CPM = (Total Campaign Spend / Number of Impressions) * 1000

CPC = Total Cost (spend) / Number of Clicks**.**

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

1--2)-b)Check if there are any outliers

#Check for presence of outliers in each feature

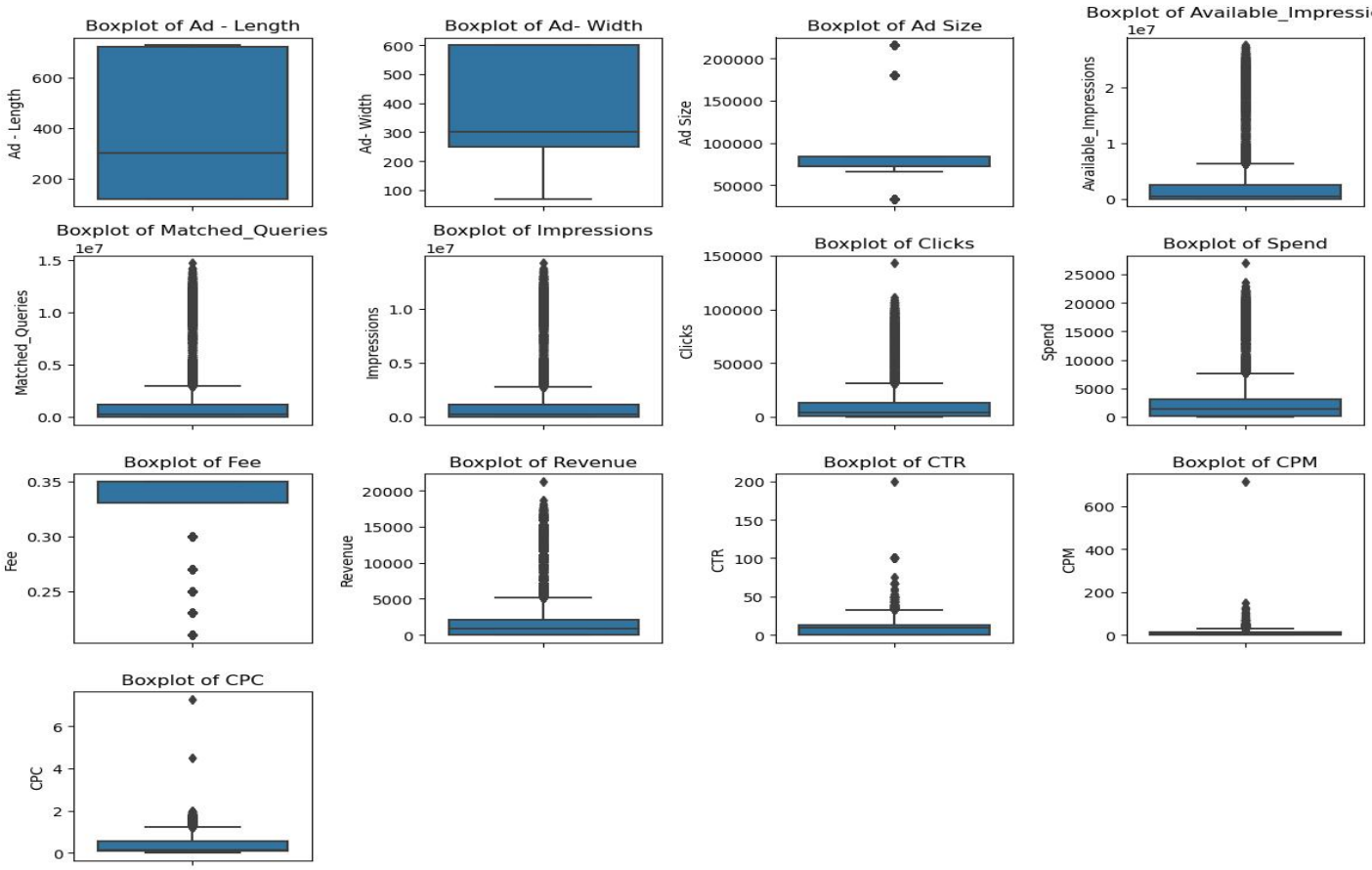


Fig-21-check outliers

 # treate outliers using IQR method

1--2-c)-z-score scaling

#scaling data frame using z-score method, Scaled data

	Ad - Length	Ad- Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
0	-0.36449	-0.43279	-0.352218	-0.512407	-0.515248	-0.51091	-0.61531	-0.66537	0.46544 7	-0.61969	-0.87459	-0.92705	-0.98661

Ad-Length	Ad-Width	Ad Size	Available_Impressions	Matched_Queries	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC	
	6	7				8	1	2		3	3	4	5
1	-0.364496	-0.432797	-0.352218	-0.512413	-0.515264	-0.510933	-0.611311	-0.665372	0.465447	-0.619693	-0.870136	-0.927054	-0.986615
2	-0.364496	-0.432797	-0.352218	-0.512213	-0.515235	-0.510905	-0.611311	-0.665372	0.465447	-0.619693	-0.877606	-0.927054	-0.986615
3	-0.364496	-0.432797	-0.352218	-0.512276	-0.515179	-0.510847	-0.611311	-0.665372	0.465447	-0.619693	-0.886208	-0.927054	-0.986615
4	-0.364496	-0.432797	-0.352218	-0.512531	-0.515281	-0.510951	-0.611311	-0.665372	0.465447	-0.619693	-0.863404	-0.927054	-0.986615
...	...	...	...	...	...	...	...	...	...	...	...	...	...
23061	1.433093	-0.186599	1.939086	-0.512788	-0.515377	-0.511050	-0.611311	-0.665355	0.465447	-0.619678	9.888962	6.801294	-0.781484
23062	1.433093	-0.186599	1.939086	-0.512787	-0.515376	-0.511050	-0.611311	-0.665362	0.465447	-0.619684	4.490471	1.281046	-0.869397
23063	1.433093	-0.186599	1.939086	-0.512788	-0.515377	-0.511050	-0.611311	-0.665360	0.465447	-0.619682	9.888962	4.593195	-0.840092
23064	-1.134891	1.290590	-0.400970	-0.512787	-0.515377	-0.511050	-0.611311	-0.665355	0.465447	-0.619678	9.888962	6.801294	-0.781484
23065	1.433093	-0.186599	1.939086	-0.512788	-0.515376	-0.511050	-0.611311	-0.665350	0.465447	-0.619674	4.490471	4.041170	-0.722875

23066 rows × 13 columns Table-3- Scaled data

Part 1:--3) Clustering: Hierarchical Clustering

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

1---3-a)-Construct a dendrogram using Ward linkage and Euclidean distance

#plot dendrogram

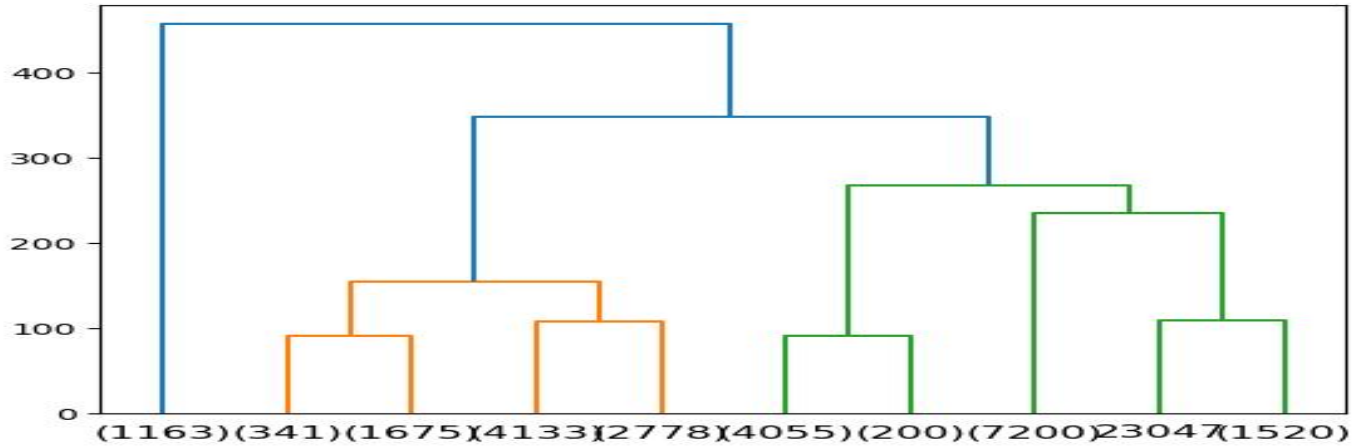


Fig-22-dendrogram

1-3-b) Identify the optimum number of Clusters

In this example, you might decide to cut the dendrogram at a certain height (horizontal line) where the vertical lines are longest, and you would consider the number of clusters as the number of branches you've crossed. we can extract the optimal number of clusters by looking dendrogram.By looking at the dendrogram ,we can say that the optimmal number of clusters is 2.because the vertical distance is high for 2 clusters

Optimum number of clusters=2

Part 1--4)-: Clustering: K-means Clustering

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

1--4--a)Apply K-means Clustering

Elbow plot(up to n=10)

Wss=

```
[299858.0,
 188281.49,
 130712.22,
 94685.77,
 66289.08,
 55262.54,
 49134.2,
 44094.3,
 40161.91,
 34878.56]
```

1--4-b)-Plot the Elbow curve¶

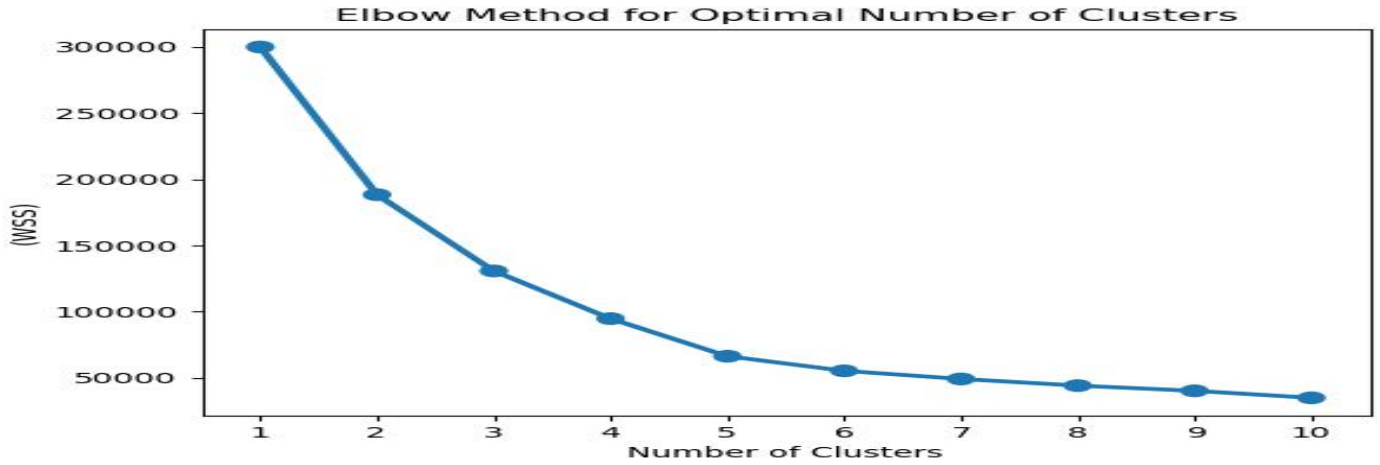


Fig-23-Elbow curve

We can see from the plot that there is a consistent dip from 2 to 10 and there doesn't seem to be a clear 'elbow' here. We may choose any from 2 to 8 as our # of clusters. So, let's look at another method to get a 'second opinion from maths'. Let's create a plot with Silhouette scores to see how it varies with k.

