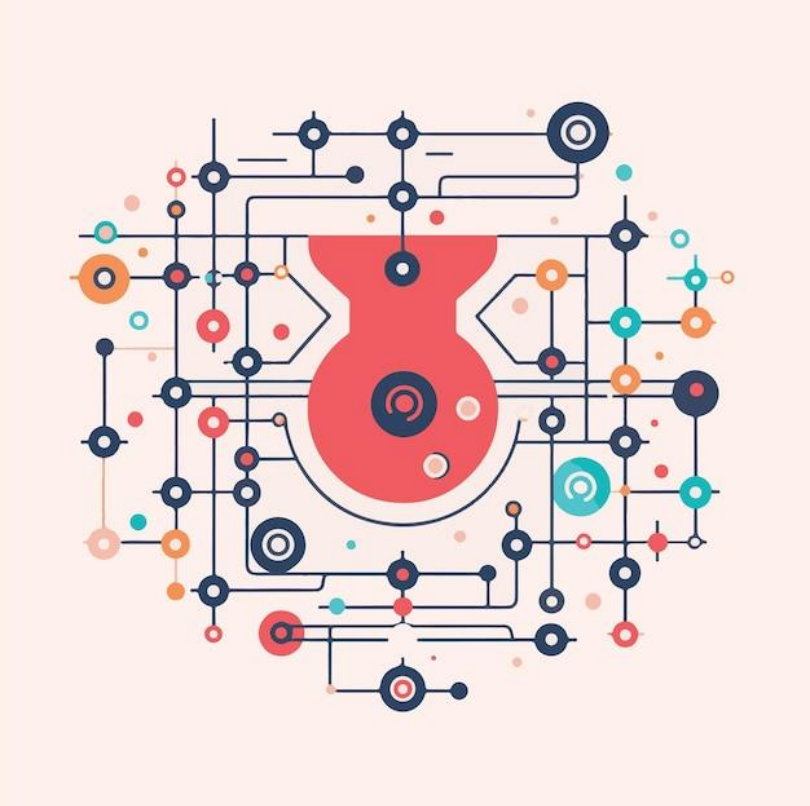

MACHINE LEARNING - 1 PROJECT

By Kurt Warren Mario Gilby On May 12th 2024

Submitted to

Great Learning
POWER AHEAD

As a part of the requirements for completion of PGP-DSBA offered in affiliation with

TEXAS McCombs
The University of Texas at Austin
McCombs School of Business

Table of Contents

List of Figures

List of Tables

List of Equations

Introduction

While doing the **"Machine Learning -1 course"**, we have seen and practised some tools and techniques that fall under the unsupervised learning criteria, which means that we look at all the independent features, and try and find some meaning.

We do this by generating and selecting the most important features which explain most of the variance in the data "PCA", Or/And find similar observations/records and group them together, label them based on their similarities by the use of "Clustering" techniques.

I will in the following pages use the said tools and techniques to review the problem sets given and answer the questions posed. In the following pages of this document, I have given the following sections:

- **Overview**: High level overview of the problem statement/case study
- **Dataset Description**: Definition of the dataset provided
- **Objective**: detail list of the steps that will be taken when answering the questions asked.
- **Questions Asked**: List of all the questions asked with their answers and supporting material like Figures Tables etc.

The first problem set here showcases two Clustering techniques "Hierarchical Clustering" and "K-means Clustering". While the second problem set looks at the dataset given and showcases the Principal Component Analysis or "PCA" technique.

Along with using and showcasing the above I will also showcase statistical techniques and good practices learnt in pervious courses, such as Exploratory Data Analysis (EDA), and Data Preprocessing.

Problem 1

Overview

A Digital Marketing company **"ads24x7"** would like to **segment ads into homogeneous groups**, Groups having similar features.

To do this I will use the data provided by the **"Marketing Intelligence"** team of the company, and use **Clustering procedures** to create the groups and try and **find actionable insights and recommendations** for the input provided.

Objective

Using the data provided will perform the following steps:

1. Define the problem
2. Explore the data
3. Get the statistical summary of the data.
4. Perform data preprocessing
5. Perform Hierarchical Clustering
6. Perform K-means Clustering
7. Do Cluster Profiling
8. Derive Actionable Insights and Recommendations

Dataset Description

This is the Definition of the data provided by the "Marketing Intelligence" team. The features commonly used in digital marketing are three, namely CTR, CPM and CPC the definition is in the below table:

Sl. No	Column Name	Column Description
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 Advertisement.
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 particular 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 counts 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.

Table 1 Ads Data Definitions

Questions Asked

1.1 Define the problem and perform Exploratory Data Analysis.

1.1.1 Problem definition

Using the dataset given for ads campaigns run we need to find campaigns that are similar to one and other and derive actionable Insights and recommendations, for the said groups of campaigns.

1.1.2 Checking the Data and doing Exploratory Data Analysis on the same.

The Ads Data set has **23066** Observations(rows) and **19** Features(columns). **6** out of the **19** features are of type string these are:

- Timestamp
- InventoryType
- Ad Type
- Platform
- Device Type
- Format

The remaining **13** features at of type numbers these are:

- Ad – Length
- Ad -Width
- Ad Size
- Available_Impressions
- Matched_Queries
- Impressions
- Clicks
- Spend
- Fee
- Revenue
- CTR
- CPM
- CPC

1.1.2.1 Key Observations and Treatments of the data

The datatype for **Timestamp** needed to be corrected to datetime and used as categorical variable, the datatypes **of InventoryType, Platform, Device Type and Format** too are to be used and set to categorical variables, upon checking we see that there are Null or missing values in the fields of **CTR, CPM and CPC**, these can be filled in using the formulars given for these:

$$CTR = \left(\frac{Clicks}{Impressions}\right) * 100$$

Equation 1 Click Through Rate

$$CPM = \left(\frac{Spend}{Impressions}\right) * 1000$$

Equation 2 Cost Per 1000 Impressions

$$CPC = \left(\frac{Spend}{Clicks}\right)$$

Equation 3 Cost Per Click

!!Please Note: "Since the data is being used for Clustering, we will ignore the Categorical values for now, ideally, we would treat these too using one hot encoding or numeric label substitution, but for this exercise we will not be doing this. This will hold good for all the activities we do for Problem 1 in this exercise."

Reviewing the data and doing a describe on the dataset we see that:

- The values for the columns in the data are not in the same scale.
- There are potential outliers present in the dataset, as for some features the max values are far from the 75-percentile value and the median.
- There are missing values for CTR, CPM and CPC.

	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	23066.0	1241519.52	2429399.96	1.00	7990.50	225290.00	1112428.50	14194774.00
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

Figure 1Describe of the Ads Data using python pandas.

1.1.2.2 Exploratory Data Analysis (Univariate and Bivariate)

We perform **"Univariate"** Analysis on the all the numerical variables by plotting boxplots and histograms of these variables and these are the findings.

Boxplots observations:

- **Ad – Length** and **Ad- Width** do not have outliers
- The rest [***'Ad Size' 'Available_Impressions' 'Matched_Queries' 'Impressions' 'Clicks' 'Spend' 'Fee' 'Revenue' 'CTR' 'CPM' 'CPC'***] have outliers.
- ***!!Please Note: "Even though outliers are available I would want to keep them as removing the same would affect the classification in Clustering, and these outliers are not due to data errors and are the actual measurement/input of the variable"***
- *if I was looking to treat these outliers, I would have used the method of finding the values that were* ***+- 1.5 IQR*** *distance and set the value for them as* ***the 5 percentiles and 95 percentiles*** *respectively,* ***but I'm keeping the outliers without treating them here.***

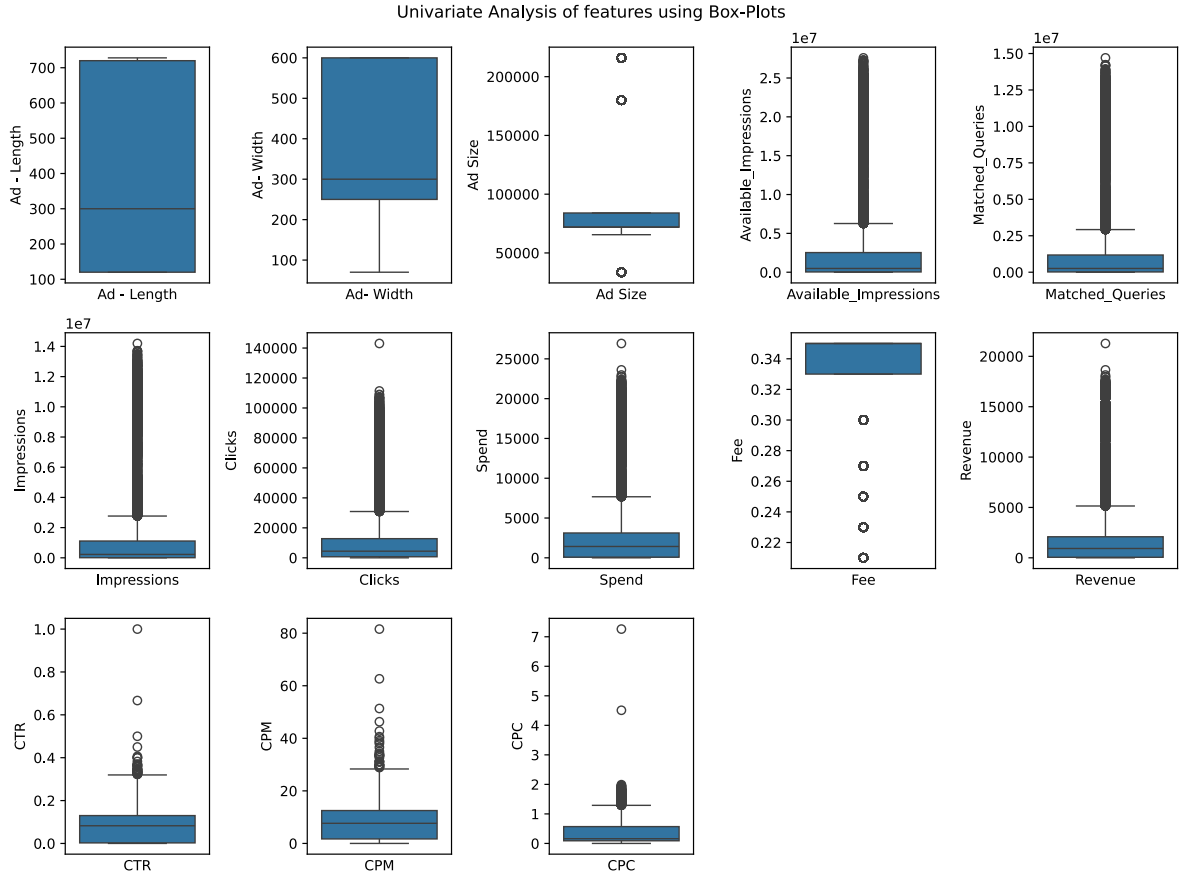

Figure 2 Ads data numerical variables Boxplots

Histograms Observations:

- **Ad - Length, Ad – Width, Ad Size and Fee** have a few discreet values
- Rest of the features follow a bell shape distribution will are being right skewed.
- **CTR, CPM and CPC** are bi-nodal and right skewed.

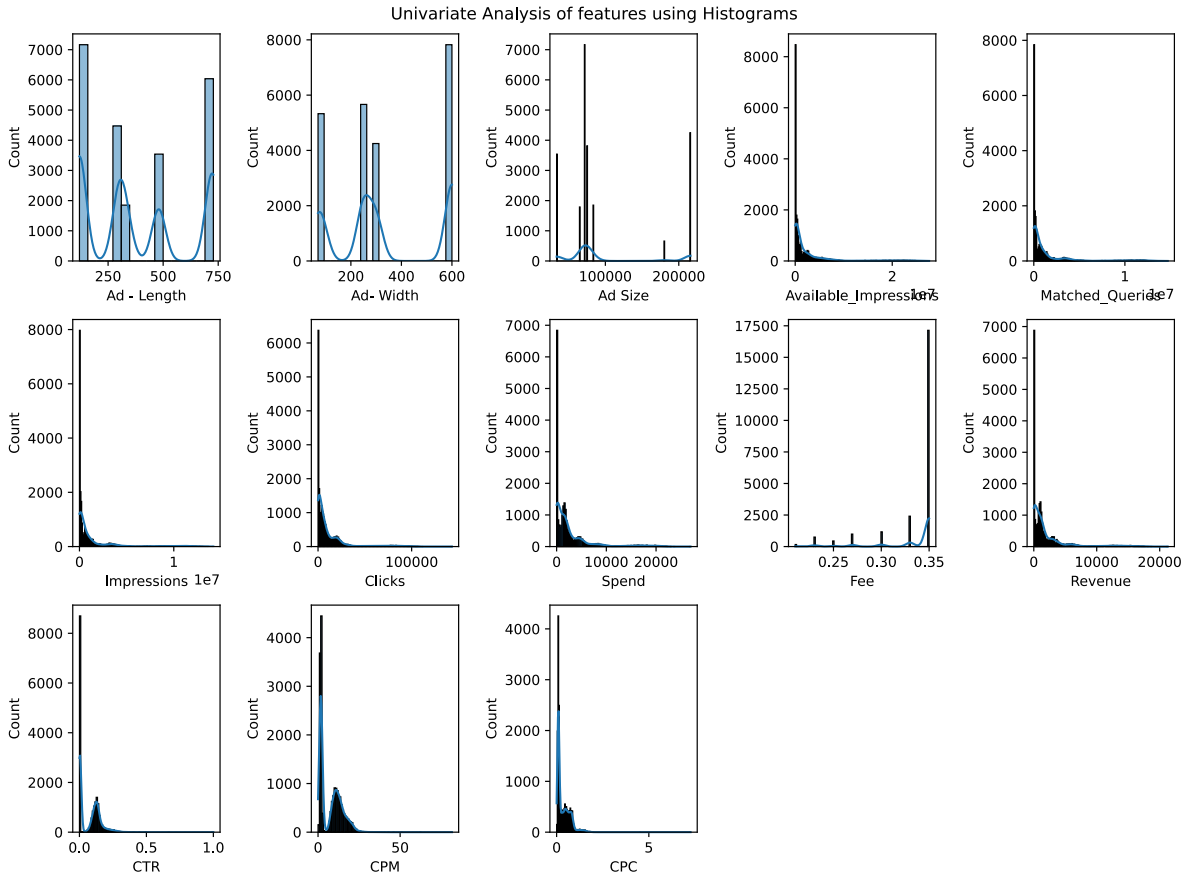

Figure 3 Ads Data numerical variables Histogram Plots

Since we see that there is a **skew** on most of the numerical variables lets, calculate the skew and plot the same in a graph.

- We see that most of the variables have a very high positive skew, and Fee has a very high negative skew.
- ***!!Please Note: "Ideally in these cases, we would look to transform the numerical variables using transformations such as log, square-root, cube-root, box-cox transformations etc.., for this exercise I'm not transforming the data and taking the same as given."***

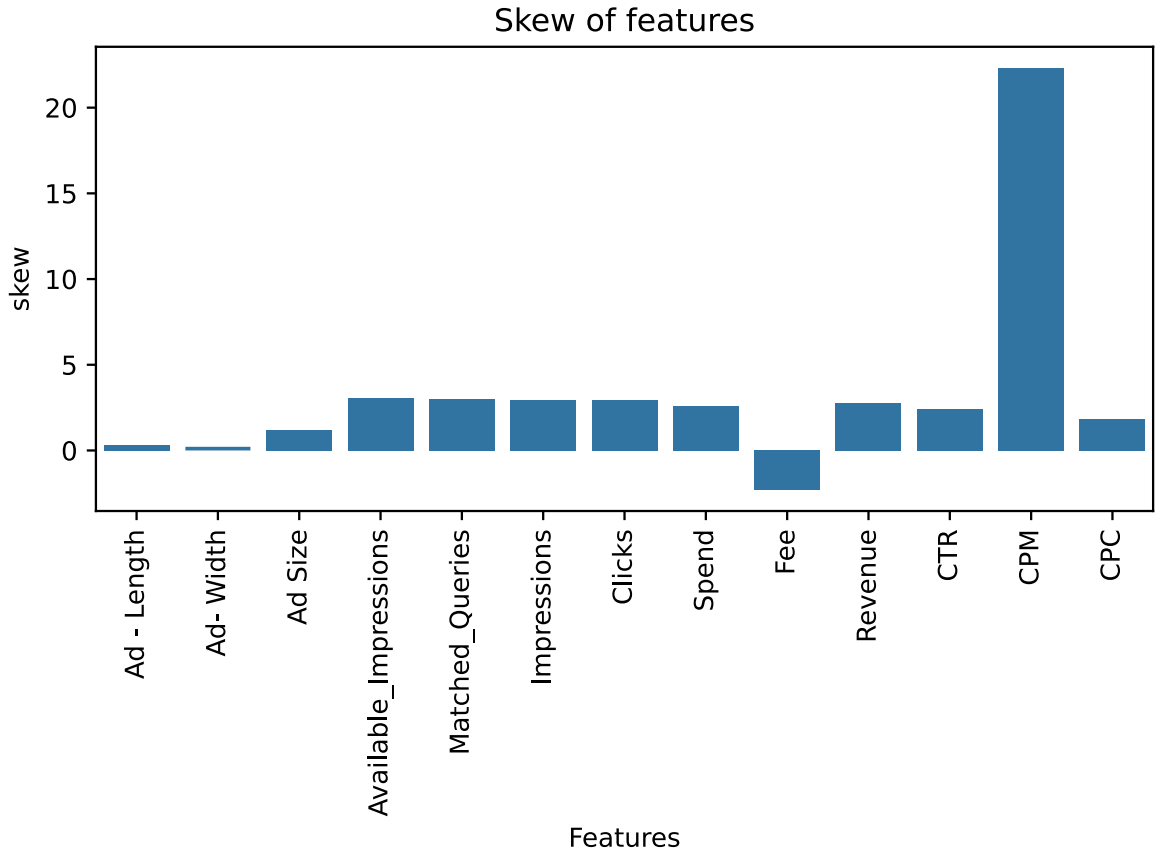

Figure 4 Ads data numerical variables Skew values plotted

Next, we do **Bivariate Analysis** using a **pair plot** and **heatmap** for all the numerical variables.

