**PROJECT ON**

**CLUSTERING AND PRINCIPAL COMPONENT ANALYSIS(PCA)**

**CONTENTS:**

|  |  |  |
| --- | --- | --- |
|  | **CLUSTERING**  **ads24x7 - Digital Marketing company** |  |
|  |  |  |
|  | Read the data and perform basic analysis such as printing a few rows (head and tail), info, data summary, null values duplicate values, etc. | 7 |
|  | Treat missing values in CPC, CTR and CPM using the formula given. | 9 |
|  | Check if there are any outliers. Do you think treating outliers is necessary for K-Means clustering? Based on your judgement decide whether to treat outliers and if yes, which method to employ | 10 |
|  | Perform z-score scaling and discuss how it affects the speed of the algorithm. | 10 |
|  | Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance. | 11 |
|  | Make Elbow plot (up to n=10) and identify optimum number of clusters for k-means algorithm. | 14 |
|  | Print silhouette scores for up to 10 clusters and identify optimum number of clusters. | 14 |
|  | Profile the ads based on optimum number of clusters using silhouette score and your domain understanding | 15 |
|  | Conclude the project by providing summary of your learnings. | 17 |
|  |  |  |
|  | **PCA**  **Primary census** | 20 |
|  |  |  |
|  | Read the data and perform basic checks like checking head, info, summary, nulls, and duplicates, etc. | 22 |
|  | Perform detailed Exploratory analysis by creating certain questions like (i) Which state has highest gender ratio and which has the lowest? (ii) Which district has the highest & lowest gender ratio | 25 |
|  | We choose not to treat outliers for this case. Do you think that treating outliers for this case is necessary? | 28 |
|  | Scale the Data using z-score method. Does scaling have any impact on outliers? Compare boxplots before and after scaling and comment. | 28 |
|  | Perform all the required steps for PCA (use sklearn only) Create the covariance Matrix Get eigen values and eigen vector. | 30 |
|  | Identify the optimum number of PCs (for this project, take at least 90% explained variance). Show Scree plot. | 32 |
|  | Compare PCs with Actual Columns and identify which is explaining most variance. Write inferences about all the principal components in terms of actual variables. | 33 |
|  | Write linear equation for first PC. | 37 |

**LIST OF TABLES:**

|  |  |  |
| --- | --- | --- |
| S.no | Name of the table | Pg.no |
|  | First five rows of Ads data | 7 |
|  | Basic information of the Ads data | 8 |
|  | Null Values in Ads data | 9 |
|  | After treatment of Null values in Ads data | 10 |
|  | Scaled data of Ads data | 11 |
|  | After Hierarchical cluster for Ads data | 13 |
|  | Kmeans data with averages | 15 |
|  | First five rows of census data | 22 |
|  | Basic information of the census data | 23 |
|  | Statistics of census data | 24 |
|  | Gender ration vs State | 25 |
|  | Area vs Gender ration | 26 |
|  | State vs ST population | 26 |
|  | State vs SC population | 27 |
|  | State vs Literates | 28 |
|  | Eigen Vector for scaled data | 31 |
|  | Eigen Values for scaled data | 31 |
|  | Eigen values for each component | 32 |
|  | PCA applied census data | 32 |
|  | Cumulative Explained ratio | 33 |
|  | PCA of census data | 33 |
|  | Final PCA applied on census data | 35 |
|  | Scaled census data | 36 |

**LIST OF FIGURES:**

|  |  |  |
| --- | --- | --- |
| S.no | Name of the Figure | Pg.no |
|  | Dendrogram for 10 clusters | 12 |
|  | Dendrogram for 4 clusters | 12 |
|  | Dendrogram for 5 clusters | 13 |
|  | Elbow plot for ads data | 14 |
|  | Cluster vs various parameters | 16 |
|  | Boxplot of 5 variables of census data before scaling | 29 |
|  | Boxplot of 5 variables of census data after scaling | 29 |
|  | Scree plot for census data | 32 |
|  | Loading of PC1, PC2, PC3, PC4 | 34 |
|  | Loading of PC5, PC6, PC7, PC8 | 34 |
|  | Final correlation after PCA | 36 |