1-4-c)-Check Silhouette Scores

Let us now find the Silhouette Score for the values of K from 2 to 10

```
# Silhouette Analysis
For n_clusters=2, the silhouette score is 0.6122159547090302
For n_clusters=3, the silhouette score is 0.38939835242959714
For n_clusters=4, the silhouette score is 0.5033463240941357
For n_clusters=5, the silhouette score is 0.557117025455947
For n_clusters=6, the silhouette score is 0.5291811252973383
For n_clusters=7, the silhouette score is 0.5367029944777137
For n_clusters=8, the silhouette score is 0.5381765938866728
For n_clusters=9, the silhouette score is 0.5531121077559789
For n_clusters=10, the silhouette score is 0.5534060301063022
```

1-4-d)Figure out the appropriate number of clusters

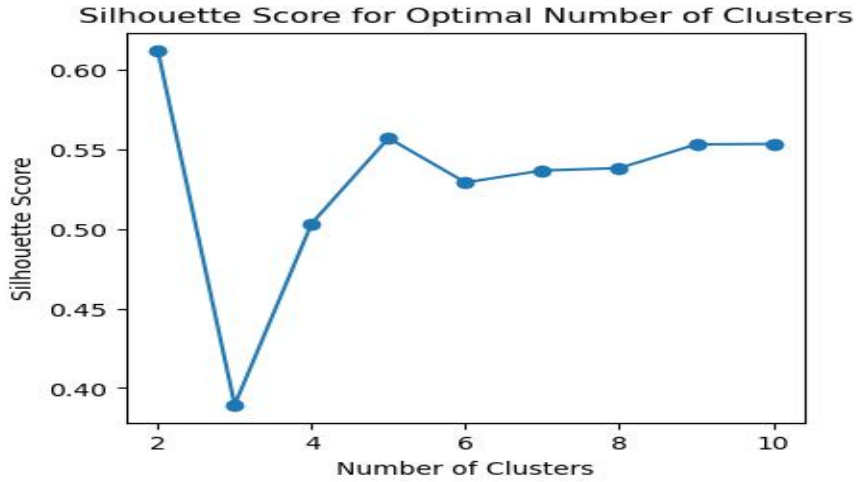


Fig-24-plot of silhouette score

1--4--e)Cluster Profiling

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.]

We can see from the plot that silhouette score is highest for k=2. Well that makes it slightly easy for us and we can start with first understanding these 2 clusters. So let's take the number of clusters as 2.

#let's take the number of clusters as 2.

#Adding predicted labels to the original data and scaled data

#counts for clusters

Observations:

This looks like a very skewed clustering . Let's check out the profiles of these clusters

#Calculating mean and median of the original data for each label

	group_0 Mean	group_1 Mean	group_0 Median	group_1 Median
Ad - Length	3.645779e+02	6.841050e+02	300.000000	7.280000e+02
Ad- Width	3.531738e+02	1.160296e+02	300.000000	9.000000e+01
Ad Size	9.852045e+04	6.986673e+04	72000.000000	6.552000e+04
Available_Impressions	1.349530e+06	1.815254e+07	408718.000000	1.894390e+07
Matched_Queries	7.173359e+05	9.685496e+06	221739.000000	1.024037e+07
Impressions	6.848326e+05	9.325842e+06	188945.500000	9.857084e+06
Clicks	1.019593e+04	1.768678e+04	3912.000000	1.896250e+04

	group_0 Mean	group_1 Mean	group_0 Median	group_1 Median
Spend	1.818878e+03	1.559868e+04	1299.980000	1.607917e+04
Fee	3.417966e-01	2.382100e-01	0.350000	2.300000e-01
Revenue	1.234364e+03	1.194295e+04	844.985000	1.238096e+04
CTR	8.976118e+00	1.878036e-01	10.178568	1.904888e-01
CPM	8.857767e+00	1.703308e+00	9.106158	1.667699e+00
CPC	2.968454e-01	9.151287e-01	0.127471	8.573520e-01

Table-4- Calculating mean and median of the original data for each label

Observations:

It looks like Cluster 1 with high impression and CPC, Cluster 0 is of less impression,CPC than cluster 1 .

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.]

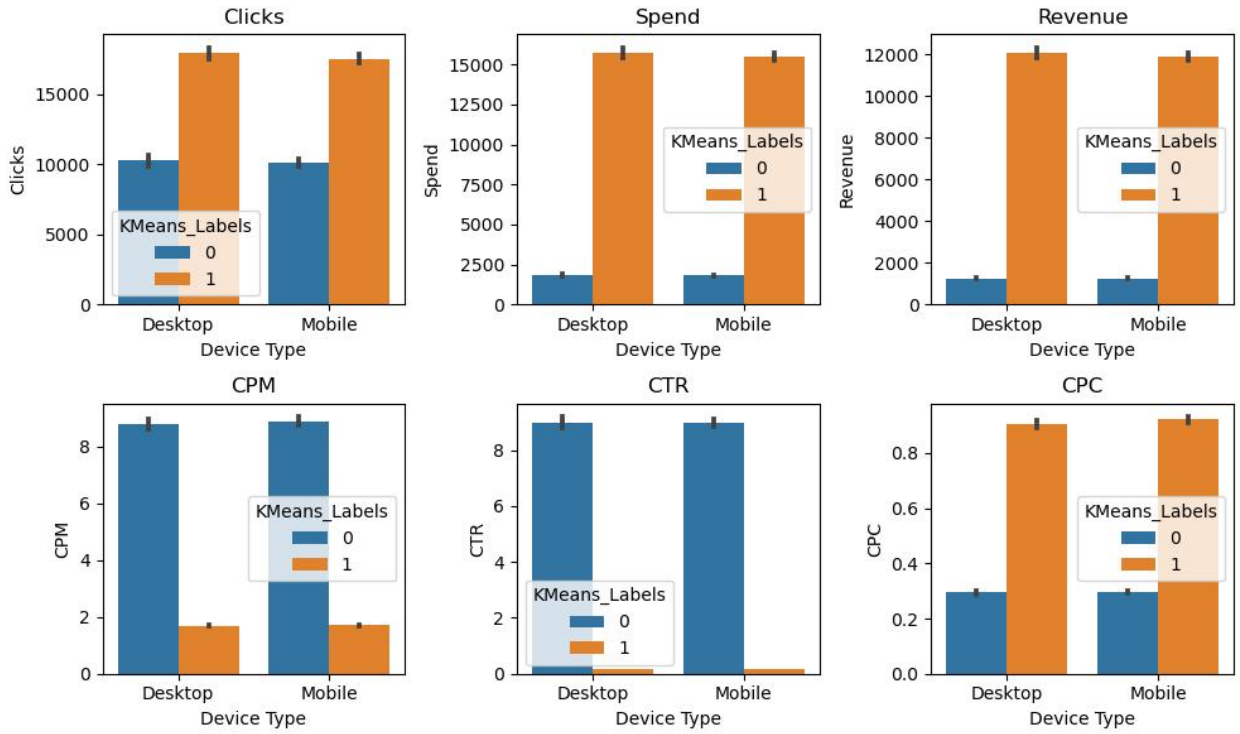


Fig-25-barplot for *Group the data by clusters and take sum or mean to identify trends in clicks, spend, revenue, CPM, CTR, & CPC based on Device Type*

Conclude the project by providing summary of your learnings.

- There are 23066 rows and 19 columns into the dataset.
- There are no duplicated values in data frame.
- There are 4736 null values in CTR,CPM,CPC columns.
- I have treated missing values in CPC,CTR,CPM columns using the given formula
- It seems thet there are outliers into the dataset
- We treated outliers using IQR method
- I have applied z-score metod on data frame for scaling .
- I have plotted Dendrogram for value of P=10
- plotted elbow plot and got the optimum value of cluster is 2.
- As per Elbow plot/scree plot ,we conclude that the optimal nummber of clusters should be 5.
- I have create 2 clusters for the data set.

conclusion after clustering ``

- When clicks on Ads gets increases then revenue is also increases.
- When amount of money spent on specific Ad variation within a specific campaign or ad set is increases then revenue is also increases.
- when impression count of the particular Advertisement increasa then revenue is also increases.

- the clicks for both device type have approximately same in both clusters.
- all variables are same in both devices type.

- click,spend,revenue,cpc are high in cluster 1 type.
- CPM and CTR are high in cluster 0 type .

Problem 2:

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

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

DATA DICTIONARY

State Code: Code representing the state.

Dist. Code: Code representing the district within the state.

State, Area Name: Name of the state AND district.

No_HH: Number of households.

TOT_M: Total male population.

TOT_F: Total female population.

M_06, F_06: Male and female population in the age group 0-6.

M_SC, F_SC: Male and female population belonging to the Scheduled Caste.

M_ST, F_ST: Male and female population belonging to the Scheduled Tribe.

M_LIT, F_LIT: Male and female literate population.

M_ILL, F_ILL: Male and female illiterate population.

TOT_WORK_M, TOT_WORK_F: Total working male and female population.

MAINWORK_M, MAINWORK_F: Main working male and female population.

MAIN_CL_M, MAIN_CL_F: Main working male and female population engaged in cultivation.

MAIN_AL_M, MAIN_AL_F: Main working male and female population engaged in agriculture and allied activities.

MAIN_HH_M, MAIN_HH_F: Main working male and female population engaged in household industry.

MAIN_OT_M, MAIN_OT_F: Main working male and female population engaged in other occupations.

MARGWORK_M, MARGWORK_F: Marginal working male and female population.

MARG_CL_M, MARG_CL_F: Marginal working male and female population engaged in cultivation.

MARG_AL_M, MARG_AL_F: Marginal working male and female population engaged in agriculture and allied activities.

MARG_HH_M, MARG_HH_F: Marginal working male and female population engaged in household industry.

MARG_OT_M, MARG_OT_F: Marginal working male and female population engaged in other occupations.

MARGWORK_3_6_M, MARGWORK_3_6_F: Marginal working male and female population in the age group 3-6.

MARG_CL_3_6_M, MARG_CL_3_6_F: Marginal working male and female population in the age group 3-6 engaged in cultivation.

MARG_AL_3_6_M, MARG_AL_3_6_F Marginal working male and female population in the age group 3-6 engaged in agriculture and allied activities.

MARG_HH_3_6_M, MARG_HH_3_6_F: Marginal working male and female population in the age group 3-6 engaged in household industry.

MARG_OT_3_6_M, MARG_OT_3_6_F: Marginal working male and female population in the age group 3-6 engaged in other occupations.

MARGWORK_0_3_M, MARGWORK_0_3_F: Marginal working male and female population in the age group 0-3.

MARG_CL_0_3_M, MARG_CL_0_3_F: Marginal working male and female population in the age group 0-3 engaged in cultivation.

MARG_AL_0_3_M, MARG_AL_0_3_F: Marginal working male and female population in the age group 0-3 engaged in agriculture and allied activities.

MARG_HH_0_3_M, MARG_HH_0_3_F: Marginal working male and female population in the age group 0-3 engaged in household industry.

MARG_OT_0_3_M, MARG_OT_0_3_F: Marginal working male and female population in the age group 0-3 engaged in other occupations.

NON_WORK_M, NON_WORK_F: Non-working male and female population.