Pair plot Observations:

- **Available_impressions** seem to be correlated **with Matched_Queries, Impressions, Clicks, Spend, Revenue.**
- **Matched_Queries** seems to be correlated with **Available_impressions, Impressions, Clicks, Spend, Revenue**
- **Impressions** seems to correlated with **Available_impressions, Matched_Queries, Clicks, Spend, Revenue**
- **Clicks** seems to correlated with **Available_impressions, Matched_Queries, Impressions, Spend, Revenue**
- **Spend** seems to correlated with **Available_impressions, Matched_Queries, Impressions, Clicks, Revenue**
- **Revenue** seems to correlated with **Available_impressions, Matched_Queries, Impressions, Clicks, Spend**
- **Fee** seems to have a negative correlation with **all the above-mentioned attributes**.

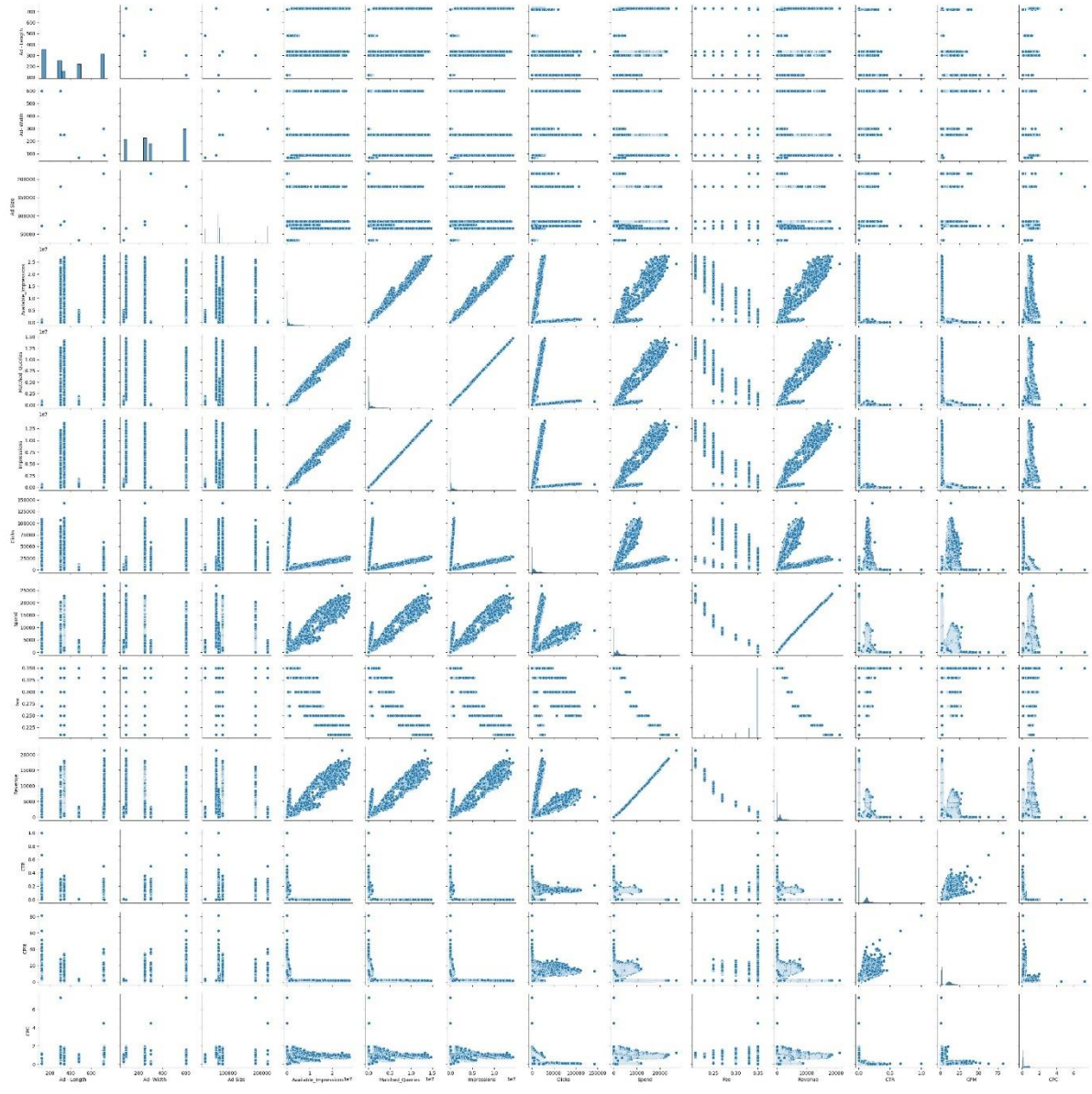

Figure 5 Ads Data numerical variables Pair Plot

Heatmap Observations:

- The heat map shows the same inferences as the pair plot.
- ***!!Please Note: "For cases where columns are correlated, we would look to make sure to reduce all the correlations in the data, by using techniques like PCA that would give us features which are less correlated to each other, but this is at the cost of explainability of the model and also would make it harder to do cluster profiling, so in this case we will not do this."***

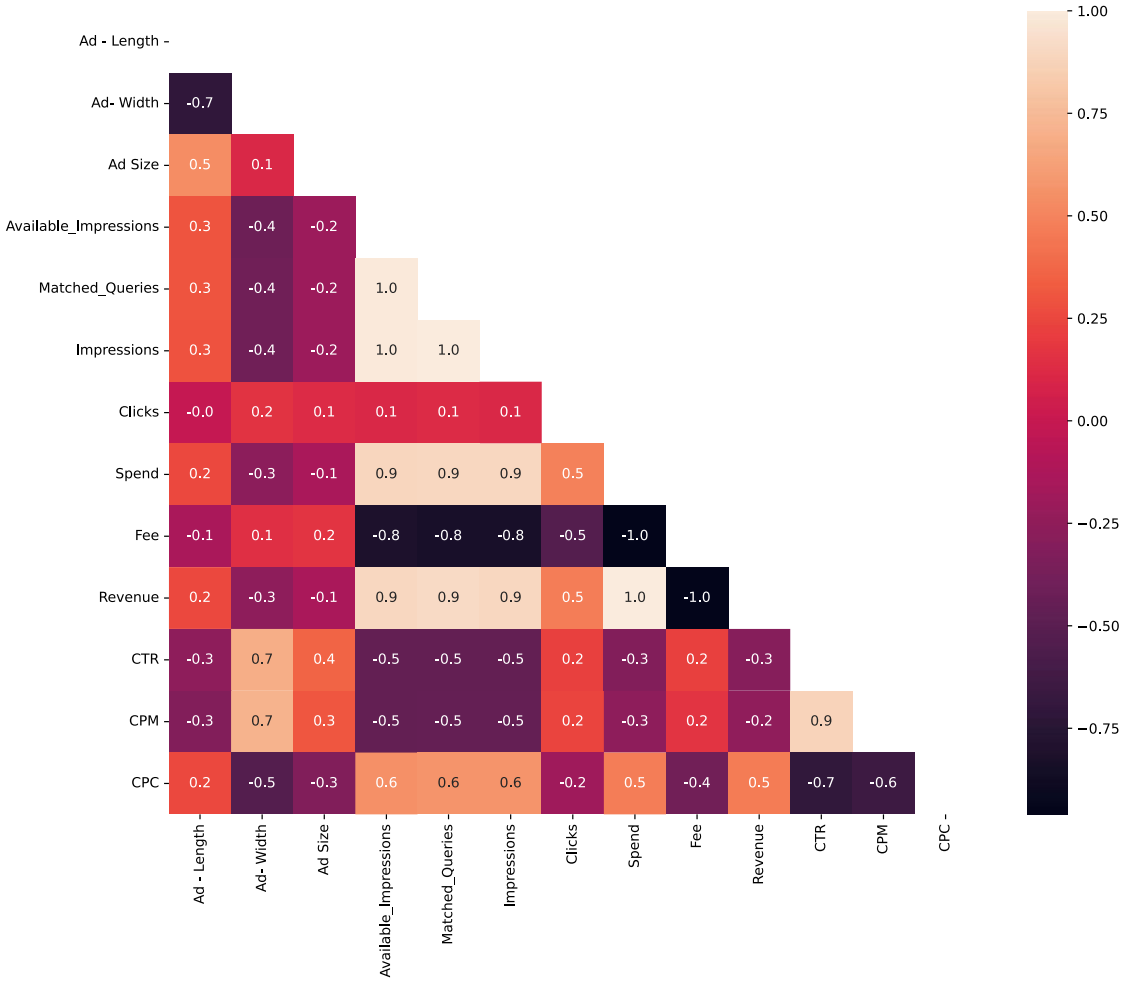

Figure 6 Ads Data numerical variables Heatmap

1.1.3 Data Preprocessing

In the Data Preprocess steps we will do the following:

- We will focus only on the numerical columns as stated above, for this clustering exercise we will not be considering the categorical columns.
- We check for duplicates, this step is done as soon as the data in imported, but called out here as it is a part of data preprocessing steps
- We see that there are Null or missing values in the fields of **CTR, CPM and CPC**, these can be filled in using the formulars given for these:
  - $CTR = \left(\frac{Clicks}{Impressions}\right) * 100$
  - $CPM = \left(\frac{Spend}{Impressions}\right) * 1000$
  - $CPC = \left(\frac{Spend}{Clicks}\right)$
- Outlier treatment:
  - ***!!Please Note: "Even though outliers are available I would want to keep them as removing the same would affect the classification in Clustering, and these outliers are not due to data errors and are the actual measurement/input of the variable"***
- Scaling:
  - The numerical variables are on different scales, e.g. **Ad Size, Clicks, Spend, Revenue**
  - I have used Z score scaling method here which replace all the values in a variable/column of the data using the formula.

Equation 4 Z-Score formula.

- $Z_i = \frac{(x_i - \mu)}{\sigma}$

1.1.4 Identification of Features to use for Clustering:

- We will not use any of the categorical features: **exclude: Timestamp, InventoryType, Ad Type, Platform, Device Type, and Format.** Reason: Categorical Columns will need separate treatment like one hot key encoding or K-Modes for example, which we will not explore in this exercise.
- Ad-Size is made up as a calculated field of Ad - Length and Ad- Width: **exclude: Ad- Length, Ad-Width.** Reason: Both Length and Width is capture in size.
- Avalable_Impressions and Matched_Queries are highly correlated "0.99": **exclude: Matched_Queries** Reason: Matched_Queries is not something the Ad agency can directly control.
- Avalable_Impressions and Impressions are highly correlated "0.99": **exclude: Avalable_Impressions** Reason: Impression is a better gauge of how a particular Advertisement is doing, as compared to the overall of available impressions.
- **Spend and Revenue** are highly correlated "1": But we will **keep both** check in case we have less spend high revenue clusters.
- **Final Columns used for Clustering: Ad Size, Impressions, Clicks, Spend, Fee, Revenue, CTR, CPM, CPC**

1.2 Hierarchical Clustering

1.2.1 Construct a dendrogram using ward linkage and Euclidean distance

The first step of doing a **"Hierarchical Clustering"** is to construct a **dendrogram.**

Hierarchical Clustering: An Overview:

The initial step in performing Hierarchical Clustering involves the construction of a dendrogram. This process begins by examining the data and presuming that each observation is an individual cluster.

Subsequently, we employ the Ward linkage method to identify the nearest observations and group or cluster them together. This process is repeated until we are left with a single cluster. This iterative procedure can be conveniently executed using the scipy.cluster.hierarchy package in Python.

The output of this process is a dendrogram, which provides a visual representation of how the observations are grouped at varying Euclidean distances. By studying the dendrogram, we can determine an appropriate cut-off point to obtain the optimal number of clusters.

Identify optimal number of Clusters.

Looking at the dendrogram it looks like 3 or 4 clusters should be optimal, but instead of relying on this we will calculate the Silhouette Score at various values of clusters to check what would be optimal. Doing this we see that it suggests that 2 or 5 clusters would be optimal with 2 having more clear boundaries between the clusters than 5, but 5 is also slightly higher than .5 and closer to 1 so would be good to investigate

Plotting the Dendrogram:

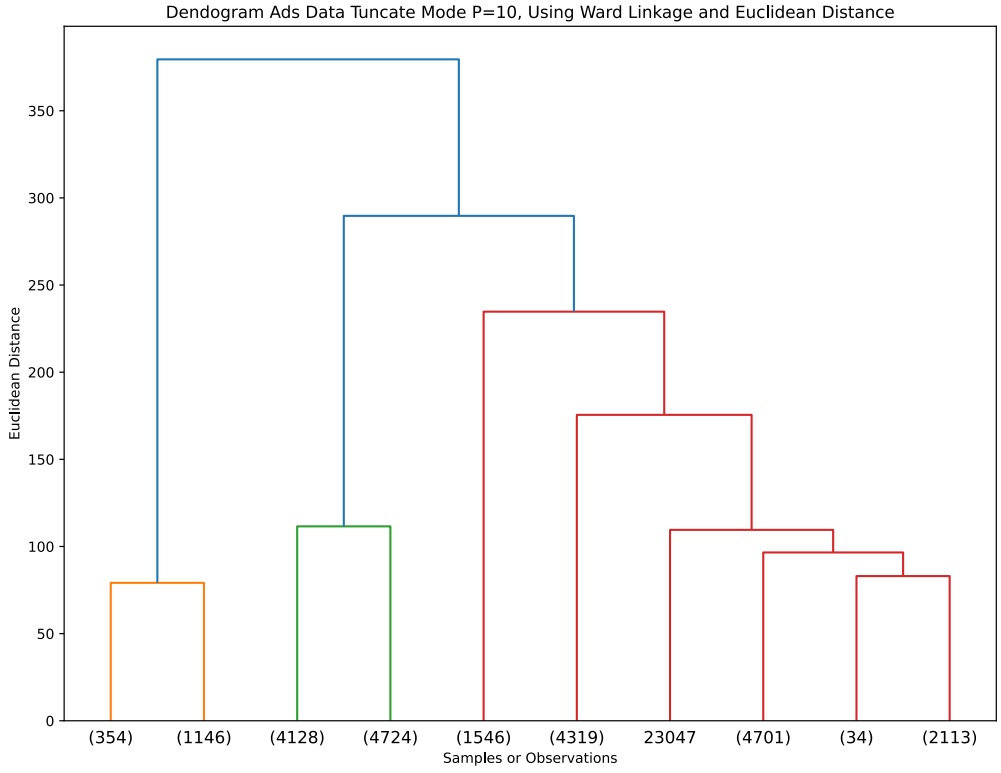

Figure 7 Ads Data Only Clustering numerical variables Dendrogram

Plotting the Silhouette Score chart for different values of K for Hierarchical Clustering

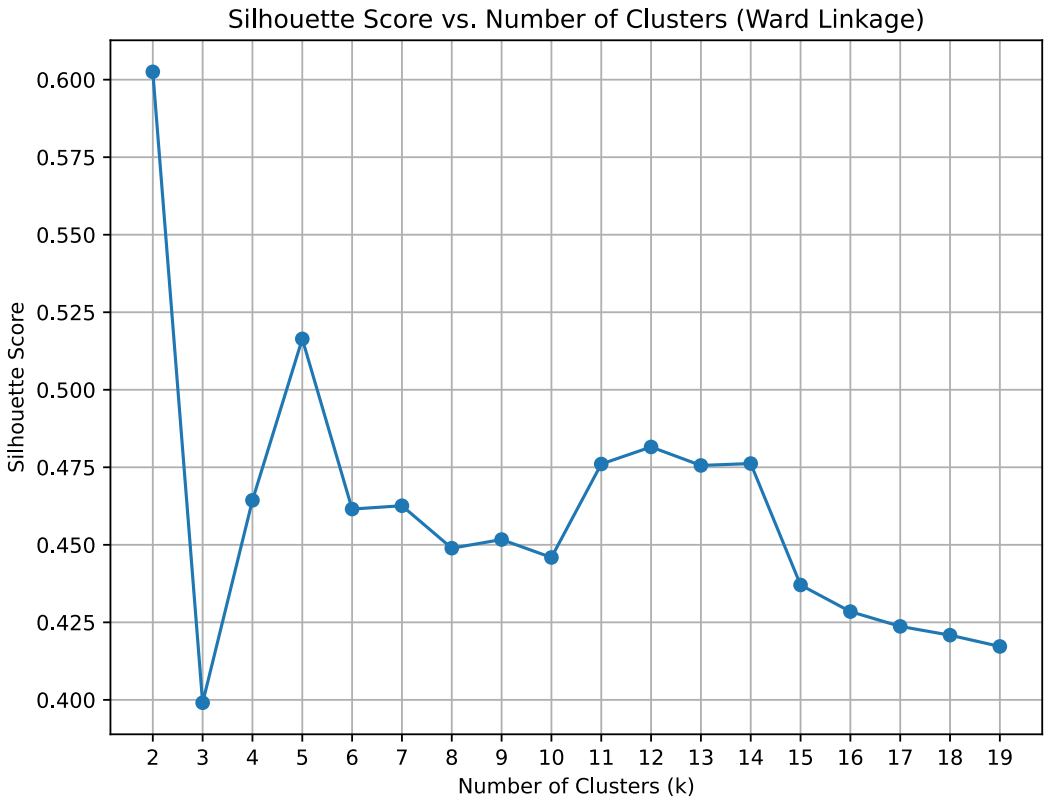

Figure 8 Silhouette Score chart for different values of K for Hierarchical Clustering

1.3 K-means Clustering

K-Means Clustering: An Overview

The initial step in performing K-Means Clustering is to determine the number of clusters, denoted as 'k'. This is often done using methods such as the Elbow Method or the Silhouette Method.