**CLUSTERING:**

**K-MEANS & HIERARCHICAL**

**Digital Ads Data:**

The ads24x7 is a Digital Marketing company that 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 want you (their newly appointed data analyst) to segment type of ads based on the features provided. Use the Clustering procedure to segment ads into homogeneous groups.

|  |  |  |
| --- | --- | --- |
| **Sl. No** | **Column Name** | **Column Description** |
| 1 | Timestamp | The Timestamp of the particular Advertisement. |
| 2 | Inventory Type | 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 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. |

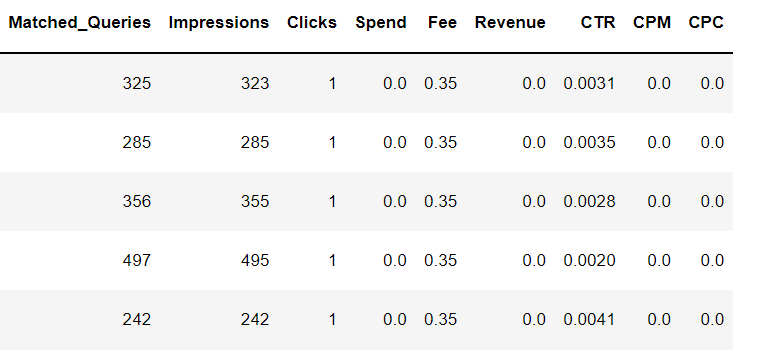
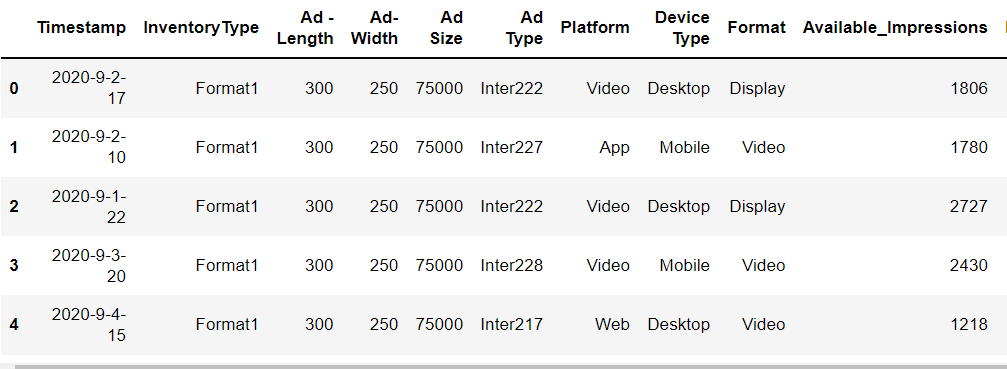
1. **Read the data and perform basic analysis such as printing a few rows (head and tail), info, data summary, null values duplicate values, etc.**

First, we upload data into and import the required packages and set the working directory.

The data set has 23066 rows and 19 columns, which can be identified by using the shape attribute.

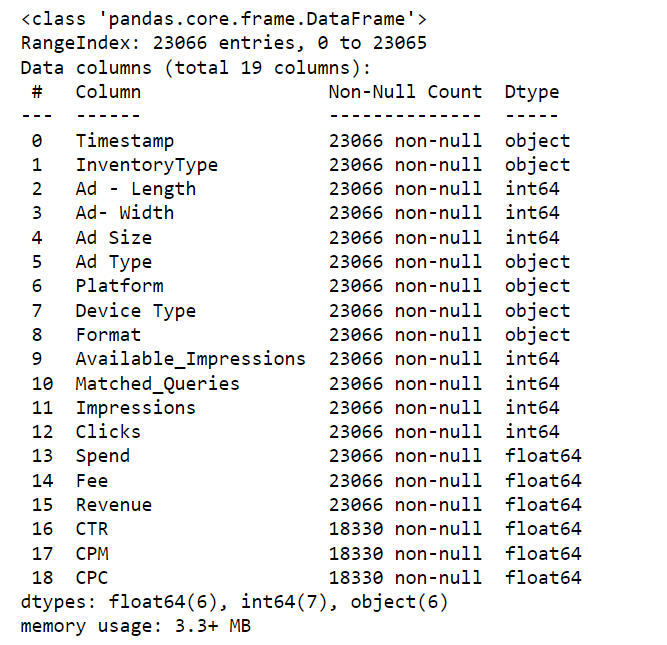
Before starting, we should have a glance at the data in the data set by using head () function.