2-1: PCA: Define the problem and perform Exploratory Data Analysis

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

ANSWER

#read the data set

2-1-b) Check shape

```
(640, 61)-640 rows,61 columns
```

2-1-c) Data types

```
<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
```

2-1-d)statistical summary

	State Code	Dist.Code	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	...	MARG_CL_0_3_M	MARG_CL_0_3_F	MARG_AL_0_3_M	MARG_AL_0_3_F	MARG_HH_0_3_M	MARG_HH_0_3_F	MARG_OT_0_3_M	MARG_OT_0_3_F	NON_WORK_M	NON_WORK_F
count	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	...	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000
mean	17.114062	320.500000	51222.871875	79940.576563	122372.084375	12309.098438	11942.300000	13820.946875	20778.392188	6191.807813	...	1392.973438	2757.050000	250.889062	558.098438	560.690625	1293.431250	71.379688	200.742188	510.014063	704.778125
std	9.426448	184.896	48135.405475	73384.511114	113600.717728	11500.906881	11326.294567	14426.373130	21727.887713	9912.668948	...	1489.707052	2788.776676	453.336594	1117.642748	762.578991	1585.377936	107.897627	309.740854	610.603187	910.209225

	State Code	Dist. Code	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	. .	MARG_CL_0_3_M	MARG_CL_0_3_F	MARG_AL_0_3_M	MARG_AL_0_3_F	MARG_HH_0_3_M	MARG_HH_0_3_F	MARG_OT_0_3_M	MARG_OT_0_3_F	NON_WORK_M	NON_WORK_F
	6	367			2																
min	1.000000	1.000000	350.000000	391.000000	698.000000	56.000000	56.000000	0.000000	0.000000	0.000000	. . .	4.000000	30.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	5.000000
25%	9.000000	160.750000	1948.000000	3022.000000	4651.750000	4733.750000	4672.250000	3466.250000	5603.250000	293.750000	. . .	4895.000000	9572.500000	47.000000	109.000000	136.500000	298.000000	14.000000	43.000000	161.000000	220.500000
50%	18.000000	320.500000	3583.700000	5833.900000	8724.500000	9159.000000	8663.000000	9591.500000	13709.000000	2333.500000	. . .	9490.000000	1928.000000	114.500000	247.500000	308.000000	717.000000	35.000000	113.000000	326.000000	464.500000
75%	24.000	480.2	6889.200	10791.85	16425.17	16520.25	15902.25	19429.75	29180.00	7658.000	. . .	1714.0000	3599.7500	270.7500	568.7500	642.0000	1710.7500	79.0000	240.0000	604.5000	853.5000

	State Code	Dist. Code	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	...	MARG_CL_0_3_M	MARG_CL_0_3_F	MARG_AL_0_3_M	MARG_AL_0_3_F	MARG_HH_0_3_M	MARG_HH_0_3_F	MARG_OT_0_3_M	MARG_OT_0_3_F	NON_WORK_M	NON_WORK_F
	00000	50000	0000	0000000	5000	0000	0000	0000	0000	0		00	00	0	0	0	00	00	0	00	00
max	35.000000	640.000000	3104500.000000	4854170.000000	7503920.000000	9622 3.000000	9512 9.000000	1033 07.000000	1564 29.000000	9678 5.000000	...	9875. 00000	2161 1.000000	5775. 00000	1715 3.000000	6116. 00000	1371 4.000000	895.0 00000	3354. 00000	6456 .000000	1053 3.000000

8 rows × 59 columns

Table-5-statical summary

2-1-e)-Perform an EDA on the data to extract useful insights Note:

2-1-e)-1 . Pick 5 variables out of the given 24 variables below for EDA: No_HH, TOT_M, TOT_F, M_06, F_06, M_SC, F_SC, M_ST, F_ST, M_LIT, F_LIT, M_ILL, F_ILL, TOT_WORK_M, TOT_WORK_F, MAINWORK_M, MAINWORK_F, MAIN_CL_M, MAIN_CL_F, MAIN_AL_M, MAIN_AL_F, MAIN_HH_M, MAIN_HH_F, MAIN_OT_M, MAIN_OT_F

Answers

I have picked 5 variables such as TOT_M, TOT_F, M_LIT, F_LIT and TOT_WORK_M and comparing those 5 variables aganist "state","Dist.code"

TOT_M-total population male

TOT_F-total population female

M_LIT-literates population male

F_LIT-literates popuation female

TOT_WORK_M-total work population male

State-state code

District-district code

2-1-e)-2 . Example questions to answer from EDA -

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

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

Answers:

2-1-e)2-(i) Which state has highest gender ratio and which has the lowest?
using the bar plot we can find which state has highest gender ratio and which has the lowest
ans:
gender ratio=(total population of female/total population of male)*1000
create a column in data for store the gender ratio
#barplot for state based by gender ratio

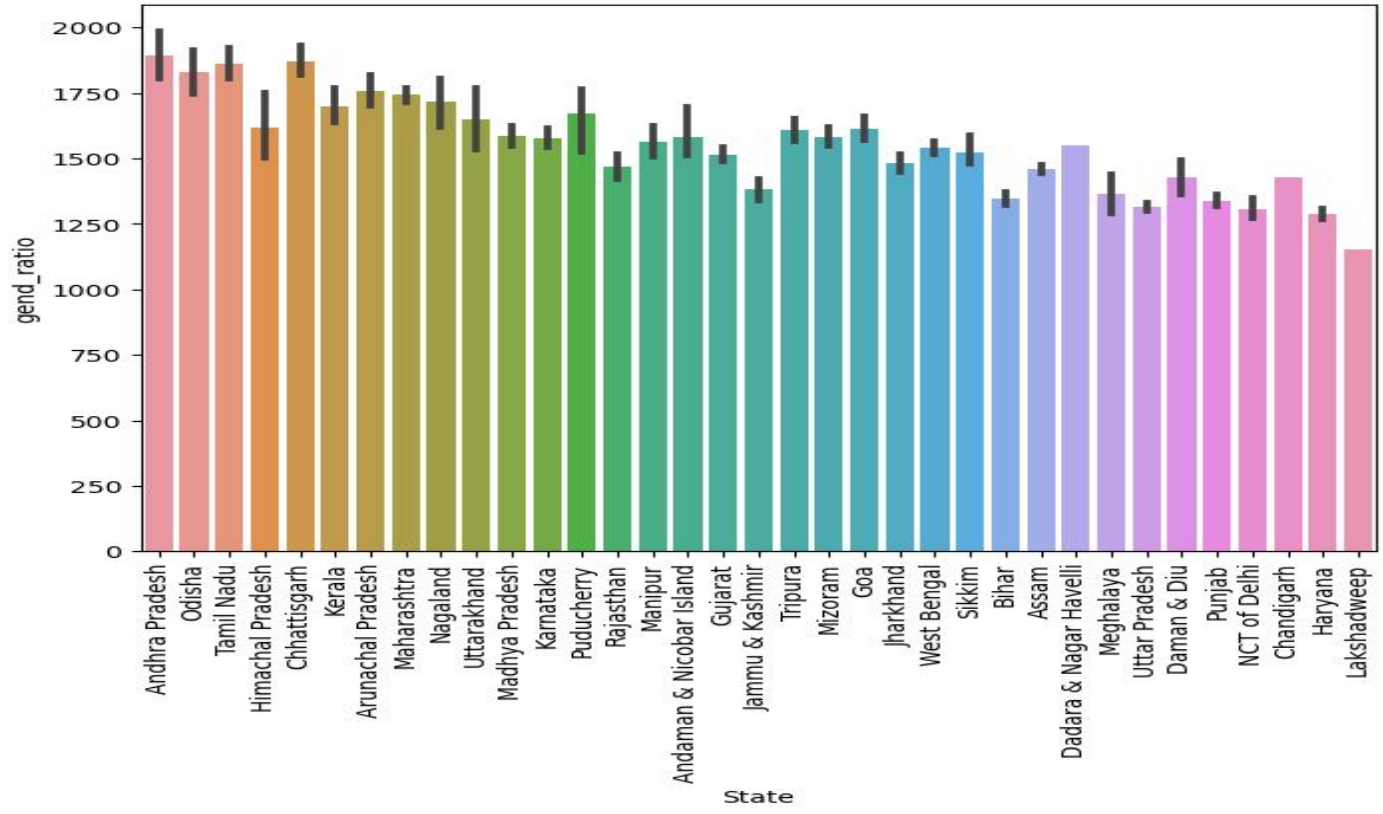


Fig-26-barplot for state based by gender ratio

```
546    Andhra Pradesh
397            Odisha
624        Tamil Nadu
545    Andhra Pradesh
390            Odisha
```

```
            ...
138      Uttar Pradesh
105          Rajasthan
143      Uttar Pradesh
1       Jammu & Kashmir
586         Lakshadweep
Name: State, Length: 640, dtype: object
```

Andhra pradesh state have higest gender ratio

Lakshadweep state have lowest gender ratio

2-1-e)2-(ii) Which district has the highest & lowest gender ratio?¶

Ans:sort values Area name based on gender ratio

```
546           Krishna

397           Koraput
624       Virudhunagar
545      West Godavari
390              Baudh
             ...
138            Baghpat
105           Dhaulpur
143     Mahamaya Nagar
1               Badgam
586        Lakshadweep
Name: Area Name, Length: 640, dtype: object
```

krishna have high gender ratio

lakshadweep have low gender ratio

Part 2-2: PCA: Data Preprocessing

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

2-2-a)Check for and treat (if needed) missing values

there is no null values in the dataset

2-2-b)Check for and treat (if needed) data irregularities

there is no duplicated values in the data set

check count the counts and unique values of columns for checking irregularities

#create ne data frame with drop the State Code","Dist.Code","State","Area Name","gend_ratio"]