Once 'k' is determined, the algorithm randomly assigns each observation to a cluster. The centroid (the arithmetic mean) of each cluster is then computed. The observations are then reassigned to the cluster whose centroid is closest, and the centroids are recalculated. This process is repeated until the assignments no longer change.

This iterative procedure can be conveniently executed using the **sklearn.cluster.KMeans** package in Python.

The output of this process is a set of 'k' cluster centroids and a labelling of the data that assigns each observation to the nearest centroid.

Identifying Optimal Number of Clusters

To identify the optimal number of clusters, we can use the Elbow Method, which involves plotting the explained variation as a function of the number of clusters, and picking the elbow of the curve as the number of clusters to use.

Alternatively, we can calculate the Silhouette Score at various values of 'k' to check what would be optimal. The Silhouette Score is a measure of how similar an object is to its own cluster compared to other clusters. A higher Silhouette Score indicates that the object is well matched to its own cluster and poorly matched to neighbouring clusters.

If the Silhouette Score for 'k' clusters is above 0.5 and closer to 1, it suggests that the clustering configuration is appropriate. However, it's also important to consider the interpretability and the practical use of the clusters. Therefore, even if the Silhouette Score is slightly lower for a different 'k', it might be worth considering if the clusters are more interpretable and useful for the task at hand.

Figure out the appropriate number of clusters:

We see that Silhouette Scores suggests that 2, 10, 11, 12, 13 clusters would be optimal with 2 having more clear boundaries between the clusters than the rest. and the Elbow method suggests we look at k = 6. So, we will investigate k=2 and k=6 as the appropriate number of clusters to investigate.

Plotting the Elbow curve:

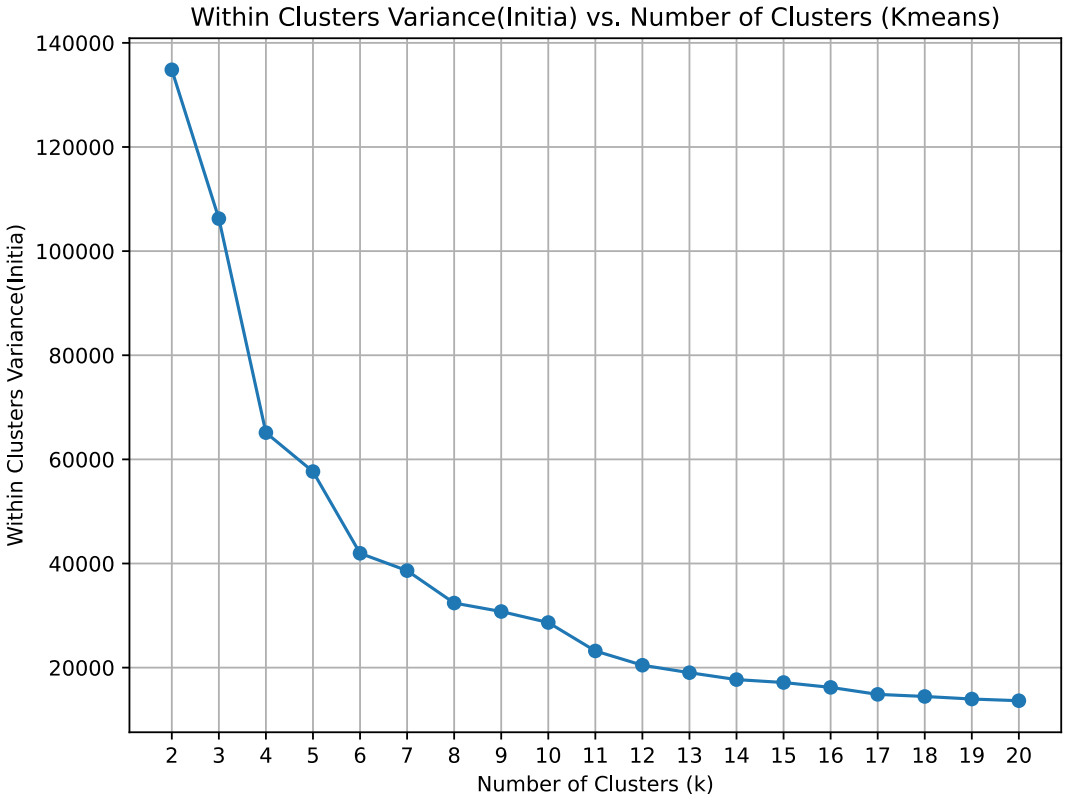

Figure 9 Elbow plot for the K-Means Clustering

Plotting the Silhouette Score chart for different values of K for K-Means Clustering

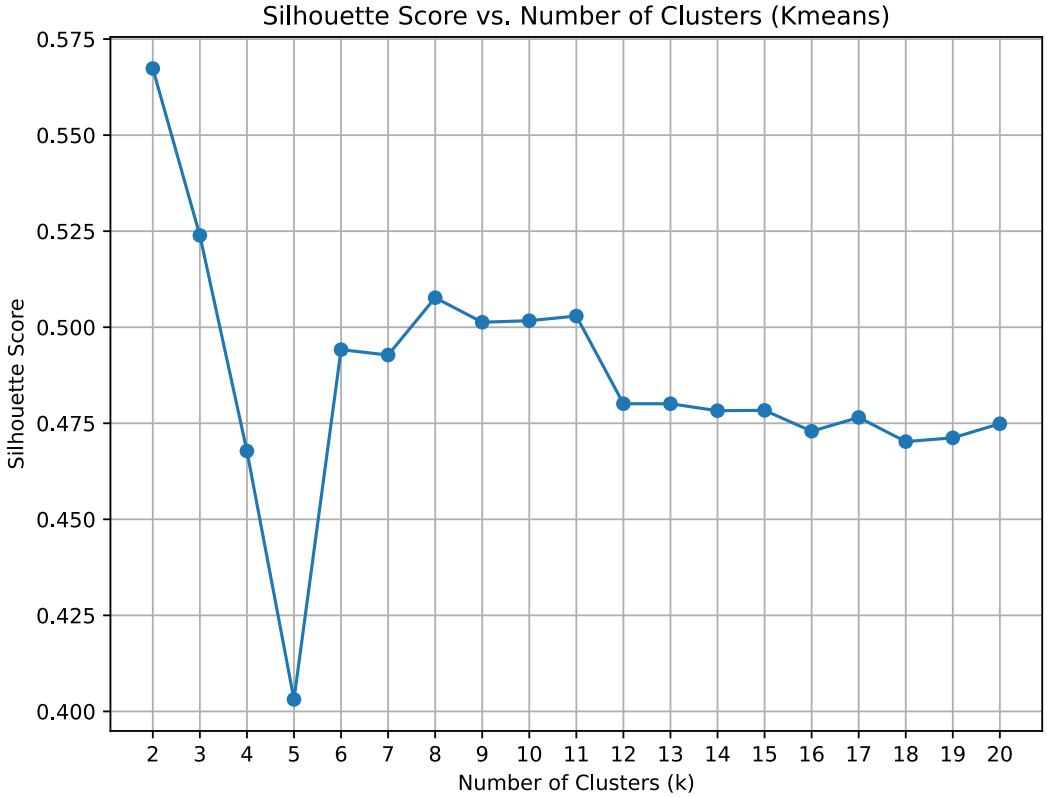

Figure 10 Silhouette Score Chart for different values of K for K-Means Clustering.

Clusters Profiling:

For this exercise we will concentrate on the k=2 clusters for the "**Hierarchical Clusters and the K-Means Cluster**", both of these give us Clusters which have the following features when we check the means of the values of the variables/columns that make up these clusters.

- Cluster 1 has **lower "Ad Size, Fee, CTR, CPM", Higer "Impressions, Clicks, Spend, Revenue, CPC"**
- Cluster 2 has **Higher "Ad Size, Fee, CTR, CPM", Lower "Impressions, Clicks, Spend, Revenue, CPC"**

!!Please Note: "It would be good to explore the other two K values, K=5 from the hierarchical clustering and K=6 from the K-Means Clustering at a later point for the business."

hcluster_k2	Ad Size	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
1	70186.72	9272826.17	17577.47	15528.92	0.24	11886.84	0.19	1.71	918.85
2	98516.79	682910.60	10198.67	1814.78	0.34	1231.32	8.98	8.86	296.19

Figure 11 Hierarchical Cluster 1 and 2 variables means compare

kmeans_k2	Ad Size	Impressions	Clicks	Spend	Fee	Revenue	CTR	CPM	CPC
0	72424.35	5777328.73	40849.54	12489.51	0.25	9436.59	5.53	6.88	598.60
1	99658.06	683459.74	6966.45	1503.00	0.35	999.98	8.76	8.58	304.45

Figure 12 K-Means Clustering Cluster 0 and 1 variable means compare

1.4 Clustering Actionable Insights & Recommendations

- There are clearly two groups**: Small_Ad_less_CPM_more_Revenue and Large_Ad_more_CPM_less_Revenue**
- More effective type of ads by size would be **smaller ads with sizes around 72276**.
- In order to effect Revenue better, should consider the avenues where the **Fee is 0.25 or lesser**.
- A **Marketing Strategy** to move form **"Large_Ad_more_CPM_less_Revenue"** segments to **"Small_Ad_less_CPM_more_Revenue"** segments would be **beneficial**.
- We should **study** the **"Small_Ad_less_CPM_more_Revenue"** more closely along with the **mix of 'InventoryType' 'Ad Type' 'Platform' 'Device Type' 'Format'** which make up this category to see **how different it is from** the **"Large_Ad_more_CPM_less_Revenue"** Group to **help us move ads to the 1st group**.

Problem 2

Overview

We have been given India Census data, the said data set has too many features to find useful details, we are tasked to use the Principal Components Analysis Statistical technique, to reduce the number of features be able to extract meaningful information.

Objective

Using the data provided will perform the following steps:

1. Define the problem
2. Explore the data
3. Get the statistical summary of the data.
4. Perform data preprocessing
5. Perform PCA

Dataset Description

The Data source is "Primary census abstract for female headed households"

The data set given has 61 Features and 640 observations, the data is by state and district code. All the features in the data are numeric expect, State and Area Name.

Name	Description
State	State Code
District	District Code
Name	Name
TRU1	Area Name
No_HH	No of Household
TOT_M	Total population Male
TOT_F	Total population Female
M_06	Population in the age group 0-6 Male
F_06	Population in the age group 0-6 Female
M_SC	Scheduled Castes population Male
F_SC	Scheduled Castes population Female
M_ST	Scheduled Tribes population Male
F_ST	Scheduled Tribes population Female
M_LIT	Literates' population Male
F_LIT	Literates' population Female
M_ILL	Illiterate Male
F_ILL	Illiterate Female
TOT_WORK_M	Total Worker Population Male
TOT_WORK_F	Total Worker Population Female
MAINWORK_M	Main Working Population Male
MAINWORK_F	Main Working Population Female
MAIN_CL_M	Main Cultivator Population Male
MAIN_CL_F	Main Cultivator Population Female
MAIN_AL_M	Main Agricultural Labourers Population Male
MAIN_AL_F	Main Agricultural Labourers Population Female
MAIN_HH_M	Main Household Industries Population Male
MAIN_HH_F	Main Household Industries Population Female
MAIN_OT_M	Main Other Workers Population Male
MAIN_OT_F	Main Other Workers Population Female
MARGWORK_M	Marginal Worker Population Male
MARGWORK_F	Marginal Worker Population Female
MARG_CL_M	Marginal Cultivator Population Male
MARG_CL_F	Marginal Cultivator Population Female
MARG_AL_M	Marginal Agriculture Labourers Population Male
MARG_AL_F	Marginal Agriculture Labourers Population Female
MARG_HH_M	Marginal Household Industries Population Male
MARG_HH_F	Marginal Household Industries Population Female
MARG_OT_M	Marginal Other Workers Population Male
MARG_OT_F	Marginal Other Workers Population Female

MARGWORK_3_6_M	Marginal Worker Population 3-6 Male
MARGWORK_3_6_F	Marginal Worker Population 3-6 Female
MARG_CL_3_6_M	Marginal Cultivator Population 3-6 Male
MARG_CL_3_6_F	Marginal Cultivator Population 3-6 Female
MARG_AL_3_6_M	Marginal Agriculture Labourers Population 3-6 Male
MARG_AL_3_6_F	Marginal Agriculture Labourers Population 3-6 Female
MARG_HH_3_6_M	Marginal Household Industries Population 3-6 Male
MARG_HH_3_6_F	Marginal Household Industries Population 3-6 Female
MARG_OT_3_6_M	Marginal Other Workers Population Person 3-6 Male
MARG_OT_3_6_F	Marginal Other Workers Population Person 3-6 Female
MARGWORK_0_3_M	Marginal Worker Population 0-3 Male
MARGWORK_0_3_F	Marginal Worker Population 0-3 Female
MARG_CL_0_3_M	Marginal Cultivator Population 0-3 Male
MARG_CL_0_3_F	Marginal Cultivator Population 0-3 Female
MARG_AL_0_3_M	Marginal Agriculture Labourers Population 0-3 Male
MARG_AL_0_3_F	Marginal Agriculture Labourers Population 0-3 Female
MARG_HH_0_3_M	Marginal Household Industries Population 0-3 Male
MARG_HH_0_3_F	Marginal Household Industries Population 0-3 Female
MARG_OT_0_3_M	Marginal Other Workers Population 0-3 Male
MARG_OT_0_3_F	Marginal Other Workers Population 0-3 Female
NON_WORK_M	Non-Working Population Male
NON_WORK_F	Non-Working Population Female

Table 2 Census Data Definitions

Questions Asked

2.1 Define the problem and perform Exploratory Data Analysis

2.1.1 Problem Definition

We have been given India Census data, the said data set has too many features to find useful details, we are tasked to use the Principal Components Analysis Statistical technique, to reduce the number of features be able to extract meaningful information.

2.1.2 Check Data Shape, Data Types and Statistical Summary.

The dataset provided has 640 observations with 61 features. The dataset has 2 columns with string values, 59 columns with int values.

2.1.2 Perform EDA on the data (Note: 5 variables chosen, "No_HH, TOT_M, TOT_F, M_06, F_06")

I will pick No_HH, TOT_M, TOT_F, M_06, F_06 Checking the above 5 variables this is what we find.

- State with the most households is Uttar Pradesh with ~4 million households which ~12% of the total households, State with the least households is Dadara & Nagar Havelli with 4288 households making up ~0.013% of the total households.
- District with the most households is North Twenty-Four Parganas in West Bengal with ~300K households which is about .94% of the overall households, District with the least number of households is the Dibang Valley in Arunachal Pradesh with 350 households making up 0.001% of the total households
- Since the Dataset is an "abstract for female headed households", the Gender ratio= TOT_F/TOT_M is higher than 100% for all states.
- State with the highest Gender Ratio is Andra Pradesh with the GR at ~1.9, State with the least Gender Ratio is Lakshadweep with the GR at ~1.2
- District "['Krishna']" from the State "['Andhra Pradesh']"has the most Gender Ratio with the GR ~2.3, District "['Lakshadweep']" from the State "['Lakshadweep']"has the least Gender Ratio with the GR at ~1.2.
- State with the highest Child Gender Ratio (CGR) = F_06/M_06 is Arunachal Pradesh with CGR ~1.1, State with the least CGR is Haryana with CGR ~0.9
- District "['East Kameng']" from the State "['Arunachal Pradesh']"has the most Child Gender Ratio with CGR ~1.2, District "['Samba']" from the State "['Jammu & Kashmir']"has the least Gender Ratio with CGR ~0.8.