**Table 1: First five rows of Ads data**



Let’s check the basic information of the dataset using the info () function which shows the data types of columns, the number of non-null values in the column.

**Table 2: Basic information the Ads data**

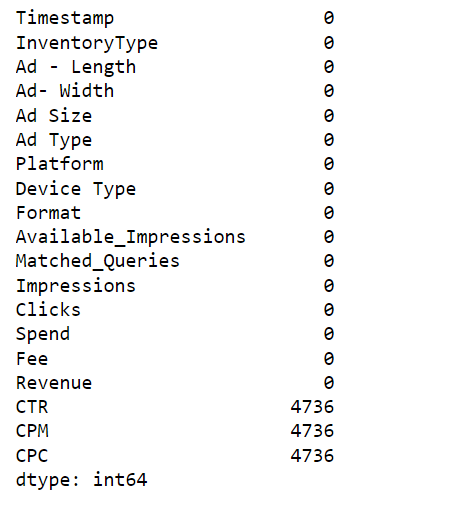


From a quick look at the information, we can see there are 6 categorical or object-type data and 13 numerical data columns.

Also, we can observe that CTR, CPM, CPC are having few null values.

Let’s confirm it by using is null () and sum() functions.

**Table 3: Null values in Ads data**



As we see there are 4,736 null values in CTR, CPM, CPC we should treat these null values.

1. **Treat missing values in CPC, CTR and CPM**

To treat these values, we should use following formula

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

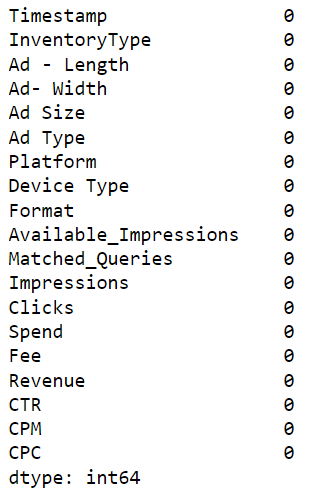
**CPC = Total Cost (spend) / Number of Clicks**

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

Here we will not treat only the missing values, but the above formula will be applied to the entire column.

After applying the formula, the null values will be treated.

**Table 4: After treatment for Null-values in Ads data**



Now there are no null values in the data set.

### Check if there are any outliers. Do you think treating outliers is necessary for K-Means clustering? Based on your judgment decide whether to treat outliers and if yes, which method to employ.

The cause of outliers can be data corruption or failure to record data. The handling of outliers is very important during the data preprocessing pipeline as the presence of outliers can prevent the model to perform best.

Here we could see that there are outliers. As we are dealing with K-Means. The k-Means algorithm does not give the best results. It is sensitive to outliers. K-Means deals with the averages, which can get affected by the outliers. Hence here we are treating the outliers.

Here the data points which are above the upper limit are limited to that upper limit point and data points that are less than the lower limit are limited to the lower limit itself. Therefore, the method I used is to limit the extreme values to upper and lower limits.

### Perform z-score scaling and discuss how it affects the speed of the algorithm.

Scaling is important as:

All such distance-based algorithms are affected by the scale of the variables.

The most common way is to calculate the Euclidean distance

We do not want our algorithm to be affected by the magnitude of these variables. The algorithm should not be biased towards variables with higher magnitude.

To overcome this problem, we can bring down all the variables to the same scale.