Check the outliers

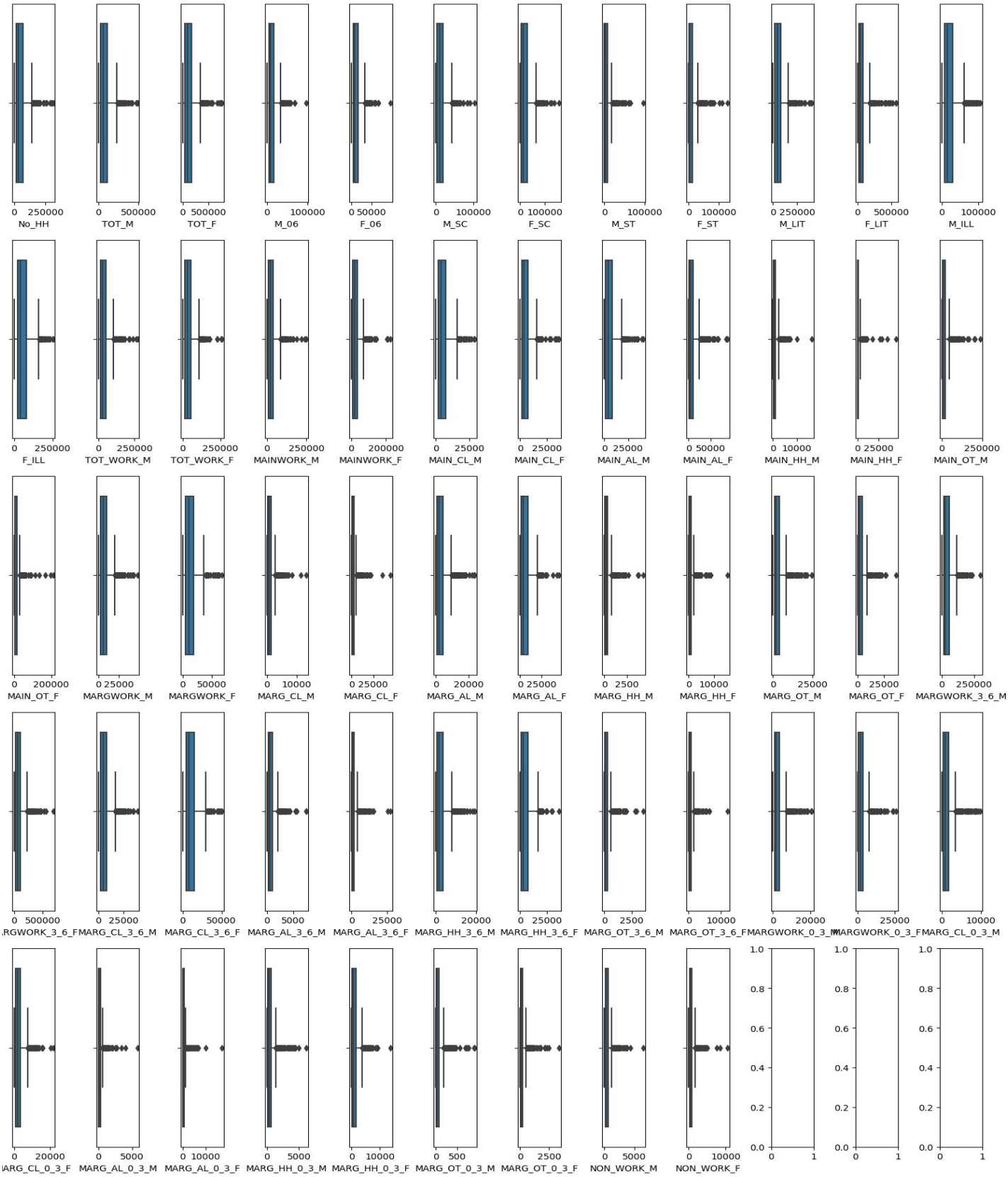

No_HH
TOT_M
TOT_F
M_06
F_06
M_SC
F_SC
M_ST
F_ST
M_LIT
F_LIT
M_ILL
F_ILL
TOT_WORK_M
TOT_WORK_F
MAINWORK_M
MAINWORK_F
MAIN_CL_M
MAIN_CL_F
MAIN_AL_M
MAIN_AL_F
MAIN_HH_M
MAIN_HH_F
MAIN_OT_M
MAIN_OT_F
MARGWORK_M
MARGWORK_F
MARG_CL_M
MARG_CL_F
MARG_AL_M
MARG_AL_F
MARG_HH_M
MARG_HH_F
MARG_OT_M
MARG_OT_F
MARGWORK_3_6_M
RGWORK_3_6_F
MARG_CL_3_6_M
MARG_CL_3_6_F
MARG_AL_3_6_M
MARG_AL_3_6_F
MARG_HH_3_6_M
MARG_HH_3_6_F
MARG_OT_3_6_M
MARG_OT_3_6_F
MARGWORK_0_3_M
MARGWORK_0_3_F
MARG_CL_0_3_M
ARG_CL_0_3_F
MARG_AL_0_3_M
MARG_AL_0_3_F
MARG_HH_0_3_M
MARG_HH_0_3_F
MARG_OT_0_3_M
MARG_OT_0_3_F
NON_WORK_M
NON_WORK_F

Fig-27-checking outliers

#treat the outliers

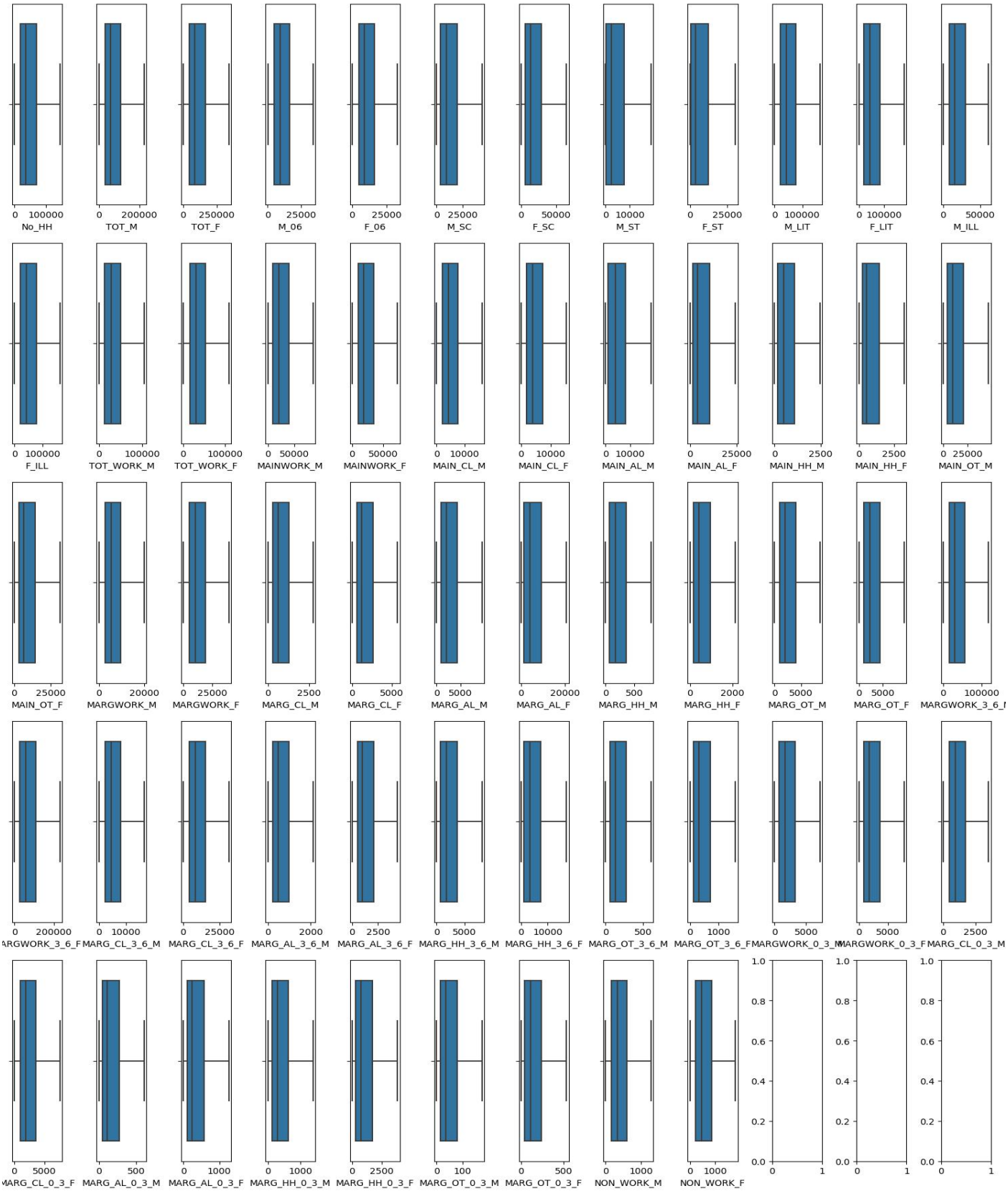


figure-28-treated outliers

2-2-c)scaling data using z-score method¶

Scaling the data using 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	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	-1.038986	-0.874837	-0.937027	-0.624685	-0.561282	-0.080201	-0.079963	-0.510441	-0.574198	-0.939617	...	-0.093587	-0.860882	-0.041800	-0.423378	0.523468	-0.793688	-0.443385	-0.756991	-0.497493	-0.774865
1	-1.076896	-0.938023	-0.009723	-0.773932	-0.835623	-0.079863	-0.079635	-0.771833	-0.782092	-1.005083	...	-0.719169	-0.877096	-0.347040	-0.443790	-0.634478	-0.848805	0.358782	0.088391	-0.813235	-0.880662
2	-1.218558	-1.154665	-1.414153	-1.164642	-1.381104	-0.080201	-0.079635	0.122588	0.137599	-1.414156	...	-1.130551	-1.128423	-0.770091	-0.800999	-1.063434	-0.984684	-1.026779	-1.025976	-1.008568	-1.074622
3	-1.201599	-2.117171	-2.149301	-1.977720	-1.760911	-0.080447	-0.079963	-0.399531	-0.437330	-2.030090	...	-1.050477	-1.100286	-0.679055	-0.701491	-1.039994	-0.970666	-0.953855	-0.987550	-0.877454	-1.078541

No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	F_ST	M_LIT	...	MARG_CL_03_M	MARG_CL_03_F	MARG_AL_03_M	MARG_AL_03_F	MARG_HH_03_M	MARG_HH_03_F	MARG_OT_03_M	MARG_OT_03_F	NON_WORK_M	NON_WORK_F	
4	-0.938495	-0.921309	-0.935018	-0.700931	-0.740523	-1.078807	-1.078160	0.432534	0.249489	-0.942767	...	-0.369844	-0.298617	1.484398	1.633130	-0.589942	-0.749882	-0.589234	-0.379131	-0.706204	-0.257637
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
635	-1.150348	-1.127949	-1.132667	-1.134720	-1.117495	-1.078725	-1.078324	-0.842834	-0.833741	-0.871892	...	-1.212628	-1.195666	-1.057114	-1.028083	-1.070466	-1.014472	-1.026779	-1.051594	-1.102220	-1.102051
636	-0.965024	-1.058299	-1.025046	-1.091676	-1.080734	-0.897218	-0.852926	-0.842834	-0.833741	-1.010127	...	-1.089513	-1.057365	-0.989649	-0.992362	-0.981393	-0.900575	-0.953855	-0.904293	-0.893509	-0.861070
637	-1.202745	-1.237690	-1.232045	-1.234110	-1.226443	-1.080478	-1.079963	-0.674559	-0.658914	-1.217960	...	-1.140561	-1.154175	-0.957519	-1.017877	-1.065778	-1.009215	-0.716851	-0.750587	-0.984485	-1.043275

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	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	
638	-1.139425	-1.177029	-1.173597	-1.180468	-1.177539	-0.804479	-0.799631	-0.838178	-0.828746	-1.149048	...	-1.108531	-1.136053	-0.877193	-0.915817	-1.044682	-0.960733	-1.085348	-1.025976	-0.920267	-0.992336
639	-1.032162	-0.064430	-0.645970	-0.998068	-1.002296	-0.804479	-0.799631	-0.816076	-0.807367	-0.014624	...	-1.071496	-1.159898	-0.973584	-1.022980	-1.030617	-0.995978	-0.903017	-1.025976	-0.791829	-1.001773

640 rows × 57 columns

Table-6-scaled data

2-2-4)Visualize the data before and after scaling and comment on the impact on outliers

i have already treated outliers in above question 2-2-b) ,but still i applied the z-score for the scaling of the data set .please find below outputs by boxplot and describe function for before and after

before scaling

```
# Assuming dx_new is your DataFrame,  plot boxplot for checking outliers
```