Below are the plots used to get this information:

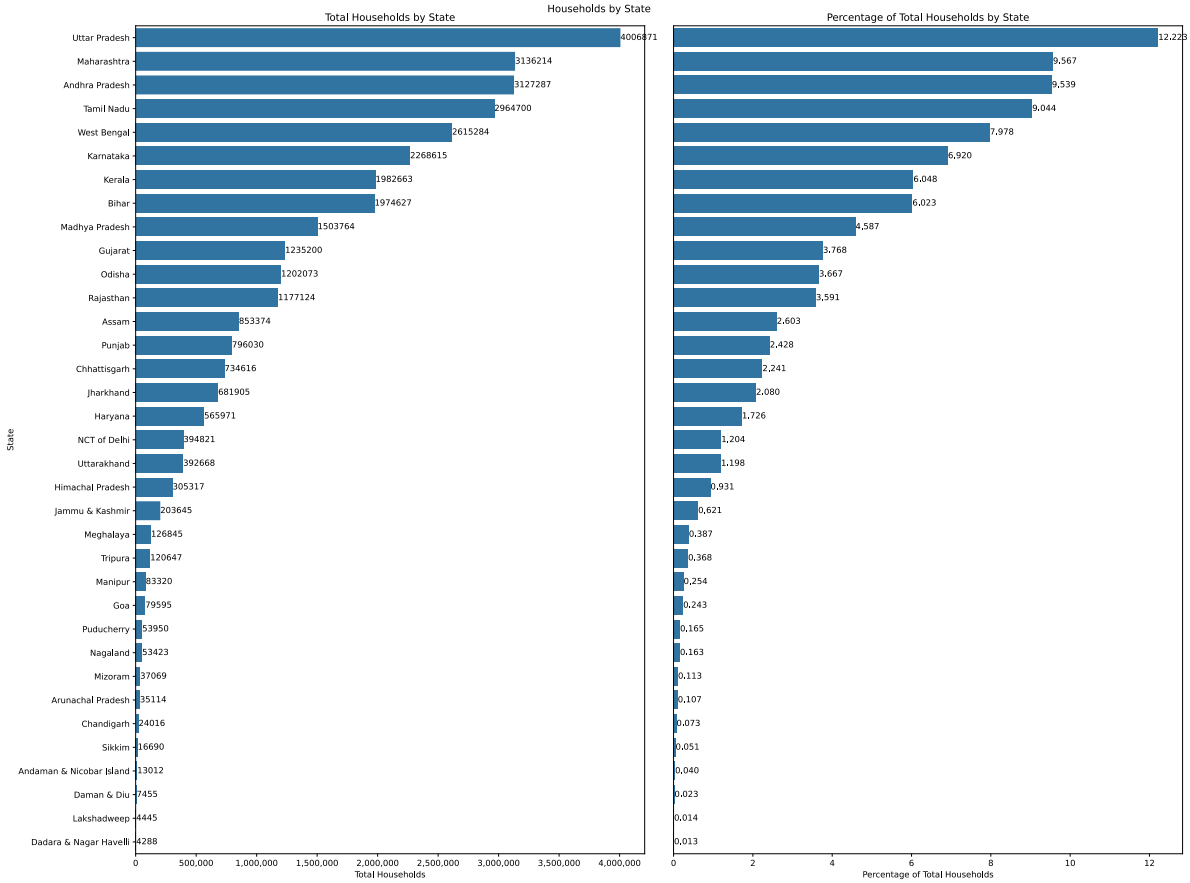

Figure 13 Households by States

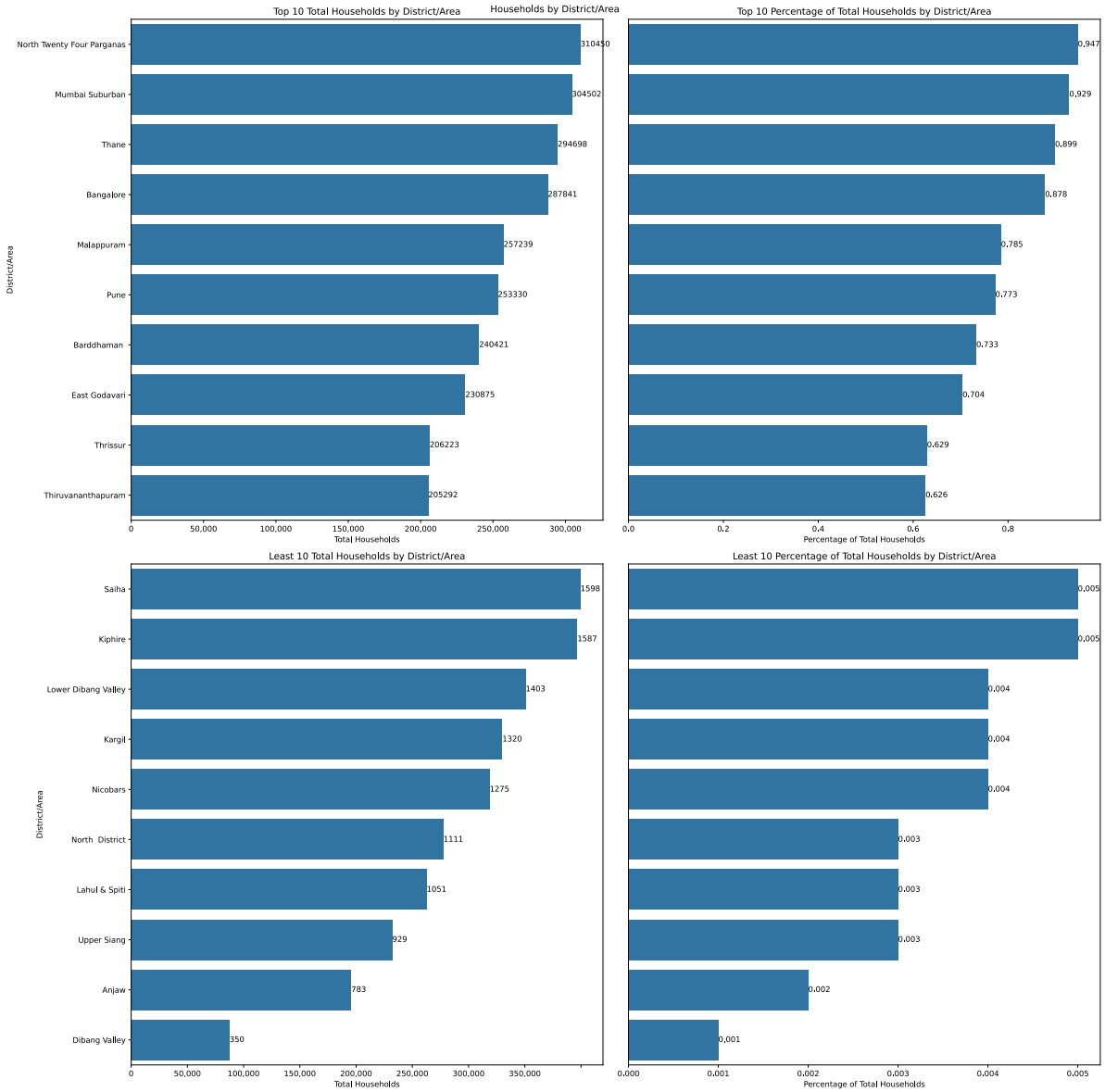

Figure 14 Top 10 and Least 10 Districts by House Hold Count.

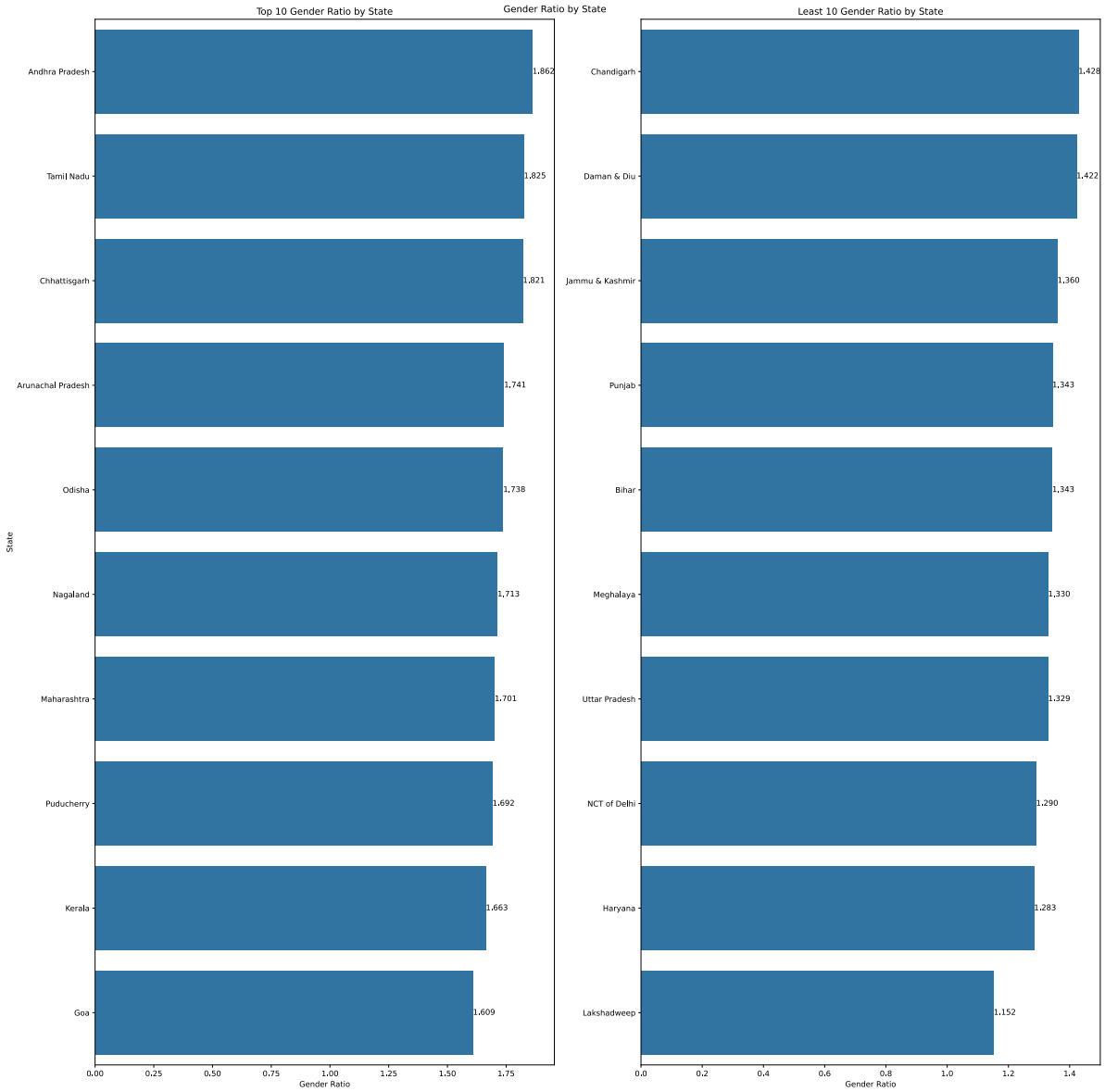

Figure 15 Top 10 and Least 10 States ranked by Gender Ratio

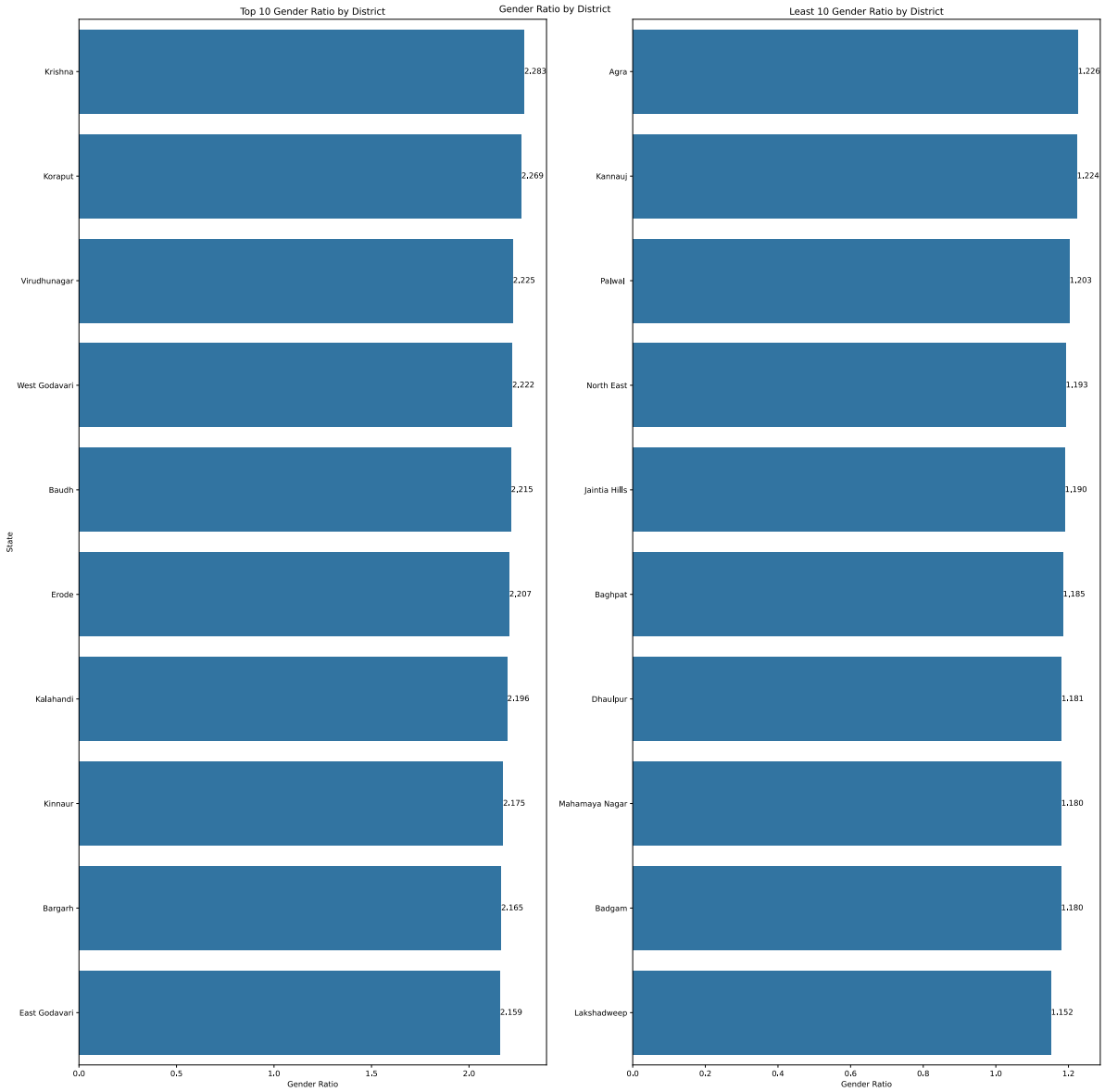

Figure 16 Top 10 and Least 10 Districts ranked by Gender Ratio

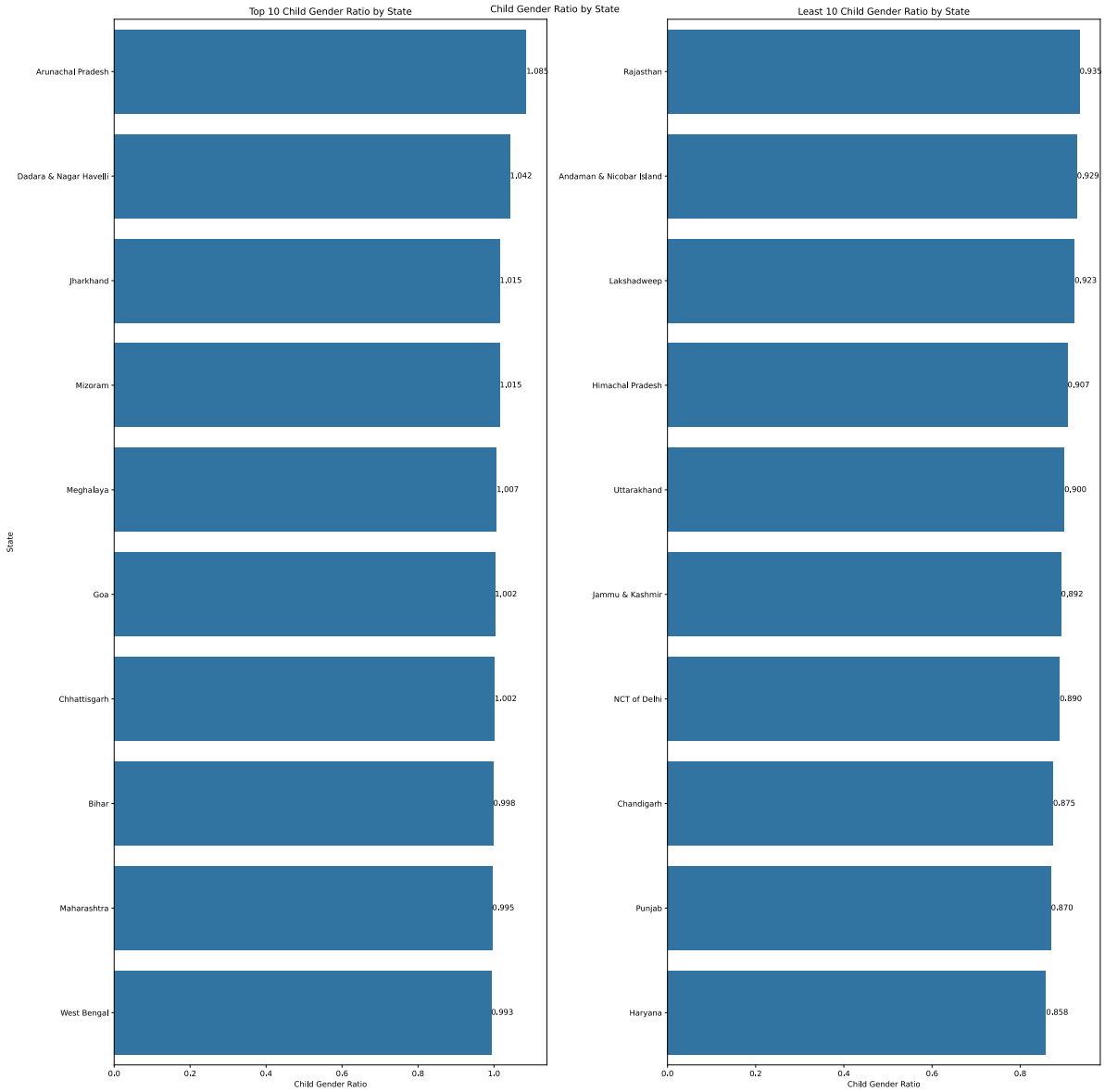

Figure 17 Top 10 and Least 10 States Ranked by Child Gender Ration (Child = 0-6 years)