If we could see the data which is not scaled, the data is of different scales. The scale of Ad-length is in hundreds, Ad-size is in Thousands, and some other columns are with very less scales.

This affects the Euclidean distance and the weights may differ.

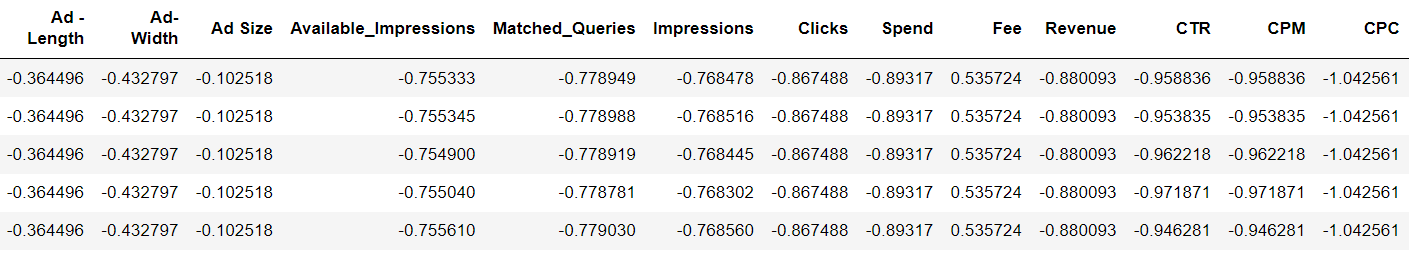
Before performing the z-score, I created another data frame with only the numerical data.

I have dropped the categorical data and time stamp column.

The new data frame is cluster\_df.

After applying the z-score this scaled data is load into a new data frame df\_scaled.

**Table 5: Scaled data of Ads data set**



Now the complete data is scaled.

### Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance

### For performing the Hierarchical clustering, we should import a few packages.

### The linkage method we are using is Ward linkage.

### Ward’s procedure is a variance method that attempts to generate clusters to minimize the within-cluster variance.

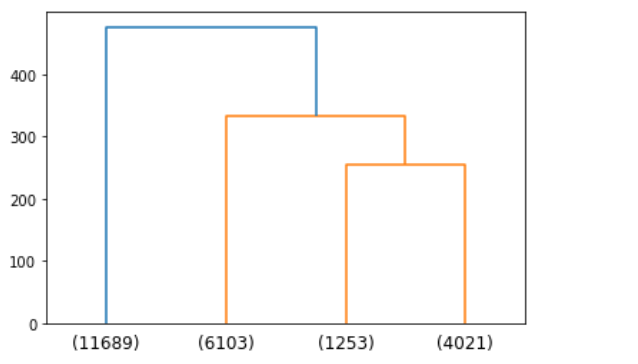
### Let’s construct a dendrogram for ten clusters.

**Figure 1: Dendrogram for 10 clusters**

### 

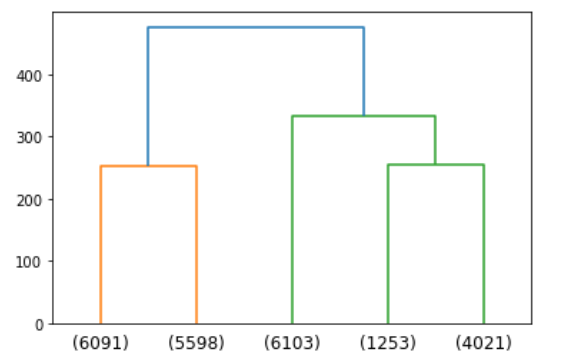
### Check for 4 clusters

**Figure 2: Dendrogram for 4 clusters**



Here we can observe that 1 cluster has 11,689 and other clusters have less data points. So, let’s check for 5 clusters.