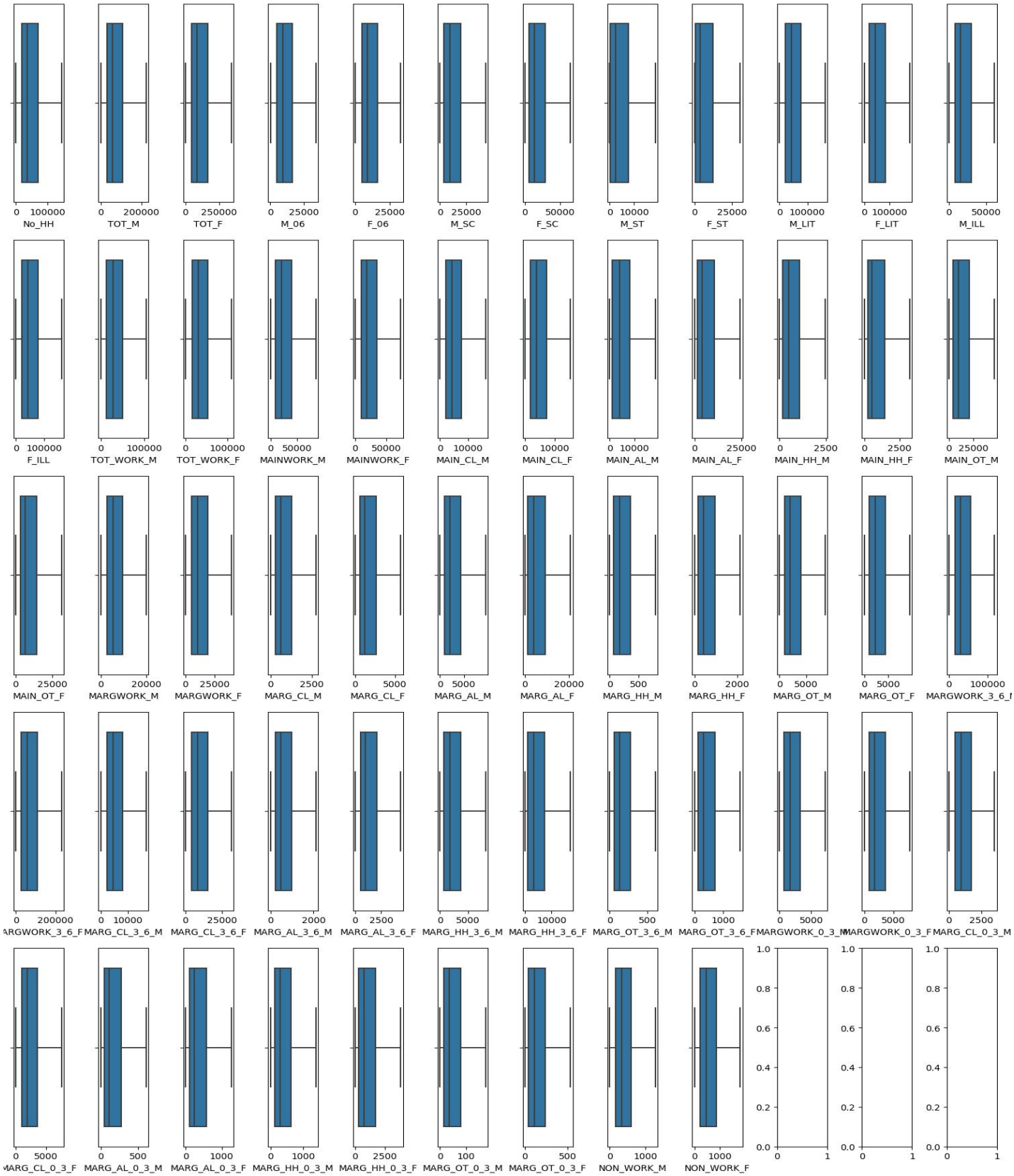


Fig-29-before scalingboxplot for checking outliers

After scaling:*# Assuming dx_new is your DataFrame, after scalingboxplot for checking outliers*

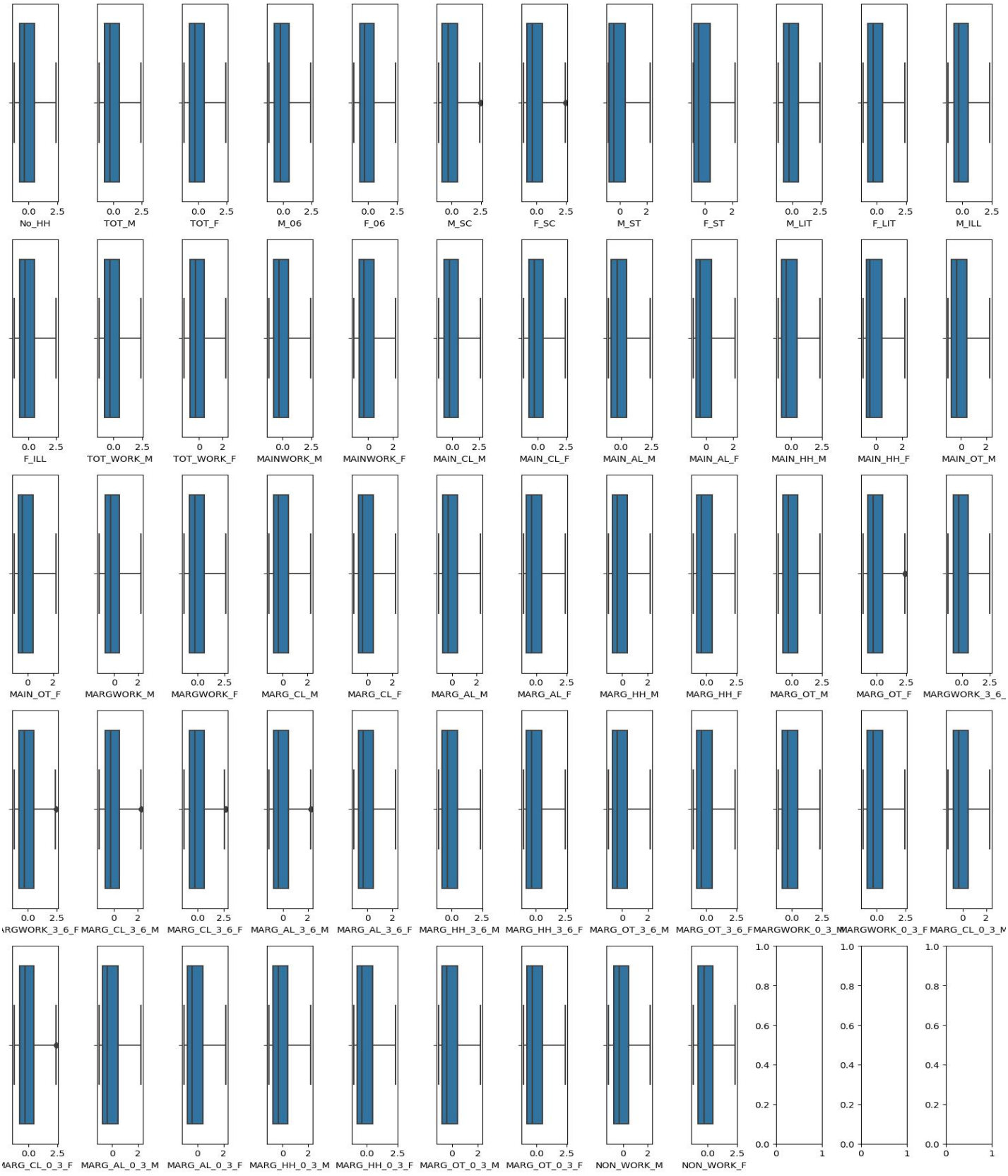

No_HH
TOT_M
TOT_F
M_06
F_06
M_SC
F_SC
M_ST
F_ST
M_LIT
F_LIT
M_ILL
F_ILL
TOT_WORK_M
TOT_WORK_F
MAINWORK_M
MAINWORK_F
MAIN_CL_M
MAIN_CL_F
MAIN_AL_M
MAIN_AL_F
MAIN_HH_M
MAIN_HH_F
MAIN_OT_M
MAIN_OT_F
MARGWORK_M
MARGWORK_F
MARG_CL_M
MARG_CL_F
MARG_AL_M
MARG_AL_F
MARG_HH_M
MARG_HH_F
MARG_OT_M
MARG_OT_F
MARGWORK_3_6_
RGWORK_3_6_F
MARG_CL_3_6_M
MARG_CL_3_6_F
MARG_AL_3_6_M
MARG_AL_3_6_F
MARG_HH_3_6_M
MARG_HH_3_6_F
MARG_OT_3_6_M
MARG_OT_3_6_F
MARGWORK_0_3_M
MARGWORK_0_3_F
MARG_CL_0_3_M
ARG_CL_0_3_F
MARG_AL_0_3_M
MARG_AL_0_3_F
MARG_HH_0_3_M
MARG_HH_0_3_F
MARG_OT_0_3_M
MARG_OT_0_3_F
NON_WORK_M
NON_WORK_F

Fig-30-after scalingboxplot for checking outliers

The range of the outliers changes after the scaling.

before scaling

	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	F_ST	M_LIT	...	MARG_CL_03_M	MARG_CL_03_F	MARG_AL_03_M	MARG_AL_03_F	MARG_HH_03_M	MARG_HH_03_F	MARG_OT_03_M	MARG_OT_03_F	NON_WORK_M	NON_WORK_F
count	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	...	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000	640.000000
mean	485.154218	760.419538	1160.798085	1163.809675	1123.450820	1317.319675	1976.436503	5068.761133	834.564804	5454.874219	...	1243.500000	2554.161719	187.805664	402.933008	456.679297	1157.905078	56.320312	164.198438	443.924219	609.501562
std	308.008223	603.383621	921.545443	925.364941	893.379265	1220.189292	1831.527610	6018.652465	1007.745145	4384.346970	...	999.851461	2098.515606	186.884611	392.232300	426.951049	1142.279691	54.894326	156.264559	374.014651	510.812596
min	350.000000	391.000000	698.000000	563.000000	563.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	4.000000	30.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	5.000000

	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	F_ST	M_LIT	...	MARG_CL_03_M	MARG_CL_03_F	MARG_AL_03_M	MARG_AL_03_F	MARG_HH_03_M	MARG_HH_03_F	MARG_OT_03_M	MARG_OT_03_F	NON_WORK_M	NON_WORK_F	
	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
25%	1	3	4	4	4	4	3	5	2	4	2	...	48	95	47	10	13	29	14	43	1	2
	9	0	6	7	6	6	4	6	9	2	1		9.	7.	.0	9.	6.	8.	.0	.0	6	2
	4	2	5	3	7	7	6	0	3.	9.	2		50	25	00	00	50	00	00	00	1.	0.
	8.	8.	1	3.	2.	2.	6.	3.	7	5	9		00	00	00	00	00	00	00	00	00	50
	0	0	7.	7	2.	2	2	2	5	0	8.		0	0	0	0	0	0	0	0	0	0
	0	0	7	5	2	5	5	5	0	0	0										0	0
	0	0	5	0	5	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	0										0	0
50%	3	5	8	9	8	9	1	1	2	3	4	...	94	19	11	24	30	71	35	11	3	4
	5	8	7	1	6	5	5	3	3	8	2		9.	28	4.	7.	8.	7.	.0	3.	2	6	
	8	3	7	5	6	9	9	7	3.	3	6		00	.0	50	50	00	00	00	00	6.	4.
	3	3	2	9	3.	1.	1.	0	5	4.	9		00	00	00	00	00	00	00	00	0	5
	7.	9.	4.	1.	0	5	5	9.	5	5	3.		00	00	00	00	00	00	00	00	0	0
	0	0	5	5	0	0	0	0	0	0	5		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	0	0	0										0	0
75%	6	1	1	1	1	1	1	2	7	1	7	...	17	35	27	56	64	17	79	24	6	8
	8	0	0	6	6	5	9	9	6	2	7		14	99	0.	8.	2.	10	.0	0.	0	5
	8	7	7	4	5	9	4	1	5	4	9		.0	.7	75	75	00	.7	00	00	4.	3.
	9	9	9	2	2	0	2	8	8	8	8		00	50	00	00	00	50	00	00	5	5
	2.	1.	1	5	0.	2.	9.	0.	0.	0.	9.		00	00	00	00	00	00	00	00	0	0
	0	0	8.	1.	2	2	7	0	0	2	5		0	0	0	0	0	0	0	0	0	0
	0	0	5	7	5	5	5	0	0	5	0										0	0
	0	0	0	5	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
m	1	2	3	3	3	4	6	1	3	1	...	35	75	60	12	14	38	17	53	1	1	

No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	F_ST	M_LIT	...	MARG_CL_03_M	MARG_CL_03_F	MARG_AL_03_M	MARG_AL_03_F	MARG_HH_03_M	MARG_HH_03_F	MARG_OT_03_M	MARG_OT_03_F	NON_WORK_M	NON_WORK_F	
ax	43000	24455	40852	42000	27474	33375	45451	87043	05563	63026		50.750000	63.500000	6.3750	58.375000	00.250000	29.875000	6.500000	5.500000	269.750000	803.000000
	4.000000e+02	4.250000e+02	2.750000e+02	0.000000e+02	7.250000e+02	5.000000e+02	5.125000e+02	4.375000e+02	6.375000e+02	6.750000e+02	...										