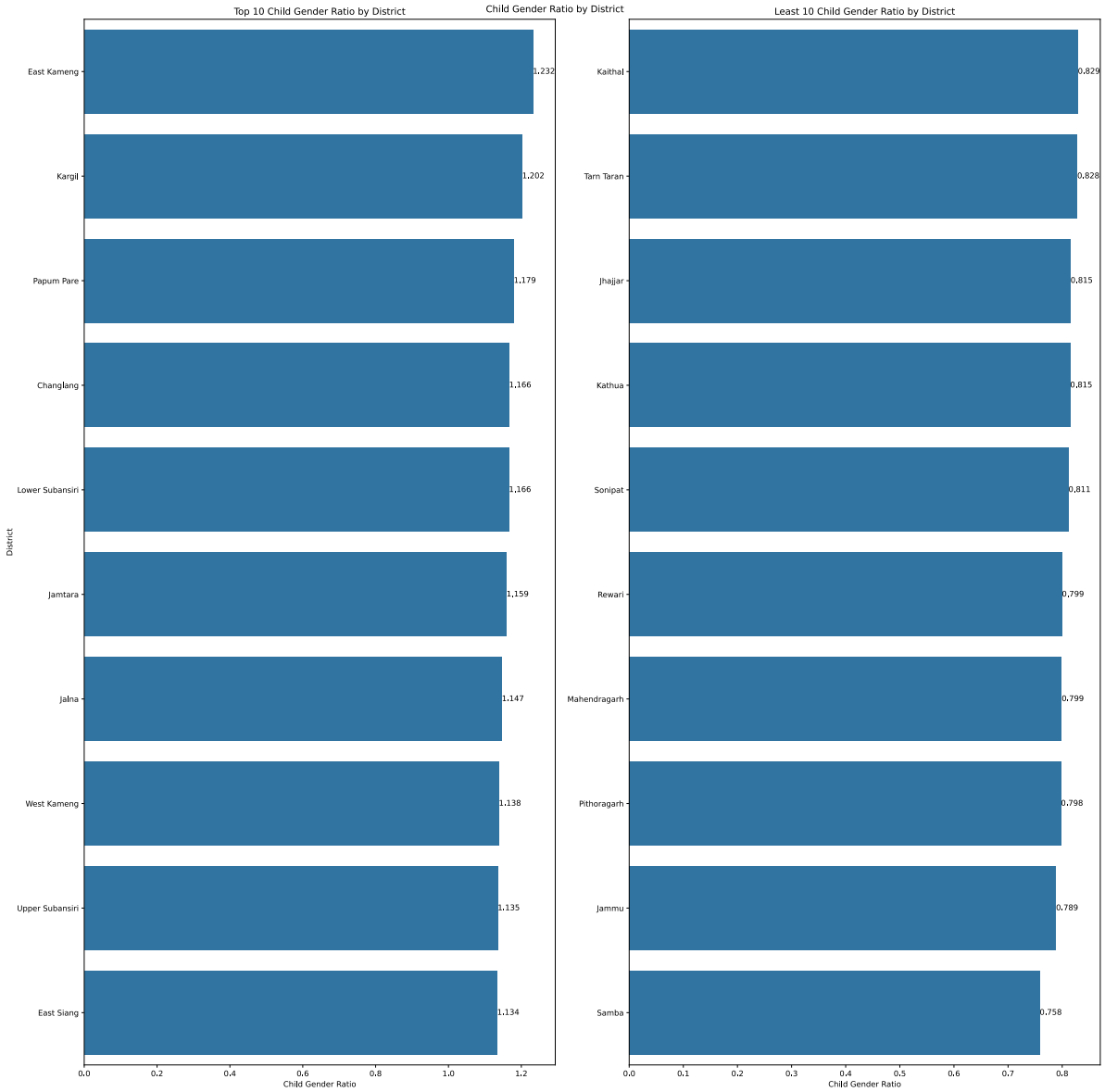

Figure 18 Top 10 and Least 10 Districts Ranked by Child Gender Ratio (Child = 0-6 years)

2.2 Data Preprocessing

We will not consider the two columns with string values, "State" and "Area Name" as there are the corresponding code values for the same.

	count	mean	std	min	25%	50%	75%	max
State Code	640.0	17.11	9.43	1.0	9.00	18.0	24.00	35.0
Dist.Code	640.0	320.50	184.90	1.0	160.75	320.5	480.25	640.0
No_HH	640.0	51222.87	48135.41	350.0	19484.00	35837.0	68892.00	310450.0
TOT_M	640.0	79940.58	73384.51	391.0	30228.00	58339.0	107918.50	485417.0
TOT_F	640.0	122372.08	113600.72	698.0	46517.75	87724.5	164251.75	750392.0
M_06	640.0	12309.10	11500.91	56.0	4733.75	9159.0	16520.25	96223.0
F_06	640.0	11942.30	11326.29	56.0	4672.25	8663.0	15902.25	95129.0
M_SC	640.0	13820.95	14426.37	0.0	3466.25	9591.5	19429.75	103307.0
F_SC	640.0	20778.39	21727.89	0.0	5603.25	13709.0	29180.00	156429.0
M_ST	640.0	6191.81	9912.67	0.0	293.75	2333.5	7658.00	96785.0
F_ST	640.0	10155.64	15875.70	0.0	429.50	3834.5	12480.25	130119.0
M_LIT	640.0	57967.98	55910.28	286.0	21298.00	42693.5	77989.50	403261.0
F_LIT	640.0	66359.57	75037.86	371.0	20932.00	43796.5	84799.75	571140.0
M_ILL	640.0	21972.60	19825.61	105.0	8590.00	15767.5	29512.50	105961.0
F_ILL	640.0	56012.52	47116.69	327.0	22367.00	42386.0	78471.00	254160.0
TOT_WORK_M	640.0	37992.41	36419.54	100.0	13753.50	27936.5	50226.75	269422.0
TOT_WORK_F	640.0	41295.76	37192.36	357.0	16097.75	30588.5	53234.25	257848.0
MAINWORK_M	640.0	30204.45	31480.92	65.0	9787.00	21250.5	40119.00	247911.0
MAINWORK_F	640.0	28198.85	29998.26	240.0	9502.25	18484.0	35063.25	226166.0
MAIN_CL_M	640.0	5424.34	4739.16	0.0	2023.50	4160.5	7695.00	29113.0
MAIN_CL_F	640.0	5486.04	5326.36	0.0	1920.25	3908.5	7286.25	36193.0
MAIN_AL_M	640.0	5849.11	6399.51	0.0	1070.25	3936.5	8067.25	40843.0
MAIN_AL_F	640.0	8926.00	12864.29	0.0	1408.75	3933.5	10617.50	87945.0
MAIN_HH_M	640.0	883.89	1278.64	0.0	187.50	498.5	1099.25	16429.0
MAIN_HH_F	640.0	1380.77	3179.41	0.0	248.75	540.5	1435.75	45979.0
MAIN_OT_M	640.0	18047.10	26068.48	36.0	3997.50	9598.0	21249.50	240855.0
MAIN_OT_F	640.0	12406.04	18972.20	153.0	3142.50	6380.5	14368.25	209355.0
MARGWORK_M	640.0	7787.96	7410.79	35.0	2937.50	5627.0	9800.25	47553.0

Figure 19 Describe of census data.

MARGWORK_F	640.0	13096.91	10996.47	117.0	5424.50	10175.0	18879.25	66915.0
MARG_CL_M	640.0	1040.74	1311.55	0.0	311.75	606.5	1281.00	13201.0
MARG_CL_F	640.0	2307.68	3564.63	0.0	630.25	1226.0	2659.25	44324.0
MARG_AL_M	640.0	3304.33	3781.56	0.0	873.50	2062.0	4300.75	23719.0
MARG_AL_F	640.0	6463.28	6773.88	0.0	1402.50	4020.5	9089.25	45301.0
MARG_HH_M	640.0	316.74	462.66	0.0	71.75	166.0	356.50	4298.0
MARG_HH_F	640.0	786.63	1198.72	0.0	171.75	429.0	962.50	15448.0
MARG_OT_M	640.0	3126.15	3609.39	7.0	935.50	2036.0	3985.25	24728.0
MARG_OT_F	640.0	3539.32	4115.19	19.0	1071.75	2349.5	4400.50	36377.0
MARGWORK_3_6_M	640.0	41948.17	39045.32	291.0	16208.25	30315.0	57218.75	300937.0
MARGWORK_3_6_F	640.0	81076.32	82970.41	341.0	26619.50	56793.0	107924.00	676450.0
MARG_CL_3_6_M	640.0	6394.99	6019.81	27.0	2372.00	4630.0	8167.00	39106.0
MARG_CL_3_6_F	640.0	10339.86	8467.47	85.0	4351.50	8295.0	15102.00	50065.0
MARG_AL_3_6_M	640.0	789.85	905.64	0.0	235.50	480.5	986.00	7426.0
MARG_AL_3_6_F	640.0	1749.58	2496.54	0.0	497.25	985.5	2059.00	27171.0
MARG_HH_3_6_M	640.0	2743.64	3059.59	0.0	718.75	1714.5	3702.25	19343.0
MARG_HH_3_6_F	640.0	5169.85	5335.64	0.0	1113.75	3294.0	7502.25	36253.0
MARG_OT_3_6_M	640.0	245.36	358.73	0.0	58.00	129.5	276.00	3535.0
MARG_OT_3_6_F	640.0	585.88	900.03	0.0	127.75	320.5	719.25	12094.0
MARGWORK_0_3_M	640.0	2616.14	3036.96	7.0	755.00	1681.5	3320.25	20648.0
MARGWORK_0_3_F	640.0	2834.55	3327.84	14.0	833.50	1834.5	3610.50	25844.0
MARG_CL_0_3_M	640.0	1392.97	1489.71	4.0	489.50	949.0	1714.00	9875.0
MARG_CL_0_3_F	640.0	2757.05	2788.78	30.0	957.25	1928.0	3599.75	21611.0
MARG_AL_0_3_M	640.0	250.89	453.34	0.0	47.00	114.5	270.75	5775.0
MARG_AL_0_3_F	640.0	558.10	1117.64	0.0	109.00	247.5	568.75	17153.0
MARG_HH_0_3_M	640.0	560.69	762.58	0.0	136.50	308.0	642.00	6116.0
MARG_HH_0_3_F	640.0	1293.43	1585.38	0.0	298.00	717.0	1710.75	13714.0
MARG_OT_0_3_M	640.0	71.38	107.90	0.0	14.00	35.0	79.00	895.0
MARG_OT_0_3_F	640.0	200.74	309.74	0.0	43.00	113.0	240.00	3354.0
NON_WORK_M	640.0	510.01	610.60	0.0	161.00	326.0	604.50	6456.0

Figure 20 Describe of census data continued.

2.2.1 Check and Treat Missing values

There are no missing values seen

2.2.2 Check and Treat data irregularities

- There is no Duplicates.
- Reviewing the data using describe, there are some features with min 0 values, but these could be valid entries as could be zero for said features, hence we will not treat these.

2.2.3 Scale Data

2.2.3.1 Data before scaling

There are outliers for all the features expect State Code and District Code which is expected as State Code and District Code are identifiers rather than counts.

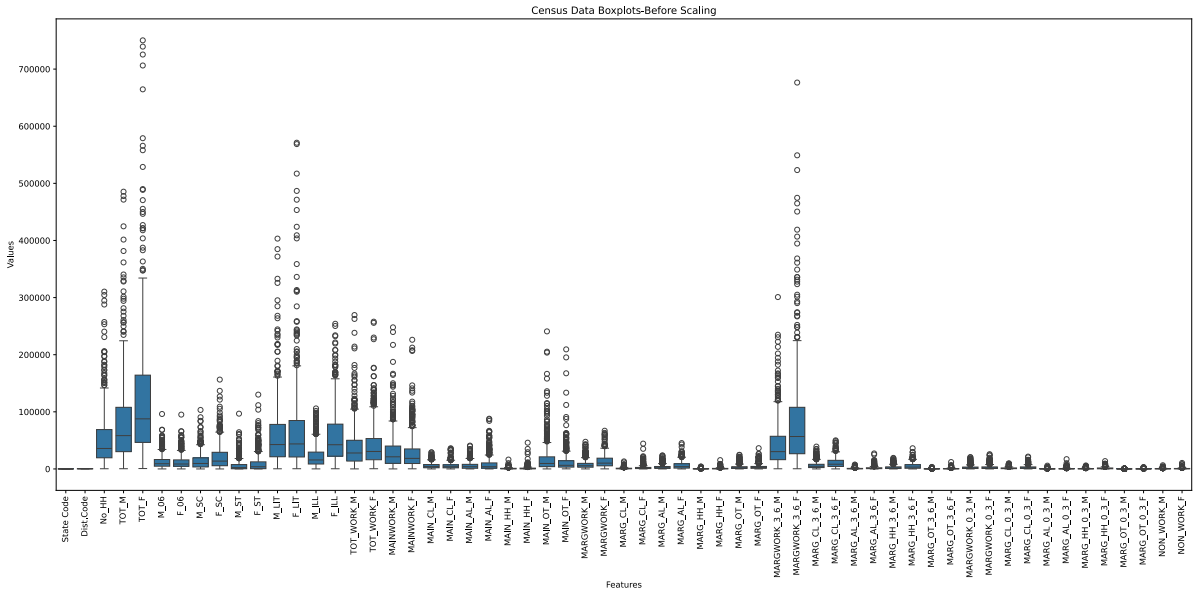

Figure 21 Boxplots of PCA Attributes before Scaling.

2.2.3.2 Data after scaling

Scale the Data using the z-score method

- We Scale the data using z-score method and plot the box plots again.
- There are outliers for all the features expect State Code and District Code which is expected as State Code and District Code are identifiers rather than counts.
- Scaling does not have an impact on outliers, we still see extreme outliers in the dataset.
- The data set has outliers and extreme outliers, we cannot remove these as these are not due to data errors, we ideally would explore the following below:(**but would not do the same for this exercise**)
    - Run transformations like square root, cube root, log, or box-cox etc, to reduce the skew in the data a potentially reduce the outliers and outlier impact
    - Post transformation and scaling if outliers/extreme outliers still exist, we would explore techniques like Robust PCA to make sure the impact of the outliers are minimized in the creation of the Principal Components.

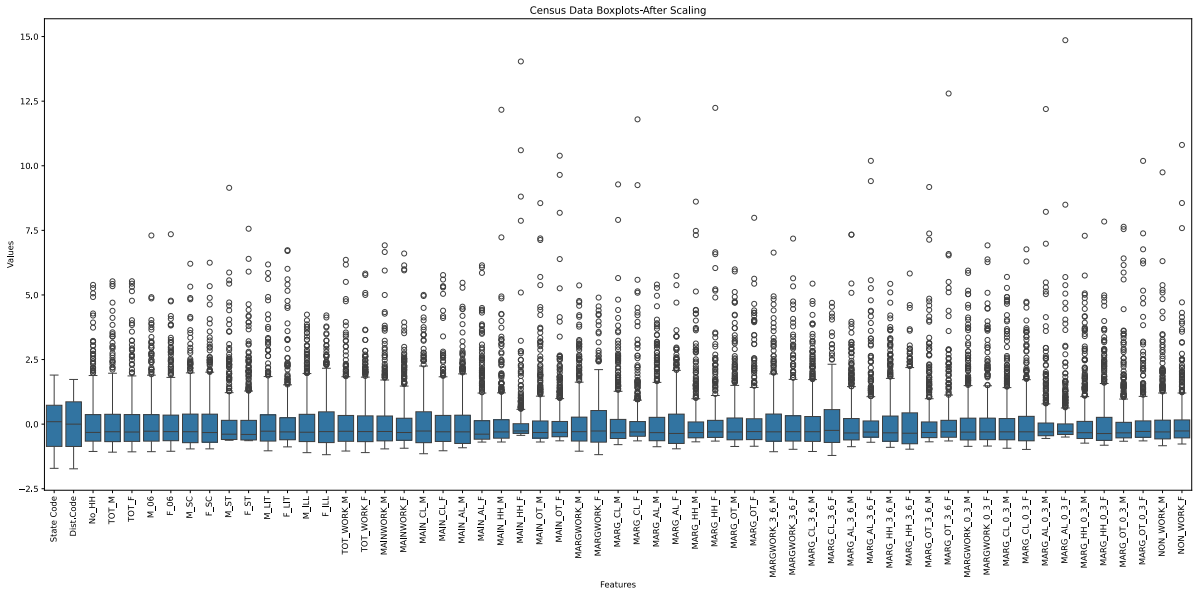

Figure 22 Boxplots PCA Attributes after Scaling

2.3 Perform Principal Component Analysis.

Note: For the scope of this project, take at least 90% explained variance.