8 rows × 57 columns

Table-7-before scaling
After scaling

No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	F_ST	M_LIT	...	MARG_CL_03_M	MARG_CL_03_F	MARG_AL_03_M	MARG_AL_03_F	MARG_HH_03_M	MARG_HH_03_F	MARG_OT_03_M	MARG_OT_03_F	NON_WORK_M	NON_WORK_F	
count	6.400000e+02	6.400000e+02	6.400000e+02	6.400000e+02	6.400000e+02	6.400000e+02	6.400000e+02	6.400000e+02	640.000000	6.400000e+02	...	6.400000e+02	6.400000e+02	6.400000e+02	6.400000e+02	6.400000e+02	640.000000	6.400000e+02	6.400000e+02	6.400000e+02	6.400000e+02
mean	-6.66	-1.333	-2.222	5.551	-3.333	2.220	-2.222	-4.444	0.000	-7.777	...	-1.38777	-3.333306	-2.22204	-6.66613	2.222046	0.000000	-2.222046	-7.21764	-6.666	4.440

No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	F_ST	M_LIT	...	MARG_CL_03_M	MARG_CL_03_F	MARG_AL_03_M	MARG_AL_03_F	MARG_HH_03_M	MARG_HH_03_F	MARG_OT_03_M	MARG_OT_03_F	NON_WORK_M	NON_WORK_F	
	1338e-17	2268e-16	0446e-17	115e-17	0669e-17	4466e-17	0446e-17	0892e-17	0000	1561e-17		79e-17	69e-17	46e-17	38e-17	e-17		46e-17	50e-17	1338e-17	892e-17
std	1.000782e+00	1.000782e+00	1.000782e+00	1.000782e+00	1.000782e+00	1.000782e+00	1.000782e+00	1.000782e+00	1.000782	...	1.000782e+00	1.000782e+00	1.000782e+00	1.000782e+00	1.000782e+00	1.000782e+00	1.000782	1.000782e+00	1.000782e+00	1.000782e+00	1.000782e+00
min	-1.226295e+00	-1.256930e+00	-1.253026e+00	-1.252604e+00	-1.245270e+00	-1.080447e+00	-1.079963e+00	-8.428341e-01	-0.833741	...	-1.240654e+00	-1.203773e+00	-1.005714e+00	-1.028083e+00	-1.070466e+00	-1.014472	-1.026779e+00	-1.051594e+00	-1.187845e+00	-1.184337e+00
25%	-7.391433e-01	-7.611904e-01	-7.554317e-01	-7.467051e-01	-7.310260e-01	-7.961502e-01	-7.737908e-01	-7.939894e-01	-0.790834	...	-7.547019e-01	-7.615673e-01	-7.540257e-01	-7.499694e-01	-7.505070e-01	-0.753386	-7.715445e-01	-7.762044e-01	-7.570439e-01	-7.621304e-01

	No_HH	TOT_M	TOT_F	M_06	F_06	M_SC	F_SC	M_ST	F_ST	M_LIT	...	MARG_CL_03_M	MARG_CL_03_F	MARG_AL_03_M	MARG_AL_03_F	MARG_HH_03_M	MARG_HH_03_F	MARG_OT_03_M	MARG_OT_03_F	NON_WORK_M	NON_WORK_F
50%	-3.227958e-01	-2.941277e-01	-3.079337e-01	-2.681143e-01	-2.864623e-01	-2.937658e-01	-3.308769e-01	-4.548195e-01	-0.450670	-2.705225e-01	...	-2.947741e-01	-2.986165e-01	-3.925577e-01	-3.963870e-01	-3.485074e-01	-0.386289	-3.886921e-01	-3.278957e-01	-3.159739e-01	-2.840865e-01
75%	5.187848e-01	5.296328e-01	5.231388e-01	5.280048e-01	5.199796e-01	5.131537e-01	5.144885e-01	4.305389e-01	0.413052	5.351530e-01	...	4.709380e-01	4.986411e-01	4.441736e-01	4.230817e-01	4.343956e-01	0.484362	4.134749e-01	4.854642e-01	4.296606e-01	4.780408e-01
max	2.405677e+00	2.465868e+00	2.440995e+00	2.440070e+00	2.396488e+00	2.477110e+00	2.446907e+00	2.267331e+00	2.218881	2.476235e+00	...	2.309398e+00	2.389854e+00	2.241473e+00	2.182658e+00	2.211749e+00	2.340985	2.191004e+00	2.377967e+00	2.209731e+00	2.338298e+00

8 rows × 57 columns table-8-after scaling

Part 2-3); PCA: PCA

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

ANSWER:

#Check for presence of correlations

Bartletts Test of Sphericity

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

H0: All variables in the data are uncorrelated

Ha: At least one pair of variables in the data are correlated If the null hypothesis cannot be rejected, then PCA is not advisable.

If the p-value is small(Reject H0 if p-value < 0.05), then we can reject the null hypothesis and agree that there is atleast one pair of vairbales in the data wihich are correlated hence PCA is recommended

```
p-value:0.0
```

the p-value is small(Reject H0 if p-value < 0.05), then we can reject the null hypothesis and agree that there is atleast one pair of vairbales in the data wihich are correlated hence PCA is recommended

KMO Test

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

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

```
MSA- 0.936189616665265
```

MSA > 0.7 is expected to provide a considerable reduction is the dimension and extraction of meaningful components.

2-3-a)-Create the covariance matrix

#Apply PCA taking all features

#co variance matrix

```
array([[-5.52816148e+00,  4.30377559e-01, -1.47382695e+00, ...,
         6.51060294e-03,  2.38391484e-03,  4.36606130e-04],
       [-5.49201646e+00, -1.06110331e-01, -2.01564100e+00, ...,
        -2.82839348e-04,  8.13609312e-03, -6.60428796e-03],
       [-7.47464297e+00, -2.17193764e-01, -2.47428211e-01, ...,
        -1.35201718e-03, -1.10109765e-03,  8.64566393e-05],
       ...,
       [-7.88626804e+00, -1.00353656e+00, -9.09284569e-01, ...,
        -2.15313673e-03,  1.45549207e-03, -4.60053251e-04],
       [-7.86425952e+00, -9.99337996e-01, -8.51569237e-01, ...,
        -2.06885382e-03, -1.22502335e-03,  1.81303381e-03],
       [-7.41622568e+00, -1.41214300e+00, -8.65921210e-01, ...,
        -1.06417476e-03, -1.66377584e-03,  1.78275792e-03]])
```

2-3-b) Get eigen values and eigen vectors

#Check the eigen values

```
array([3.56488638e+01, 7.64357559e+00, 3.76919551e+00, 2.77722349e+00,
       1.90694892e+00, 1.15490310e+00, 9.87726707e-01, 4.64629906e-01,
       3.96708513e-01, 3.22346888e-01, 2.73207369e-01, 2.35647574e-01,
       1.81401107e-01, 1.69243770e-01, 1.38592325e-01, 1.31505852e-01,
       1.03809666e-01, 9.55333831e-02, 8.58580407e-02, 8.09138742e-02,
       6.60179067e-02, 6.30797999e-02, 4.82756124e-02, 4.59506197e-02,
       4.37747566e-02, 3.19339710e-02, 2.86194563e-02, 2.75481445e-02,
       2.34340044e-02, 2.20296816e-02, 1.87487040e-02, 1.59004895e-02,
       1.39957919e-02, 1.18916465e-02, 1.11133495e-02, 9.07842645e-03,
       7.25127869e-03, 6.27213692e-03, 4.95541908e-03, 4.60667097e-03,
       3.45902033e-03, 2.18408510e-03, 2.13514664e-03, 1.92111328e-03,
       1.43840980e-03, 1.09968912e-03, 9.65752052e-04, 8.62630267e-04,
       6.51634478e-04, 5.76658846e-04, 4.35790607e-04, 3.70037468e-04,
       3.06660171e-04, 2.07854170e-04, 1.38286484e-04, 8.97034441e-05,
       4.61745385e-05])
```

#Extract eigen vectors

```
array([[ 0.14922158,  0.15916917,  0.15820921, ...,  0.14136961,
         0.14762899,  0.14210263],
       [-0.11548673, -0.08023879, -0.09371751, ...,  0.03510934,
        -0.04912234, -0.03984815],
       [ 0.1015276 , -0.03866173,  0.0289595 , ..., -0.10217491,
        -0.12667281, -0.02854464],
       ...,
       [ 0.00112879, -0.00673066,  0.02298648, ..., -0.01159627,
         0.05608352, -0.00610478],
       [ 0.00070908,  0.04637872,  0.00402434, ...,  0.01406358,
        -0.07729171, -0.00056173],
       [-0.00461221, -0.00370327,  0.00963954, ...,  0.00227908,
         0.00539901,  0.00130606]])
```

2-3-c)Identify the optimum number of PCs

```
array([6.24441446e-01, 1.33888289e-01, 6.60229147e-02, 4.86470891e-02,
       3.34029704e-02, 2.02297994e-02, 1.73014629e-02, 8.13866529e-03,
       6.94892379e-03, 5.64637229e-03, 4.78562250e-03, 4.12770833e-03,
       3.17750294e-03, 2.96454958e-03, 2.42764517e-03, 2.30351534e-03,
       1.81837655e-03, 1.67340548e-03, 1.50392785e-03, 1.41732362e-03,
       1.15639919e-03, 1.10493400e-03, 8.45617224e-04, 8.04891611e-04,
       7.66778221e-04, 5.59369722e-04, 5.01311201e-04, 4.82545623e-04,
       4.10480504e-04, 3.85881758e-04, 3.28410688e-04, 2.78520087e-04,
       2.45156553e-04, 2.08299401e-04, 1.94666401e-04, 1.59021779e-04,
       1.27016642e-04, 1.09865556e-04, 8.68013375e-05, 8.06925096e-05,
       6.05897475e-05, 3.82574118e-05, 3.74001838e-05, 3.36510796e-05,
       2.51958296e-05, 1.92626466e-05, 1.69165450e-05, 1.51102177e-05,
       1.14143210e-05, 1.01010143e-05, 7.63350323e-06, 6.48174183e-06,
       5.37159674e-06, 3.64086663e-06, 2.42228792e-06, 1.57128566e-06,
       8.08813873e-07])
```