2.3.1 Create the Covariance Matrix

Creating the covariance matrix using the cov() method, and plotting the same in a heatmap, we see that there are lot of instances where the value of the covariance is larger than 0.8. This indicates that there is a lot of covariance between the features and the dataset is a good candidate for PCA.

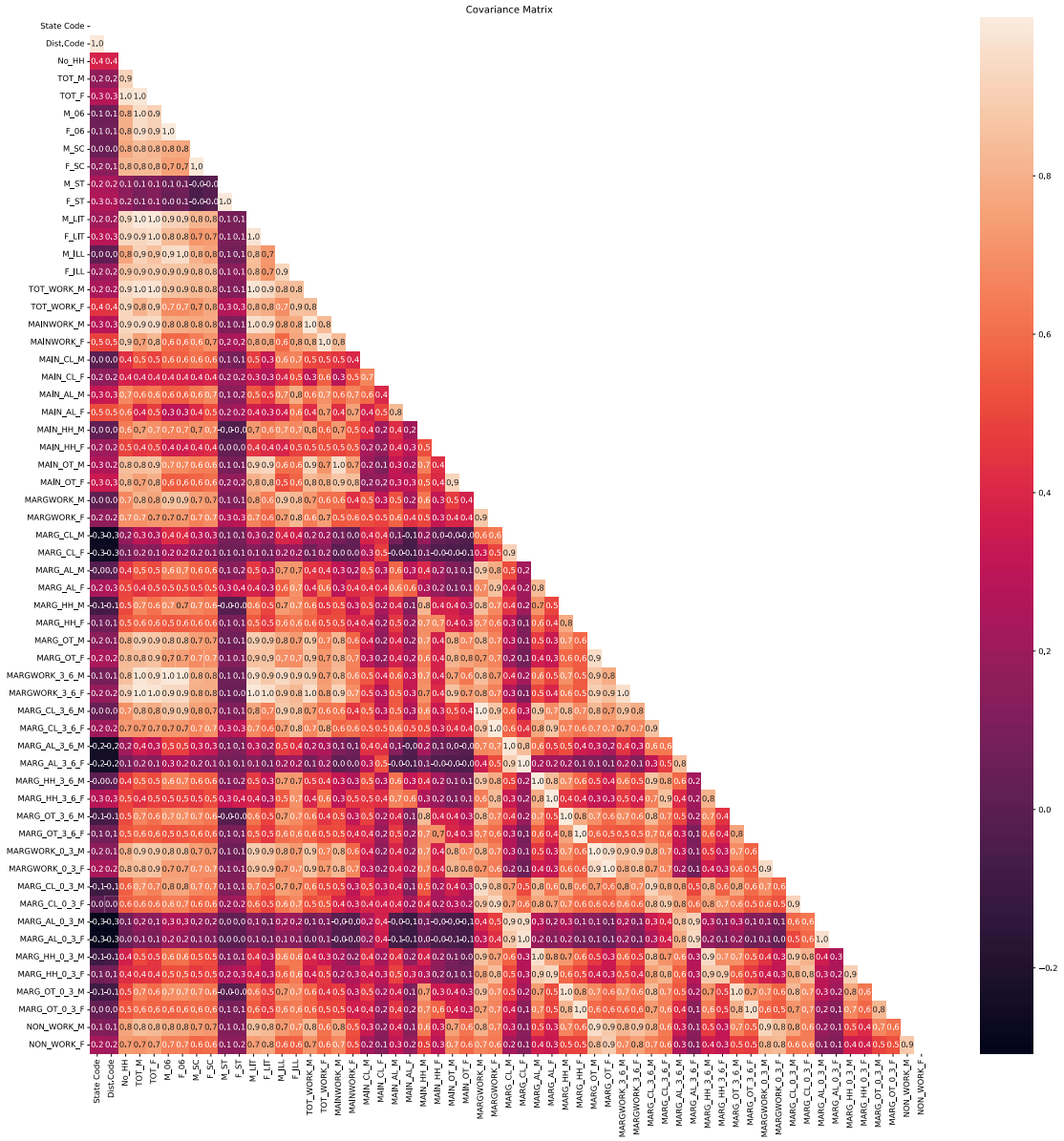

Figure 23 Census Data Variables Covariance Matrix

2.3.2 Get the eigen values and eigen vectors

Using the PCA method of the SKlearn package we can get the eigen values (explained_variance_), the eigen vectors (components_) and the eigen values ratios (explained_variance_ratio_).

- **Eigenvalues** (explained_variance_): In the context of Principal Component Analysis (PCA), eigenvalues represent the amount of variance in the data that is accounted for by each principal component. A larger eigenvalue corresponds to a direction along which the data varies more.
- **Eigenvectors** (components_): Eigenvectors are the directions in the feature space along which the original data is spread out. These are the principal components. The direction of the eigenvector captures the direction of the spread of the data, and the eigenvalue captures the magnitude of this spread.
- **Explained Variance Ratio** (explained_variance_ratio_): This is the proportion of the dataset's variance that lies along the axis of each principal component. It gives the amount of information (variance) that can be attributed to each of the principal components. This is especially useful when deciding how many principal components to keep during dimensionality reduction.

Components (components_): In sklearn's PCA, components_ returns the specific eigenvectors (principal components). Each row represents a principal component, and the columns correspond to the original features. This gives us the weights of each feature in the original data for each principal component.

2.3.3 Identify Optimum number of PCs

We take the threshold of 90% explained variance or .90 as the threshold to get the number of PCs which are Optimal, from the cumulative explained variance ratio we set this threshold and count the numbers of PCs below or equal to this threshold, we get a value of 6, which means that the 1st six PCs explain the variance in the data to a threshold of 90%.

Show Scree plot

Plotting the Scree Plot and specifically the Cumulative Scree Plot we see if we cut of the graph at 0.90 on the y axis, we have 6 PCs which would get us close to this.

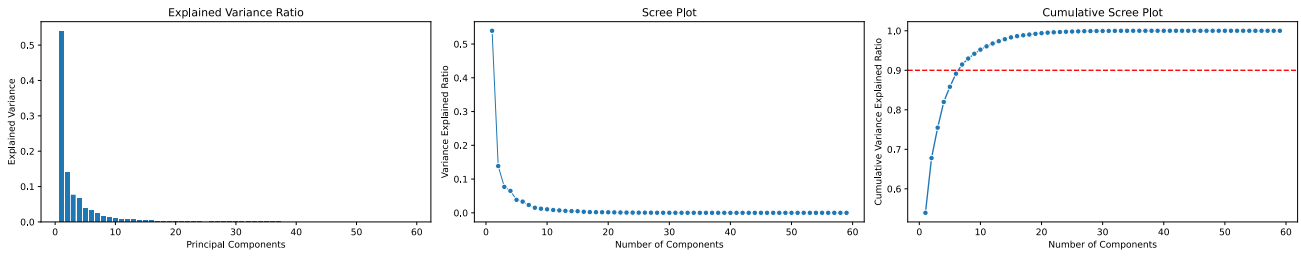

Figure 24 Explained Variance Plot, Scree Plot and Cumulative Scree Plot

2.3.4 Compare PCs with Actual Columns and Identify which is explaining most variance

Plotted the Heat map of the loadings for the Actual Columns (Feature) vs first 6 PCs

This gives us visibility to the influence of the feature within a PC, higher the magnitude more the influence, the sign lets us know if it's a positive influence or negative.

Heat Map of Loadings by Acutal Columns by Pricinpal Components, First 6 components are choosen.

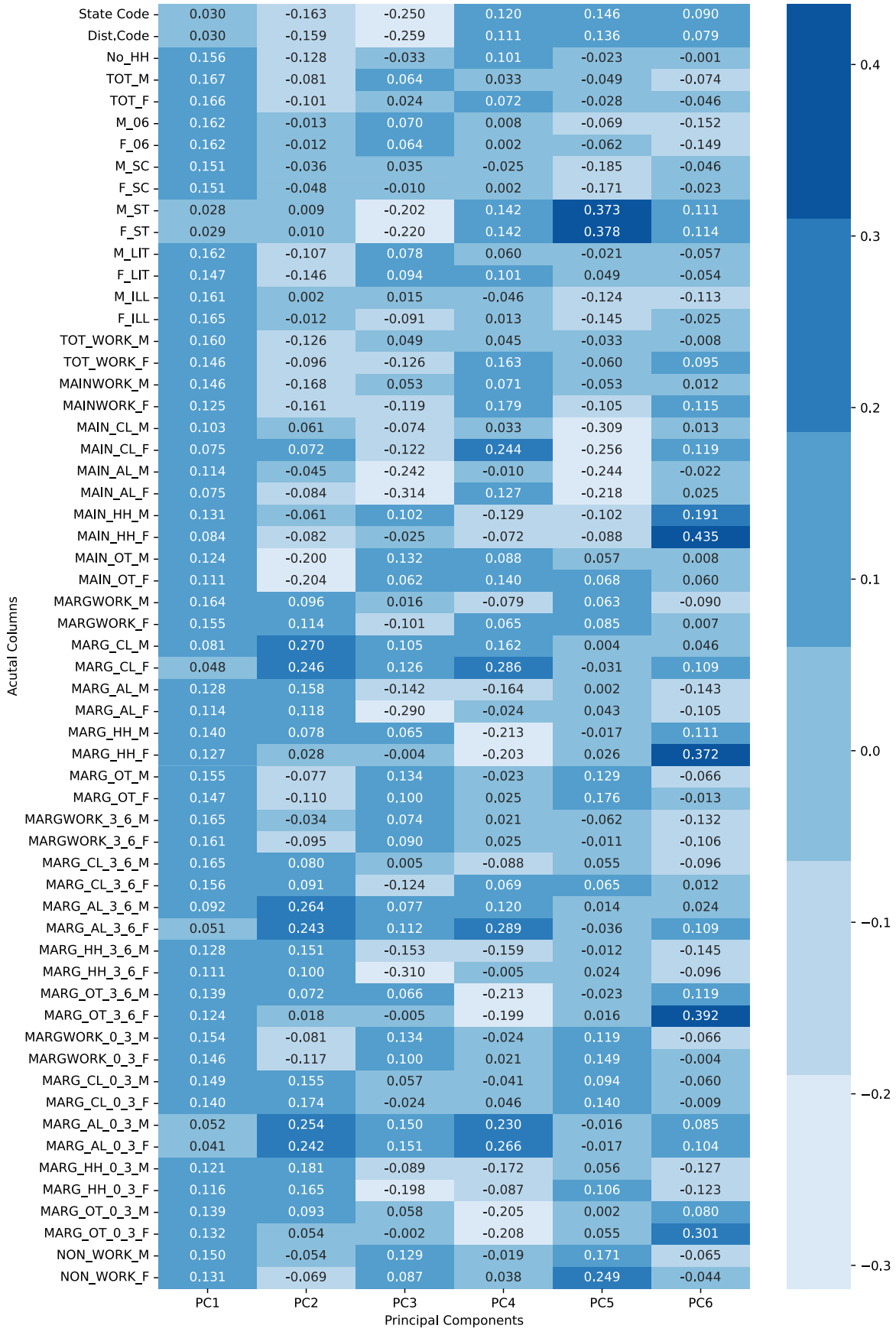

Figure 25 loadings for the Actual Columns (Feature) vs first 6 PCs

Plotted for each of the first 6 PC's the square of the loadings (tells us the contribution to variance) for the actual columns(features)

This shows us for each PC, what is the features and what is they contribution in explaining the variance with in the PC, we only looked at features which explained ~0.90 or ~90% of the variance within the PC.

- The list of the features explaining most of the variance in each of the 6 PCs are:
  - The feature explaining most that the variance for PC1 is, TOT_M, it explains 2.79% of the variance in PC1.
  - The feature explaining most that the variance for PC2 is, MARG_CL_M, it explains 7.29% of the variance in PC2.
  - The feature explaining most that the variance for PC3 is, MAIN_AL_F, it explains 9.83% of the variance in PC3.
  - The feature explaining most that the variance for PC4 is, MARG_AL_3_6_F, it explains 8.36% of the variance in PC4.
  - The feature explaining most that the variance for PC5 is, F_ST, it explains 14.3% of the variance in PC5.
  - The feature explaining most that the variance for PC6 is, MAIN_HH_F, it explains 18.95% of the variance in PC6.

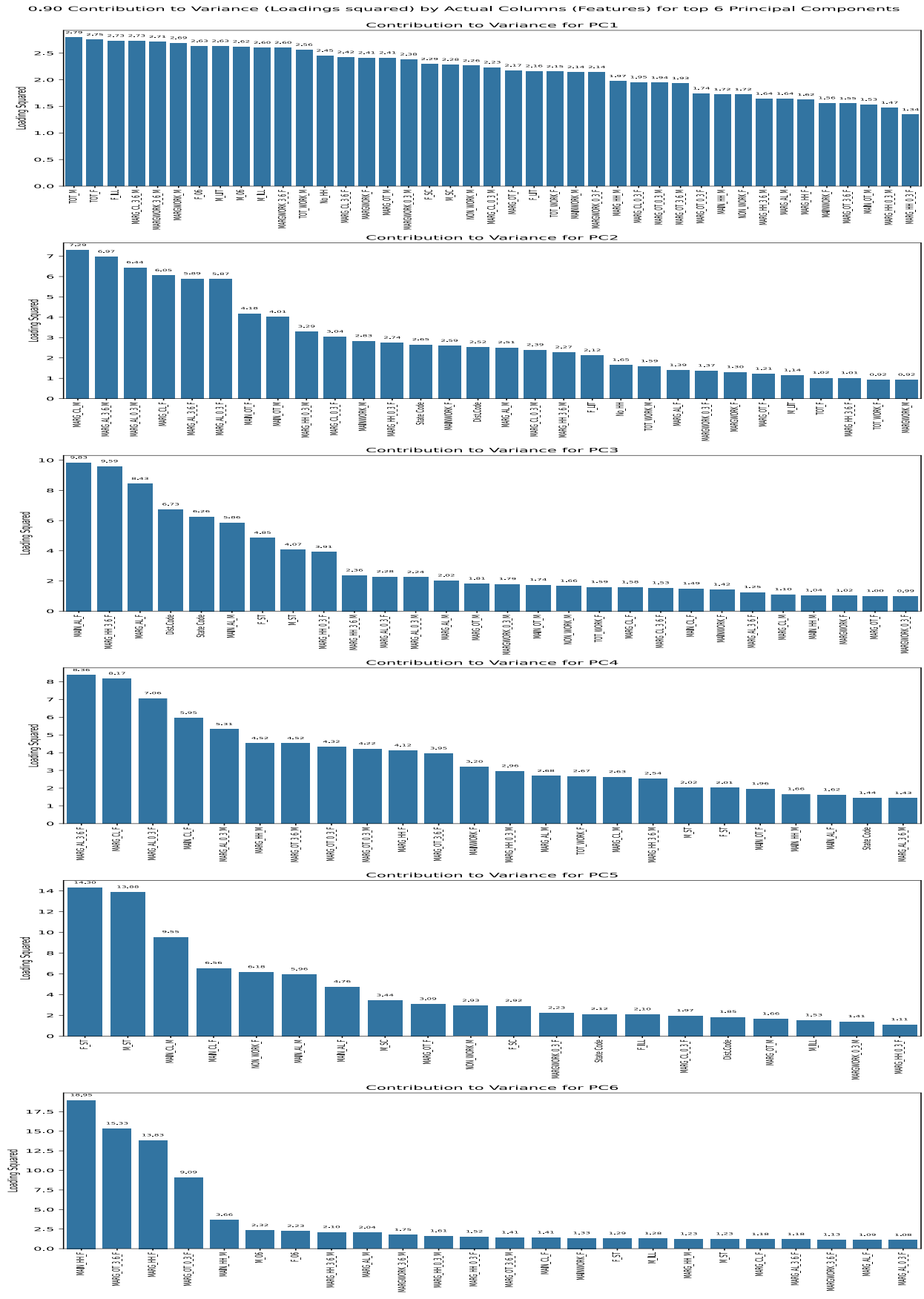

Figure 26 For first 6 PC's the square of the loadings (tells us the contribution to variance) for the actual columns(features)

2.3.5 Write Inferences about Optimum PCs in Terms of actual Variables

Inferences for PC1:

==========

- The feature explaining most that the variance for PC1 is, TOT_M, it explains 2.79% of the variance in PC1.
- There is a 16.704% Increase (if % is positive) or Decrease (if % is negative) in PC1 for every one unit increase in Scaled TOT_M