#The percentage of variance explained by each principal component

#Obtaining the Cumulative Sum of the Expalained Variance

```
Cumulative Variance Explained in Percentage: [ 62.44  75.83  82.44  87.3   90.
64  92.66  94.39  95.21  95.9   96.47
  96.95  97.36  97.68  97.97  98.22  98.45  98.63  98.79  98.95  99.09
  99.2   99.31  99.4   99.48  99.55  99.61  99.66  99.71  99.75  99.79
  99.82  99.85  99.87  99.89  99.91  99.93  99.94  99.95  99.96  99.97
  99.98  99.98  99.98  99.99  99.99  99.99  99.99 100.   100.   100.
 100.   100.   100.   100.   100.   100.   100.  ]
```

We can see above that more than 90% of the variance is explained by 5 Principal Components.

optimum number of PCs is 5

2-3-d)-Show Scree plot

 # Plot to identify the number of components to be built

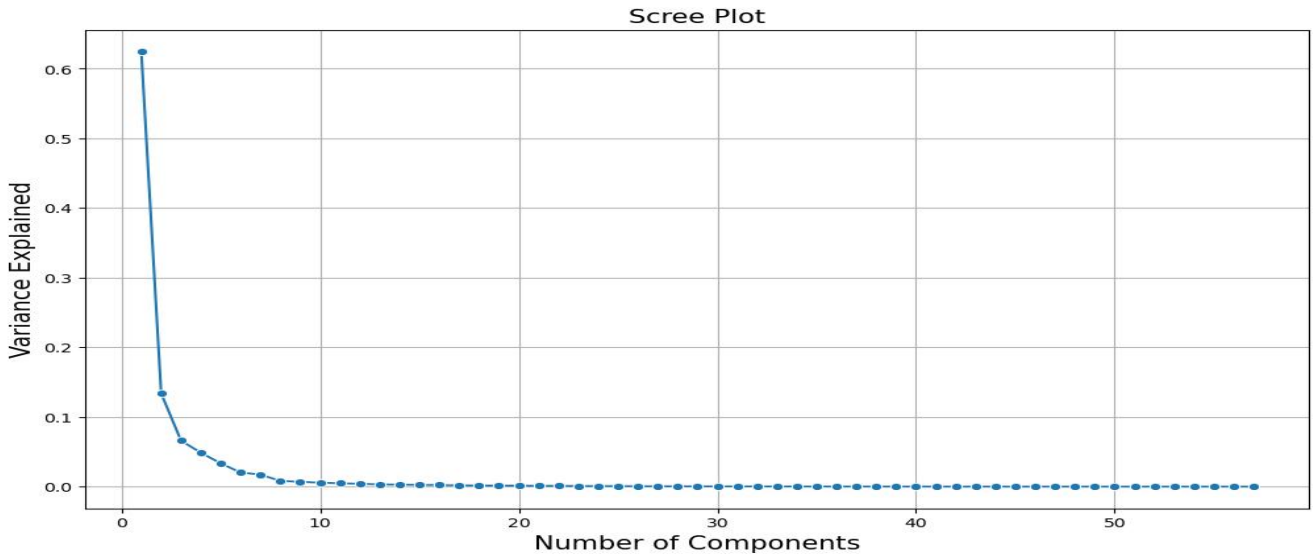


Fig-31- Scree plot

The number of components can be decided based upon the explained variance. Here, it is decided to keep the number of components as 5 as the cumulative explained variance is above 90%

 # find the least number of components that can explain more than 90% variance

```
Number of PCs that explain at least 90% variance:  5
```

Observations:

We can see that out of the 59 original features, we reduced the number of features through principal components to 5, these components explain more than 90% of the original variance. Let us now look at these principal components as a linear combination of original features.

2-3-e)-Compare PCs with Actual Columns and identify which is explaining most variance,Write inferences about all the PCs in terms of actual variables

The number of components can be decided based upon the explained variance. Here, it is decided to keep the number of components as5 as the cumulative explained variance is above 90%

	PC1	PC2	PC3	PC4	PC5
No_HH	0.15	-0.12	0.10	0.08	-0.01
TOT_M	0.16	-0.08	-0.04	0.05	-0.04
TOT_F	0.16	-0.09	0.03	0.07	-0.02
M_06	0.16	-0.02	-0.07	0.03	-0.08
F_06	0.16	-0.01	-0.07	0.02	-0.08
M_SC	0.14	-0.08	-0.04	0.01	-0.17
F_SC	0.14	-0.09	0.02	0.02	-0.16
M_ST	0.02	0.07	0.32	0.09	0.42
F_ST	0.02	0.07	0.34	0.08	0.42
M_LIT	0.16	-0.11	-0.03	0.09	-0.01

	PC1	PC2	PC3	PC4	PC5
F_LIT	0.15	-0.13	-0.01	0.13	0.03
M_ILL	0.15	-0.01	-0.05	-0.03	-0.10
F_ILL	0.16	-0.02	0.08	-0.01	-0.11
TOT_WORK_M	0.15	-0.12	-0.00	0.07	-0.02
TOT_WORK_F	0.14	-0.08	0.19	0.11	-0.02
MAINWORK_M	0.14	-0.17	0.02	0.10	-0.04
MAINWORK_F	0.13	-0.14	0.21	0.13	-0.05
MAIN_CL_M	0.11	0.04	0.03	0.08	-0.30
MAIN_CL_F	0.08	0.10	0.19	0.27	-0.26
MAIN_AL_M	0.12	-0.05	0.23	-0.12	-0.25
MAIN_AL_F	0.09	-0.07	0.36	-0.02	-0.20
MAIN_HH_M	0.14	-0.10	-0.10	-0.02	-0.06

	PC1	PC2	PC3	PC4	PC5
MAIN_HH_F	0.13	-0.11	0.02	-0.05	-0.02
MAIN_OT_M	0.12	-0.20	-0.03	0.15	0.07
MAIN_OT_F	0.12	-0.21	0.07	0.16	0.11
MARGWORK_M	0.16	0.08	-0.07	-0.08	0.07
MARGWORK_F	0.15	0.11	0.10	0.02	0.08
MARG_CL_M	0.09	0.27	-0.10	0.16	-0.02
MARG_CL_F	0.07	0.28	-0.04	0.29	-0.06
MARG_AL_M	0.13	0.16	0.07	-0.25	-0.05
MARG_AL_F	0.12	0.14	0.26	-0.15	-0.01
MARG_HH_M	0.15	0.04	-0.14	-0.17	0.01
MARG_HH_F	0.14	0.01	-0.09	-0.15	0.04
MARG_OT_M	0.15	-0.07	-0.13	0.02	0.15

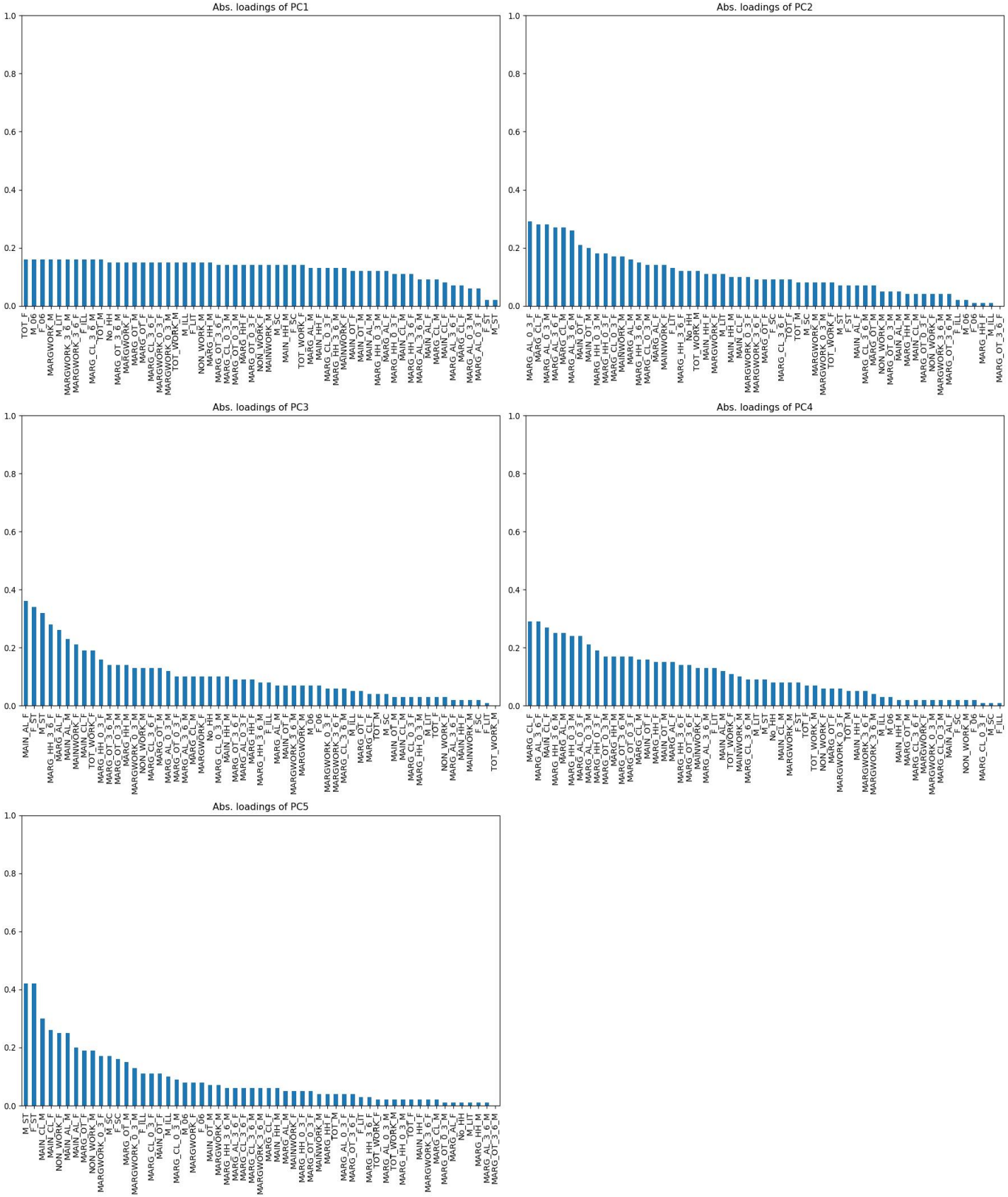
	PC1	PC2	PC3	PC4	PC5
MARG_OT_F	0.15	-0.09	-0.05	0.06	0.19
MARGWORK_3_6_M	0.16	-0.04	-0.07	0.04	-0.06
MARGWORK_3_6_F	0.16	-0.09	-0.06	0.05	-0.02
MARG_CL_3_6_M	0.16	0.07	-0.06	-0.09	0.06
MARG_CL_3_6_F	0.15	0.09	0.13	0.02	0.06
MARG_AL_3_6_M	0.09	0.26	-0.10	0.13	-0.01
MARG_AL_3_6_F	0.07	0.27	-0.02	0.29	-0.06
MARG_HH_3_6_M	0.13	0.15	0.08	-0.25	-0.06
MARG_HH_3_6_F	0.11	0.12	0.28	-0.14	-0.03
MARG_OT_3_6_M	0.15	0.04	-0.14	-0.17	0.00
MARG_OT_3_6_F	0.14	-0.00	-0.09	-0.14	0.04
MARGWORK_0_3_M	0.15	-0.08	-0.13	0.02	0.13