==========

- Features explaining ~90% of the variance of PC1 in descending order are:
- TOT_M:2.79%: There is a 16.704% Increase (if % is positive) or Decrease (if % is negative) in PC1 for every one unit increase in Scaled TOT_M.
- TOT_F:2.75%: There is a 16.57% Increase (if % is positive) or Decrease (if % is negative) in PC1 for every one unit increase in Scaled TOT_F.
- F_ILL:2.73%: There is a 16.522% Increase (if % is positive) or Decrease (if % is negative) in PC1 for every one unit increase in Scaled F_ILL.
- MARG_CL_3_6_M:2.73%: There is a 16.509% Increase (if % is positive) or Decrease (if % is negative) in PC1 for every one unit increase in Scaled MARG_CL_3_6_M.
- MARGWORK_3_6_M:2.71%: There is a 16.471% Increase (if % is positive) or Decrease (if % is negative) in PC1 for every one unit increase in Scaled MARGWORK_3_6_M.
- MARGWORK_M:2.69%: There is a 16.414% Increase (if % is positive) or Decrease (if % is negative) in PC1 for every one unit increase in Scaled MARGWORK_M.
- F_06:2.63%: There is a 16.227% Increase (if % is positive) or Decrease (if % is negative) in PC1 for every one unit increase in Scaled F_06.
- M_LIT:2.63%: There is a 16.203% Increase (if % is positive) or Decrease (if % is negative) in PC1 for every one unit increase in Scaled M_LIT.
- M_06:2.62%: There is a 16.187% Increase (if % is positive) or Decrease (if % is negative) in PC1 for every one unit increase in Scaled M_06.
- M_ILL:2.6%: There is a 16.135% Increase (if % is positive) or Decrease (if % is negative) in PC1 for every one unit increase in Scaled M_ILL.
- MARGWORK_3_6_F:2.6%: There is a 16.121% Increase (if % is positive) or Decrease (if % is negative) in PC1 for every one unit increase in Scaled MARGWORK_3_6_F.
- TOT_WORK_M:2.56%: There is a 15.999% Increase (if % is positive) or Decrease (if % is negative) in PC1 for every one unit increase in Scaled TOT_WORK_M.
- No_HH:2.45%: There is a 15.643% Increase (if % is positive) or Decrease (if % is negative) in PC1 for every one unit increase in Scaled No_HH.
- MARG_CL_3_6_F:2.42%: There is a 15.562% Increase (if % is positive) or Decrease (if % is negative) in PC1 for every one unit increase in Scaled MARG_CL_3_6_F.
- MARGWORK_F:2.41%: There is a 15.526% Increase (if % is positive) or Decrease (if % is negative) in PC1 for every one unit increase in Scaled MARGWORK_F.
- MARG_OT_M:2.41%: There is a 15.515% Increase (if % is positive) or Decrease (if % is negative) in PC1 for every one unit increase in Scaled MARG_OT_M.
- MARGWORK_0_3_M:2.38%: There is a 15.42% Increase (if % is positive) or Decrease (if % is negative) in PC1 for every one unit increase in Scaled MARGWORK_0_3_M.
- F_SC:2.29%: There is a 15.148% Increase (if % is positive) or Decrease (if % is negative) in PC1 for every one unit increase in Scaled F_SC.
- M_SC:2.28%: There is a 15.107% Increase (if % is positive) or Decrease (if % is negative) in PC1 for every one unit increase in Scaled M_SC.

- NON_WORK_M:2.26%: There is a 15.022% Increase (if % is positive) or Decrease (if % is negative) in PC1 for every one unit increase in Scaled NON_WORK_M.
- MARG_CL_0_3_M:2.23%: There is a 14.944% Increase (if % is positive) or Decrease (if % is negative) in PC1 for every one unit increase in Scaled MARG_CL_0_3_M.
- MARG_OT_F:2.17%: There is a 14.741% Increase (if % is positive) or Decrease (if % is negative) in PC1 for every one unit increase in Scaled MARG_OT_F.
- F_LIT:2.16%: There is a 14.712% Increase (if % is positive) or Decrease (if % is negative) in PC1 for every one unit increase in Scaled F_LIT.
- TOT_WORK_F:2.15%: There is a 14.648% Increase (if % is positive) or Decrease (if % is negative) in PC1 for every one unit increase in Scaled TOT_WORK_F.
- MAINWORK_M:2.14%: There is a 14.645% Increase (if % is positive) or Decrease (if % is negative) in PC1 for every one unit increase in Scaled MAINWORK_M.
- MARGWORK_0_3_F:2.14%: There is a 14.641% Increase (if % is positive) or Decrease (if % is negative) in PC1 for every one unit increase in Scaled MARGWORK_0_3_F.
- MARG_HH_M:1.97%: There is a 14.027% Increase (if % is positive) or Decrease (if % is negative) in PC1 for every one unit increase in Scaled MARG_HH_M.
- MARG_CL_0_3_F:1.95%: There is a 13.971% Increase (if % is positive) or Decrease (if % is negative) in PC1 for every one unit increase in Scaled MARG_CL_0_3_F.
- MARG_OT_0_3_M:1.94%: There is a 13.926% Increase (if % is positive) or Decrease (if % is negative) in PC1 for every one unit increase in Scaled MARG_OT_0_3_M.
- MARG_OT_3_6_M:1.93%: There is a 13.903% Increase (if % is positive) or Decrease (if % is negative) in PC1 for every one unit increase in Scaled MARG_OT_3_6_M.
- MARG_OT_0_3_F:1.74%: There is a 13.187% Increase (if % is positive) or Decrease (if % is negative) in PC1 for every one unit increase in Scaled MARG_OT_0_3_F.
- MAIN_HH_M:1.72%: There is a 13.128% Increase (if % is positive) or Decrease (if % is negative) in PC1 for every one unit increase in Scaled MAIN_HH_M.
- NON_WORK_F:1.72%: There is a 13.118% Increase (if % is positive) or Decrease (if % is negative) in PC1 for every one unit increase in Scaled NON_WORK_F.
- MARG_HH_3_6_M:1.64%: There is a 12.819% Increase (if % is positive) or Decrease (if % is negative) in PC1 for every one unit increase in Scaled MARG_HH_3_6_M.
- MARG_AL_M:1.64%: There is a 12.817% Increase (if % is positive) or Decrease (if % is negative) in PC1 for every one unit increase in Scaled MARG_AL_M.
- MARG_HH_F:1.62%: There is a 12.742% Increase (if % is positive) or Decrease (if % is negative) in PC1 for every one unit increase in Scaled MARG_HH_F.
- MAINWORK_F:1.56%: There is a 12.47% Increase (if % is positive) or Decrease (if % is negative) in PC1 for every one unit increase in Scaled MAINWORK_F.
- MARG_OT_3_6_F:1.55%: There is a 12.433% Increase (if % is positive) or Decrease (if % is negative) in PC1 for every one unit increase in Scaled MARG_OT_3_6_F.
- MAIN_OT_M:1.53%: There is a 12.379% Increase (if % is positive) or Decrease (if % is negative) in PC1 for every one unit increase in Scaled MAIN_OT_M.
- MARG_HH_0_3_M:1.47%: There is a 12.125% Increase (if % is positive) or Decrease (if % is negative) in PC1 for every one unit increase in Scaled MARG_HH_0_3_M.
- MARG_HH_0_3_F:1.34%: There is a 11.579% Increase (if % is positive) or Decrease (if % is negative) in PC1 for every one unit increase in Scaled MARG_HH_0_3_F.

Inferences for PC2:

==========

- The feature explaining most that the variance for PC2 is, MARG_CL_M, it explains 7.29% of the variance in PC2.
- There is a 27.001% Increase (if % is positive) or Decrease (if % is negative) in PC2 for every one unit increase in Scaled MARG_CL_M

==========

- Features explaining ~90% of the variance of PC2 in descending order are:
- MARG_CL_M:7.29%: There is a 27.001% Increase (if % is positive) or Decrease (if % is negative) in PC2 for every one unit increase in Scaled MARG_CL_M.
- MARG_AL_3_6_M:6.97%: There is a 26.396% Increase (if % is positive) or Decrease (if % is negative) in PC2 for every one unit increase in Scaled MARG_AL_3_6_M.
- MARG_AL_0_3_M:6.44%: There is a 25.383% Increase (if % is positive) or Decrease (if % is negative) in PC2 for every one unit increase in Scaled MARG_AL_0_3_M.
- MARG_CL_F:6.05%: There is a 24.599% Increase (if % is positive) or Decrease (if % is negative) in PC2 for every one unit increase in Scaled MARG_CL_F.
- MARG_AL_3_6_F:5.89%: There is a 24.279% Increase (if % is positive) or Decrease (if % is negative) in PC2 for every one unit increase in Scaled MARG_AL_3_6_F.
- MARG_AL_0_3_F:5.87%: There is a 24.222% Increase (if % is positive) or Decrease (if % is negative) in PC2 for every one unit increase in Scaled MARG_AL_0_3_F.
- MAIN_OT_F:4.18%: There is a -20.443% Increase (if % is positive) or Decrease (if % is negative) in PC2 for every one unit increase in Scaled MAIN_OT_F.
- MAIN_OT_M:4.01%: There is a -20.026% Increase (if % is positive) or Decrease (if % is negative) in PC2 for every one unit increase in Scaled MAIN_OT_M.
- MARG_HH_0_3_M:3.29%: There is a 18.128% Increase (if % is positive) or Decrease (if % is negative) in PC2 for every one unit increase in Scaled MARG_HH_0_3_M.
- MARG_CL_0_3_F:3.04%: There is a 17.443% Increase (if % is positive) or Decrease (if % is negative) in PC2 for every one unit increase in Scaled MARG_CL_0_3_F.
- MAINWORK_M:2.83%: There is a -16.833% Increase (if % is positive) or Decrease (if % is negative) in PC2 for every one unit increase in Scaled MAINWORK_M.
- MARG_HH_0_3_F:2.74%: There is a 16.54% Increase (if % is positive) or Decrease (if % is negative) in PC2 for every one unit increase in Scaled MARG_HH_0_3_F.
- State Code:2.65%: There is a -16.278% Increase (if % is positive) or Decrease (if % is negative) in PC2 for every one unit increase in Scaled State Code.
- MAINWORK_F:2.59%: There is a -16.104% Increase (if % is positive) or Decrease (if % is negative) in PC2 for every one unit increase in Scaled MAINWORK_F.
- Dist.Code:2.52%: There is a -15.882% Increase (if % is positive) or Decrease (if % is negative) in PC2 for every one unit increase in Scaled Dist. Code.
- MARG_AL_M:2.51%: There is a 15.84% Increase (if % is positive) or Decrease (if % is negative) in PC2 for every one unit increase in Scaled MARG_AL_M.
- MARG_CL_0_3_M:2.39%: There is a 15.451% Increase (if % is positive) or Decrease (if % is negative) in PC2 for every one unit increase in Scaled MARG_CL_0_3_M.
- MARG_HH_3_6_M:2.27%: There is a 15.06% Increase (if % is positive) or Decrease (if % is negative) in PC2 for every one unit increase in Scaled MARG_HH_3_6_M.
- F_LIT:2.12%: There is a -14.565% Increase (if % is positive) or Decrease (if % is negative) in PC2 for every one unit increase in Scaled F_LIT.
- No_HH:1.65%: There is a -12.832% Increase (if % is positive) or Decrease (if % is negative) in PC2 for every one unit increase in Scaled No_HH.

- TOT_WORK_M:1.59%: There is a -12.602% Increase (if % is positive) or Decrease (if % is negative) in PC2 for every one unit increase in Scaled TOT_WORK_M.
- MARG_AL_F:1.39%: There is a 11.772% Increase (if % is positive) or Decrease (if % is negative) in PC2 for every one unit increase in Scaled MARG_AL_F.
- MARGWORK_0_3_F:1.37%: There is a -11.721% Increase (if % is positive) or Decrease (if % is negative) in PC2 for every one unit increase in Scaled MARGWORK_0_3_F.
- MARGWORK_F:1.3%: There is a 11.406% Increase (if % is positive) or Decrease (if % is negative) in PC2 for every one unit increase in Scaled MARGWORK_F.
- MARG_OT_F:1.21%: There is a -11.015% Increase (if % is positive) or Decrease (if % is negative) in PC2 for every one unit increase in Scaled MARG_OT_F.
- M_LIT:1.14%: There is a -10.671% Increase (if % is positive) or Decrease (if % is negative) in PC2 for every one unit increase in Scaled M_LIT.
- TOT_F:1.02%: There is a -10.111% Increase (if % is positive) or Decrease (if % is negative) in PC2 for every one unit increase in Scaled TOT_F.
- MARG_HH_3_6_F:1.01%: There is a 10.031% Increase (if % is positive) or Decrease (if % is negative) in PC2 for every one unit increase in Scaled MARG_HH_3_6_F.
- TOT_WORK_F:0.92%: There is a -9.617% Increase (if % is positive) or Decrease (if % is negative) in PC2 for every one unit increase in Scaled TOT_WORK_F.
- MARGWORK_M:0.92%: There is a 9.573% Increase (if % is positive) or Decrease (if % is negative) in PC2 for every one unit increase in Scaled MARGWORK_M.

Inferences for PC3:

==========

- The feature explaining most that the variance for PC3 is, MAIN_AL_F, it explains 9.83% of the variance in PC3.
- There is a -31.353% Increase (if % is positive) or Decrease (if % is negative) in PC3 for every one unit increase in Scaled MAIN_AL_F

==========

- Features explaining ~90% of the variance of PC3 in descending order are:
- MAIN_AL_F:9.83%: There is a -31.353% Increase (if % is positive) or Decrease (if % is negative) in PC3 for every one unit increase in Scaled MAIN_AL_F.
- MARG_HH_3_6_F:9.59%: There is a -30.974% Increase (if % is positive) or Decrease (if % is negative) in PC3 for every one unit increase in Scaled MARG_HH_3_6_F.
- MARG_AL_F:8.43%: There is a -29.027% Increase (if % is positive) or Decrease (if % is negative) in PC3 for every one unit increase in Scaled MARG_AL_F.
- Dist.Code:6.73%: There is a -25.936% Increase (if % is positive) or Decrease (if % is negative) in PC3 for every one unit increase in Scaled Dist. Code.
- State Code:6.26%: There is a -25.013% Increase (if % is positive) or Decrease (if % is negative) in PC3 for every one unit increase in Scaled State Code.
- MAIN_AL_M:5.86%: There is a -24.198% Increase (if % is positive) or Decrease (if % is negative) in PC3 for every one unit increase in Scaled MAIN_AL_M.
- F_ST:4.85%: There is a -22.013% Increase (if % is positive) or Decrease (if % is negative) in PC3 for every one unit increase in Scaled F_ST.
- M_ST:4.07%: There is a -20.176% Increase (if % is positive) or Decrease (if % is negative) in PC3 for every one unit increase in Scaled M_ST.
- MARG_HH_0_3_F:3.91%: There is a -19.78% Increase (if % is positive) or Decrease (if % is negative) in PC3 for every one unit increase in Scaled MARG_HH_0_3_F.

- MARG_HH_3_6_M:2.36%: There is a -15.35% Increase (if % is positive) or Decrease (if % is negative) in PC3 for every one unit increase in Scaled MARG_HH_3_6_M.
- MARG_AL_0_3_F:2.28%: There is a 15.108% Increase (if % is positive) or Decrease (if % is negative) in PC3 for every one unit increase in Scaled MARG_AL_0_3_F.
- MARG_AL_0_3_M:2.24%: There is a 14.959% Increase (if % is positive) or Decrease (if % is negative) in PC3 for every one unit increase in Scaled MARG_AL_0_3_M.
- MARG_AL_M:2.02%: There is a -14.207% Increase (if % is positive) or Decrease (if % is negative) in PC3 for every one unit increase in Scaled MARG_AL_M.
- MARG_OT_M:1.81%: There is a 13.447% Increase (if % is positive) or Decrease (if % is negative) in PC3 for every one unit increase in Scaled MARG_OT_M.
- MARGWORK_0_3_M:1.79%: There is a 13.389% Increase (if % is positive) or Decrease (if % is negative) in PC3 for every one unit increase in Scaled MARGWORK_0_3_M.
- MAIN_OT_M:1.74%: There is a 13.207% Increase (if % is positive) or Decrease (if % is negative) in PC3 for every one unit increase in Scaled MAIN_OT_M.
- NON_WORK_M:1.66%: There is a 12.896% Increase (if % is positive) or Decrease (if % is negative) in PC3 for every one unit increase in Scaled NON_WORK_M.
- TOT_WORK_F:1.59%: There is a -12.615% Increase (if % is positive) or Decrease (if % is negative) in PC3 for every one unit increase in Scaled TOT_WORK_F.
- MARG_CL_F:1.58%: There is a 12.565% Increase (if % is positive) or Decrease (if % is negative) in PC3 for every one unit increase in Scaled MARG_CL_F.
- MARG_CL_3_6_F:1.53%: There is a -12.358% Increase (if % is positive) or Decrease (if % is negative) in PC3 for every one unit increase in Scaled MARG_CL_3_6_F.
- MAIN_CL_F:1.49%: There is a -12.193% Increase (if % is positive) or Decrease (if % is negative) in PC3 for every one unit increase in Scaled MAIN_CL_F.
- MAINWORK_F:1.42%: There is a -11.931% Increase (if % is positive) or Decrease (if % is negative) in PC3 for every one unit increase in Scaled MAINWORK_F.
- MARG_AL_3_6_F:1.25%: There is a 11.176% Increase (if % is positive) or Decrease (if % is negative) in PC3 for every one unit increase in Scaled MARG_AL_3_6_F.
- MARG_CL_M:1.1%: There is a 10.467% Increase (if % is positive) or Decrease (if % is negative) in PC3 for every one unit increase in Scaled MARG_CL_M.
- MAIN_HH_M:1.04%: There is a 10.21% Increase (if % is positive) or Decrease (if % is negative) in PC3 for every one unit increase in Scaled MAIN_HH_M.
- MARGWORK_F:1.02%: There is a -10.119% Increase (if % is positive) or Decrease (if % is negative) in PC3 for every one unit increase in Scaled MARGWORK_F.
- MARG_OT_F:1.0%: There is a 9.977% Increase (if % is positive) or Decrease (if % is negative) in PC3 for every one unit increase in Scaled MARG_OT_F.
- MARGWORK_0_3_F:0.99%: There is a 9.966% Increase (if % is positive) or Decrease (if % is negative) in PC3 for every one unit increase in Scaled MARGWORK_0_3_F.