	PC1	PC2	PC3	PC4	PC5
MARGWORK_0_3_F	0.15	-0.10	-0.06	0.06	0.17
MARG_CL_0_3_M	0.14	0.14	-0.10	-0.02	0.09
MARG_CL_0_3_F	0.13	0.17	0.03	0.01	0.11
MARG_AL_0_3_M	0.06	0.28	-0.12	0.21	-0.02
MARG_AL_0_3_F	0.06	0.29	-0.09	0.24	-0.04
MARG_HH_0_3_M	0.12	0.18	0.03	-0.24	0.02
MARG_HH_0_3_F	0.11	0.18	0.16	-0.19	0.05
MARG_OT_0_3_M	0.14	0.05	-0.14	-0.17	0.01
MARG_OT_0_3_F	0.14	0.04	-0.10	-0.17	0.05
NON_WORK_M	0.15	-0.05	-0.13	0.02	0.19
NON_WORK_F	0.14	-0.04	-0.03	0.06	0.25

Table-9-)-Compare PCs with Actual Columns and identify which is explaining most variance

#Check as to how the original features matter to each PC

#Note: Here we are only considering the absolute values


Abs. loadings of PC1
Abs. loadings of PC2
Abs. loadings of PC3
Abs. loadings of PC4
Abs. loadings of PC5

¶

Fig-32-Compare PCs with Actual Columns

	PC1	PC2	PC3	PC4	PC5
No_HH	0.150000	-0.120000	0.100000	0.080000	-0.010000
TOT_M	0.160000	-0.080000	-0.040000	0.050000	-0.040000
TOT_F	0.160000	-0.090000	0.030000	0.070000	-0.020000
M_06	0.160000	-0.020000	-0.070000	0.030000	-0.080000
F_06	0.160000	-0.010000	-0.070000	0.020000	-0.080000
M_SC	0.140000	-0.080000	-0.040000	0.010000	-0.170000
F_SC	0.140000	-0.090000	0.020000	0.020000	-0.160000
M_ST	0.020000	0.070000	0.320000	0.090000	0.420000
F_ST	0.020000	0.070000	0.340000	0.080000	0.420000
M_LIT	0.160000	-0.110000	-0.030000	0.090000	-0.010000
F_LIT	0.150000	-0.130000	-0.010000	0.130000	0.030000
M_ILL	0.150000	-0.010000	-0.050000	-0.030000	-0.100000

	PC1	PC2	PC3	PC4	PC5
F_ILL	0.160000	-0.020000	0.080000	-0.010000	-0.110000
TOT_WORK_M	0.150000	-0.120000	-0.000000	0.070000	-0.020000
TOT_WORK_F	0.140000	-0.080000	0.190000	0.110000	-0.020000
MAINWORK_M	0.140000	-0.170000	0.020000	0.100000	-0.040000
MAINWORK_F	0.130000	-0.140000	0.210000	0.130000	-0.050000
MAIN_CL_M	0.110000	0.040000	0.030000	0.080000	-0.300000
MAIN_CL_F	0.080000	0.100000	0.190000	0.270000	-0.260000
MAIN_AL_M	0.120000	-0.050000	0.230000	-0.120000	-0.250000
MAIN_AL_F	0.090000	-0.070000	0.360000	-0.020000	-0.200000
MAIN_HH_M	0.140000	-0.100000	-0.100000	-0.020000	-0.060000
MAIN_HH_F	0.130000	-0.110000	0.020000	-0.050000	-0.020000
MAIN_OT_M	0.120000	-0.200000	-0.030000	0.150000	0.070000

	PC1	PC2	PC3	PC4	PC5
MAIN_OT_F	0.120000	-0.210000	0.070000	0.160000	0.110000
MARGWORK_M	0.160000	0.080000	-0.070000	-0.080000	0.070000
MARGWORK_F	0.150000	0.110000	0.100000	0.020000	0.080000
MARG_CL_M	0.090000	0.270000	-0.100000	0.160000	-0.020000
MARG_CL_F	0.070000	0.280000	-0.040000	0.290000	-0.060000
MARG_AL_M	0.130000	0.160000	0.070000	-0.250000	-0.050000
MARG_AL_F	0.120000	0.140000	0.260000	-0.150000	-0.010000
MARG_HH_M	0.150000	0.040000	-0.140000	-0.170000	0.010000
MARG_HH_F	0.140000	0.010000	-0.090000	-0.150000	0.040000
MARG_OT_M	0.150000	-0.070000	-0.130000	0.020000	0.150000
MARG_OT_F	0.150000	-0.090000	-0.050000	0.060000	0.190000
MARGWORK_3_6_M	0.160000	-0.040000	-0.070000	0.040000	-0.060000

	PC1	PC2	PC3	PC4	PC5
MARGWORK_3_6_F	0.160000	-0.090000	-0.060000	0.050000	-0.020000
MARG_CL_3_6_M	0.160000	0.070000	-0.060000	-0.090000	0.060000
MARG_CL_3_6_F	0.150000	0.090000	0.130000	0.020000	0.060000
MARG_AL_3_6_M	0.090000	0.260000	-0.100000	0.130000	-0.010000
MARG_AL_3_6_F	0.070000	0.270000	-0.020000	0.290000	-0.060000
MARG_HH_3_6_M	0.130000	0.150000	0.080000	-0.250000	-0.060000
MARG_HH_3_6_F	0.110000	0.120000	0.280000	-0.140000	-0.030000
MARG_OT_3_6_M	0.150000	0.040000	-0.140000	-0.170000	0.000000
MARG_OT_3_6_F	0.140000	-0.000000	-0.090000	-0.140000	0.040000
MARGWORK_0_3_M	0.150000	-0.080000	-0.130000	0.020000	0.130000
MARGWORK_0_3_F	0.150000	-0.100000	-0.060000	0.060000	0.170000
MARG_CL_0_3_M	0.140000	0.140000	-0.100000	-0.020000	0.090000

	PC1	PC2	PC3	PC4	PC5
MARG_CL_0_3_F	0.130000	0.170000	0.030000	0.010000	0.110000
MARG_AL_0_3_M	0.060000	0.280000	-0.120000	0.210000	-0.020000
MARG_AL_0_3_F	0.060000	0.290000	-0.090000	0.240000	-0.040000
MARG_HH_0_3_M	0.120000	0.180000	0.030000	-0.240000	0.020000
MARG_HH_0_3_F	0.110000	0.180000	0.160000	-0.190000	0.050000
MARG_OT_0_3_M	0.140000	0.050000	-0.140000	-0.170000	0.010000
MARG_OT_0_3_F	0.140000	0.040000	-0.100000	-0.170000	0.050000
NON_WORK_M	0.150000	-0.050000	-0.130000	0.020000	0.190000
NON_WORK_F	0.140000	-0.040000	-0.030000	0.060000	0.250000

Table-10-Compare PCs with Actual Columns

OBSERVATION

PC1 is explaining most variance.

- The first principal component, PC1, is a measure of mpg, cylinders, displacement, horsepower, and weight. PC1 is associated with high scores of almost all variables('No_HH', 'TOT_M', 'TOT_F', 'M_06', 'F_06', 'M_SC', 'F_SC','M_LIT', 'F_LIT', 'M_ILL', 'F_ILL', 'TOT_WORK_M', ',TOT_WORK_F"MAINWORK_M', 'MAINWORK_F', 'MAIN_CL_M ,'MAIN_AL_M', 'MAIN_HH_M', 'MAIN_OT_M', 'MAIN_OT_F', 'MARGWORK_M', 'MARGWORK_F', 'MARG_AL_M', 'MARG_AL_F', 'MARG_HH_M', 'MARG_HH_F', 'MARG_OT_M', 'MARG_OT_F', 'MARGWORK_3_6_M', 'MARGWORK_3_6_F',

'MARG_CL_3_6_M', 'MARG_CL_3_6_F', 'MARG_HH_3_6_M', 'MARG_HH_3_6_F', 'MARG_OT_3_6_M', 'MARG_OT_3_6_F', 'MARGWORK_0_3_M', 'MARGWORK_0_3_F', 'MARG_CL_0_3_M', 'MARG_CL_0_3_F', 'MARG_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' M_ST,F_ST no negative scores in PC1.

- PC2 have less explain variance than PC2,The second principal component, PC2, is a measure of model year. PC2 is associated with low values of "No_HH','TOT_F','M_LIT', 'F_LIT','TOT_WORK_M','MAINWORK_M', 'MAINWORK_F','MAIN_OT_M', 'MAIN_OT_F','MARG_OT_F', 'MARGWORK_3_6_F','MARGWORK_0_3_F'

optimum number of PCs is 5

2-3-f)Write linear equation for first PC Note

PC1 have explain all variance and all are high

To write the linear equation for PC1 (Principal Component 1), you can express PC1 as a linear combination of the original variables. Let's denote the original variables as (X_1, X_2, X_3, ..., X_n) and their corresponding coefficients in PC1 as (a_1, a_2, a_3,....., a_n). The linear equation for PC1 can be written as:

PC1=a1x1a2x2.....anxn.

Using the coefficients provided in your data, the linear equation for PC1 would be

This equation represents the linear combination of the original variables that make up PC1. The coefficients indicate the contribution of each variable to the overall value of PC1.

PC1=a1x1*a2x2*.....anxn.

0.16*No_HH*+ *0.17*TOT_M+ 0.17*TOT_F*+ *0.16*M_06+ 0.16*F_06*+ *0.15*M_SC+ 0.15*F_SC*+ *0.03*M_ST+ 0.03*F_ST*+ *0.16*M_LIT+ 0.15*F_LIT*+ *0.16*M_ILL+ 0.17*F_ILL* + *0.16*TOT_WORK_M + 0.15*TOT_WORK_F*+
*0.15*MAINWORK_M+
0.12*MAINWORK_F*+
*0.10*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* +
*0.16*MARG_CL_M +
0.05*MARG_CL_F* +
*0.13*MARG_AL_M +
0.11*MARG_AL_F* +
*0.14*MARG_HH_M +
0.13*MARG_HH_F* +
*0.16*MARG_OT_M +
0.15*MARG_OT_F* +

*0.16*MARGWORK_3_6_M + 0.16*MARGWORK_3_6_F* +
*0.17*MARG_CL_3_6_M + 0.16*MARG_CL_3_6_F* +
*0.09*MARG_AL_3_6_M +
0.05*MARG_AL_3_6_F* +
*0.13*MARG_HH_3_6_M +
0.13*MARG_HH_3_6_F* +
*0.14*MARG_OT_3_6_M + 0.12*MARG_OT_3_6_F* + *0.15*MARGWORK_0_3_M +
0.15*MARGWORK_0_3_F* + *0.15*MARG_CL_0_3_M + 0.14*MARG_CL_0_3_F* +
*0.05*MARG_AL_0_3_M + 0.04*MARG_AL_0_3_F* + *0.12*MARG_HH_0_3_M +
0.12*MARG_HH_0_3_F* + *0.14*MARG_OT_0_3_M + 0.13*MARG_OT_0_3_F* + *0.15*NON_WORK_M
+ 0.13*NON_WORK_F