Inferences for PC4:

==========

- The feature explaining most that the variance for PC4 is, MARG_AL_3_6_F, it explains 8.36% of the variance in PC4.
- There is a 28.909% Increase (if % is positive) or Decrease (if % is negative) in PC4 for every one unit increase in Scaled MARG_AL_3_6_F

==========

- Features explaining ~90% of the variance of PC4 in descending order are:

- MARG_AL_3_6_F:8.36%: There is a 28.909% Increase (if % is positive) or Decrease (if % is negative) in PC4 for every one unit increase in Scaled MARG_AL_3_6_F.
- MARG_CL_F:8.17%: There is a 28.579% Increase (if % is positive) or Decrease (if % is negative) in PC4 for every one unit increase in Scaled MARG_CL_F.
- MARG_AL_0_3_F:7.06%: There is a 26.575% Increase (if % is positive) or Decrease (if % is negative) in PC4 for every one unit increase in Scaled MARG_AL_0_3_F.
- MAIN_CL_F:5.95%: There is a 24.396% Increase (if % is positive) or Decrease (if % is negative) in PC4 for every one unit increase in Scaled MAIN_CL_F.
- MARG_AL_0_3_M:5.31%: There is a 23.044% Increase (if % is positive) or Decrease (if % is negative) in PC4 for every one unit increase in Scaled MARG_AL_0_3_M.
- MARG_HH_M:4.52%: There is a -21.271% Increase (if % is positive) or Decrease (if % is negative) in PC4 for every one unit increase in Scaled MARG_HH_M.
- MARG_OT_3_6_M:4.52%: There is a -21.257% Increase (if % is positive) or Decrease (if % is negative) in PC4 for every one unit increase in Scaled MARG_OT_3_6_M.
- MARG_OT_0_3_F:4.32%: There is a -20.776% Increase (if % is positive) or Decrease (if % is negative) in PC4 for every one unit increase in Scaled MARG_OT_0_3_F.
- MARG_OT_0_3_M:4.22%: There is a -20.537% Increase (if % is positive) or Decrease (if % is negative) in PC4 for every one unit increase in Scaled MARG_OT_0_3_M.
- MARG_HH_F:4.12%: There is a -20.293% Increase (if % is positive) or Decrease (if % is negative) in PC4 for every one unit increase in Scaled MARG_HH_F.
- MARG_OT_3_6_F:3.95%: There is a -19.878% Increase (if % is positive) or Decrease (if % is negative) in PC4 for every one unit increase in Scaled MARG_OT_3_6_F.
- MAINWORK_F:3.2%: There is a 17.875% Increase (if % is positive) or Decrease (if % is negative) in PC4 for every one unit increase in Scaled MAINWORK_F.
- MARG_HH_0_3_M:2.96%: There is a -17.192% Increase (if % is positive) or Decrease (if % is negative) in PC4 for every one unit increase in Scaled MARG_HH_0_3_M.
- MARG_AL_M:2.68%: There is a -16.356% Increase (if % is positive) or Decrease (if % is negative) in PC4 for every one unit increase in Scaled MARG_AL_M.
- TOT_WORK_F:2.67%: There is a 16.341% Increase (if % is positive) or Decrease (if % is negative) in PC4 for every one unit increase in Scaled TOT_WORK_F.
- MARG_CL_M:2.63%: There is a 16.231% Increase (if % is positive) or Decrease (if % is negative) in PC4 for every one unit increase in Scaled MARG_CL_M.
- MARG_HH_3_6_M:2.54%: There is a -15.93% Increase (if % is positive) or Decrease (if % is negative) in PC4 for every one unit increase in Scaled MARG_HH_3_6_M.
- M_ST:2.02%: There is a 14.213% Increase (if % is positive) or Decrease (if % is negative) in PC4 for every one unit increase in Scaled M_ST.
- F_ST:2.01%: There is a 14.194% Increase (if % is positive) or Decrease (if % is negative) in PC4 for every one unit increase in Scaled F_ST.
- MAIN_OT_F:1.96%: There is a 13.995% Increase (if % is positive) or Decrease (if % is negative) in PC4 for every one unit increase in Scaled MAIN_OT_F.
- MAIN_HH_M:1.66%: There is a -12.901% Increase (if % is positive) or Decrease (if % is negative) in PC4 for every one unit increase in Scaled MAIN_HH_M.
- MAIN_AL_F:1.62%: There is a 12.731% Increase (if % is positive) or Decrease (if % is negative) in PC4 for every one unit increase in Scaled MAIN_AL_F.
- State Code:1.44%: There is a 12.005% Increase (if % is positive) or Decrease (if % is negative) in PC4 for every one unit increase in Scaled State Code.
- MARG_AL_3_6_M:1.43%: There is a 11.97% Increase (if % is positive) or Decrease (if % is negative) in PC4 for every one unit increase in Scaled MARG_AL_3_6_M.

Inferences for PC5:

==========

- The feature explaining most that the variance for PC5 is, F_ST, it explains 14.3% of the variance in PC5.
- There is a 37.813% Increase (if % is positive) or Decrease (if % is negative) in PC5 for every one unit increase in Scaled F_ST

==========

- Features explaining ~90% of the variance of PC5 in descending order are:
- F_ST:14.3%: There is a 37.813% Increase (if % is positive) or Decrease (if % is negative) in PC5 for every one unit increase in Scaled F_ST.
- M_ST:13.88%: There is a 37.254% Increase (if % is positive) or Decrease (if % is negative) in PC5 for every one unit increase in Scaled M_ST.
- MAIN_CL_M:9.55%: There is a -30.909% Increase (if % is positive) or Decrease (if % is negative) in PC5 for every one unit increase in Scaled MAIN_CL_M.
- MAIN_CL_F:6.56%: There is a -25.618% Increase (if % is positive) or Decrease (if % is negative) in PC5 for every one unit increase in Scaled MAIN_CL_F.
- NON_WORK_F:6.18%: There is a 24.857% Increase (if % is positive) or Decrease (if % is negative) in PC5 for every one unit increase in Scaled NON_WORK_F.
- MAIN_AL_M:5.96%: There is a -24.421% Increase (if % is positive) or Decrease (if % is negative) in PC5 for every one unit increase in Scaled MAIN_AL_M.
- MAIN_AL_F:4.76%: There is a -21.817% Increase (if % is positive) or Decrease (if % is negative) in PC5 for every one unit increase in Scaled MAIN_AL_F.
- M_SC:3.44%: There is a -18.539% Increase (if % is positive) or Decrease (if % is negative) in PC5 for every one unit increase in Scaled M_SC.
- MARG_OT_F:3.09%: There is a 17.575% Increase (if % is positive) or Decrease (if % is negative) in PC5 for every one unit increase in Scaled MARG_OT_F.
- NON_WORK_M:2.93%: There is a 17.116% Increase (if % is positive) or Decrease (if % is negative) in PC5 for every one unit increase in Scaled NON_WORK_M.
- F_SC:2.92%: There is a -17.082% Increase (if % is positive) or Decrease (if % is negative) in PC5 for every one unit increase in Scaled F_SC.
- MARGWORK_0_3_F:2.23%: There is a 14.934% Increase (if % is positive) or Decrease (if % is negative) in PC5 for every one unit increase in Scaled MARGWORK_0_3_F.
- State Code:2.12%: There is a 14.575% Increase (if % is positive) or Decrease (if % is negative) in PC5 for every one unit increase in Scaled State Code.
- F_ILL:2.1%: There is a -14.478% Increase (if % is positive) or Decrease (if % is negative) in PC5 for every one unit increase in Scaled F_ILL.
- MARG_CL_0_3_F:1.97%: There is a 14.043% Increase (if % is positive) or Decrease (if % is negative) in PC5 for every one unit increase in Scaled MARG_CL_0_3_F.
- Dist.Code:1.85%: There is a 13.617% Increase (if % is positive) or Decrease (if % is negative) in PC5 for every one unit increase in Scaled Dist. Code.
- MARG_OT_M:1.66%: There is a 12.877% Increase (if % is positive) or Decrease (if % is negative) in PC5 for every one unit increase in Scaled MARG_OT_M.
- M_ILL:1.53%: There is a -12.36% Increase (if % is positive) or Decrease (if % is negative) in PC5 for every one unit increase in Scaled M_ILL.
- MARGWORK_0_3_M:1.41%: There is a 11.863% Increase (if % is positive) or Decrease (if % is negative) in PC5 for every one unit increase in Scaled MARGWORK_0_3_M.

- MARG_HH_0_3_F:1.11%: There is a 10.554% Increase (if % is positive) or Decrease (if % is negative) in PC5 for every one unit increase in Scaled MARG_HH_0_3_F.

Inferences for PC6:

==========

The feature explaining most that the variance for PC6 is, MAIN_HH_F, it explains 18.95% of the variance in PC6.

There is a 43.531% Increase (if % is positive) or Decrease (if % is negative) in PC6 for every one unit increase in Scaled MAIN_HH_F

==========

- Features explaining ~90% of the variance of PC6 in descending order are:
- MAIN_HH_F:18.95%: There is a 43.531% Increase (if % is positive) or Decrease (if % is negative) in PC6 for every one unit increase in Scaled MAIN_HH_F.
- MARG_OT_3_6_F:15.33%: There is a 39.154% Increase (if % is positive) or Decrease (if % is negative) in PC6 for every one unit increase in Scaled MARG_OT_3_6_F.
- MARG_HH_F:13.83%: There is a 37.188% Increase (if % is positive) or Decrease (if % is negative) in PC6 for every one unit increase in Scaled MARG_HH_F.
- MARG_OT_0_3_F:9.09%: There is a 30.15% Increase (if % is positive) or Decrease (if % is negative) in PC6 for every one unit increase in Scaled MARG_OT_0_3_F.
- MAIN_HH_M:3.66%: There is a 19.13% Increase (if % is positive) or Decrease (if % is negative) in PC6 for every one unit increase in Scaled MAIN_HH_M.
- M_06:2.32%: There is a -15.228% Increase (if % is positive) or Decrease (if % is negative) in PC6 for every one unit increase in Scaled M_06.
- F_06:2.23%: There is a -14.923% Increase (if % is positive) or Decrease (if % is negative) in PC6 for every one unit increase in Scaled F_06.
- MARG_HH_3_6_M:2.1%: There is a -14.489% Increase (if % is positive) or Decrease (if % is negative) in PC6 for every one unit increase in Scaled MARG_HH_3_6_M.
- MARG_AL_M:2.04%: There is a -14.283% Increase (if % is positive) or Decrease (if % is negative) in PC6 for every one unit increase in Scaled MARG_AL_M.
- MARGWORK_3_6_M:1.75%: There is a -13.214% Increase (if % is positive) or Decrease (if % is negative) in PC6 for every one unit increase in Scaled MARGWORK_3_6_M.
- MARG_HH_0_3_M:1.61%: There is a -12.694% Increase (if % is positive) or Decrease (if % is negative) in PC6 for every one unit increase in Scaled MARG_HH_0_3_M.
- MARG_HH_0_3_F:1.52%: There is a -12.316% Increase (if % is positive) or Decrease (if % is negative) in PC6 for every one unit increase in Scaled MARG_HH_0_3_F.
- MARG_OT_3_6_M:1.41%: There is a 11.892% Increase (if % is positive) or Decrease (if % is negative) in PC6 for every one unit increase in Scaled MARG_OT_3_6_M.
- MAIN_CL_F:1.41%: There is a 11.86% Increase (if % is positive) or Decrease (if % is negative) in PC6 for every one unit increase in Scaled MAIN_CL_F.
- MAINWORK_F:1.33%: There is a 11.544% Increase (if % is positive) or Decrease (if % is negative) in PC6 for every one unit increase in Scaled MAINWORK_F.
- F_ST:1.29%: There is a 11.357% Increase (if % is positive) or Decrease (if % is negative) in PC6 for every one unit increase in Scaled F_ST.
- M_ILL:1.28%: There is a -11.302% Increase (if % is positive) or Decrease (if % is negative) in PC6 for every one unit increase in Scaled M_ILL.
- MARG_HH_M:1.23%: There is a 11.08% Increase (if % is positive) or Decrease (if % is negative) in PC6 for every one unit increase in Scaled MARG_HH_M.

- M_ST:1.23%: There is a 11.076% Increase (if % is positive) or Decrease (if % is negative) in PC6 for every one unit increase in Scaled M_ST.
- MARG_CL_F:1.18%: There is a 10.882% Increase (if % is positive) or Decrease (if % is negative) in PC6 for every one unit increase in Scaled MARG_CL_F.
- MARG_AL_3_6_F:1.18%: There is a 10.876% Increase (if % is positive) or Decrease (if % is negative) in PC6 for every one unit increase in Scaled MARG_AL_3_6_F.
- MARGWORK_3_6_F:1.13%: There is a -10.607% Increase (if % is positive) or Decrease (if % is negative) in PC6 for every one unit increase in Scaled MARGWORK_3_6_F.
- MARG_AL_F:1.09%: There is a -10.457% Increase (if % is positive) or Decrease (if % is negative) in PC6 for every one unit increase in Scaled MARG_AL_F.
- MARG_AL_0_3_F:1.08%: There is a 10.413% Increase (if % is positive) or Decrease (if % is negative) in PC6 for every one unit increase in Scaled MARG_AL_0_3_F.

2.3.6 Write linear equation for first PC

Below is given the linear equation in the format PC1 = c1x1+c2x2.... +ci*xi, where the c1, c2,…, ci are the coefficients, given by the eigenvectors and x1, x2,…, xi are the values of the actual columns, e.g. 'State Code','No_HH','TOT_F','M_SC' etc.

'PC1 = (0.03007*State Code) + (0.03008*Dist. Code) + (0.15643*No_HH) + (0.16704*TOT_M) + (0.1657*TOT_F) + (0.16187*M_06) + (0.16227*F_06) + (0.15107*M_SC) + (0.15148*F_SC) + (0.02766*M_ST) + (0.02866*F_ST) + (0.16203*M_LIT) + (0.14712*F_LIT) + (0.16135*M_ILL) + (0.16522*F_ILL) + (0.15999*TOT_WORK_M) + (0.14648*TOT_WORK_F) + (0.14645*MAINWORK_M) + (0.1247*MAINWORK_F) + (0.10284*MAIN_CL_M) + (0.07464*MAIN_CL_F) + (0.11376*MAIN_AL_M) + (0.07479*MAIN_AL_F) + (0.13128*MAIN_HH_M) + (0.0836*MAIN_HH_F) + (0.12379*MAIN_OT_M) + (0.1115*MAIN_OT_F) + (0.16414*MARGWORK_M) + (0.15526*MARGWORK_F) + (0.08147*MARG_CL_M) + (0.04841*MARG_CL_F) + (0.12817*MARG_AL_M) + (0.11446*MARG_AL_F) + (0.14027*MARG_HH_M) + (0.12742*MARG_HH_F) + (0.15515*MARG_OT_M) + (0.14741*MARG_OT_F) + (0.16471*MARGWORK_3_6_M) + (0.16121*MARGWORK_3_6_F) + (0.16509*MARG_CL_3_6_M) + (0.15562*MARG_CL_3_6_F) + (0.09213*MARG_AL_3_6_M) + (0.05078*MARG_AL_3_6_F) + (0.12819*MARG_HH_3_6_M) + (0.11091*MARG_HH_3_6_F) + (0.13903*MARG_OT_3_6_M) + (0.12433*MARG_OT_3_6_F) + (0.1542*MARGWORK_0_3_M) + (0.14641*MARGWORK_0_3_F) + (0.14944*MARG_CL_0_3_M) + (0.13971*MARG_CL_0_3_F) + (0.05165*MARG_AL_0_3_M) + (0.04097*MARG_AL_0_3_F) + (0.12125*MARG_HH_0_3_M) + (0.11579*MARG_HH_0_3_F) + (0.13926*MARG_OT_0_3_M) + (0.13187*MARG_OT_0_3_F) + (0.15022*NON_WORK_M) + (0.13118*NON_WORK_F)'