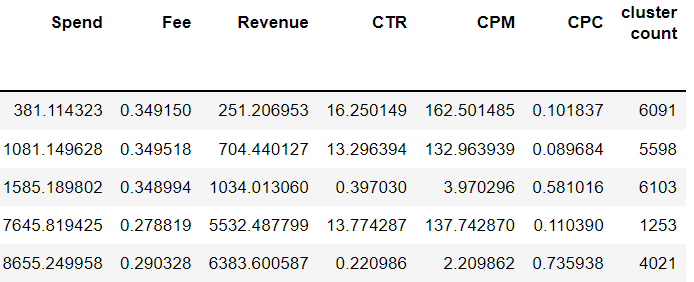
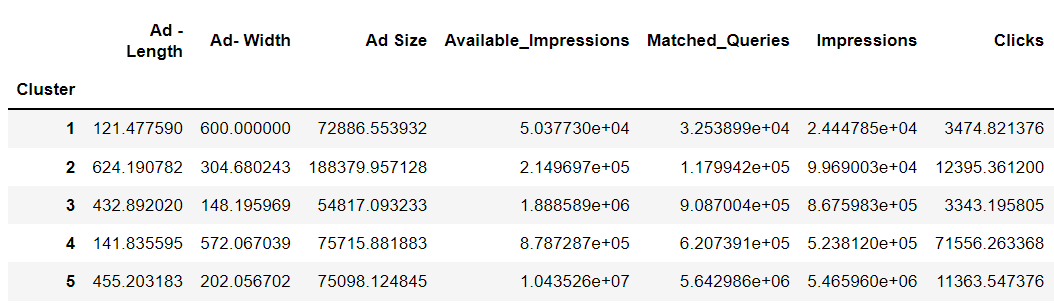
**Figure 3: Dendrogram for 5 clusters**



Now I can see some balancing among the clusters.

So, we can consider 5 clusters by observing the dendrogram.

**Table 6: After Hierarchical Clustering of Ads data**



### In the above table, we got the average data for each cluster. Cluster count is the number of data points in each cluster.

### Make Elbow plot (up to n=10) and identify optimum number of clusters for the k-means algorithm.

### To make Elbow plot, we need KMeans inertia which calculates the within sum of square.

### Inertia measures how well a dataset was clustered by K-Means. It is calculated by measuring the distance between each data point and its centroid, squaring this distance, and summing these squares across one cluster.

### As the number of clusters increases, inertia decreases.

### So, we calculate the within the sum of squares and plot. The Elbow plot is plotted between number of clusters and Kmeans inertia.

### Figure 4: Elbow plot for Ads data

### 

### The Elbow method is used to find the optimum number of clusters while doing K-Means.

### Here 5 is the optimum no. of clusters.

### As the difference or variance is more up to 5 clusters.

### Print silhouette scores for up to 10 clusters and identify the optimum number of clusters

### The silhouette value is a measure of how similar an object is to its own cluster compared to other clusters.

### The silhouette ranges from −1 to +1, where a high value indicates that the object is well-matched to its own cluster.

### The below are the Silhouette scores from 2 clusters to 10 clusters.

### [0.3880914534420389, 0.3842453039186148, 0.4495511197534906, *0.5296776454784944,* 0.527590725438368, 0.5220635854819954, 0.4968722443929237, 0.4790423315852688, 0.4922718574674112]

### As we could see the 4th value is the highest among all the values.

### So, the optimum number of clusters to be selected is 5.

### Profile the ads based on the optimum number of clusters using silhouette score and your domain understanding

### As we know the number of clusters, we apply this to the data set, so that we can know which data point belongs to which cluster.

### After applying, let’s use groupby() function to check the data according to the clusters.

### To this apply mean() function to get the average value of all the columns for each cluster.

### Table 7: KMeans Data with averages

### 

### 

**Figure 5: Cluster vs Various parameters**

### 

### 

1. **Conclude the project by providing summary of your learnings**

### Cluster 1:

### Ad length is lowest and Ad size is also low

### Money spent on Ads is the lowest

### Revenue is the lowest

### Clicks are also low, that is people are not clicking more.

### The impression rate is very low.

### CPM is the cost per 1000 impressions. The money spent/ impressions are calculated. Here CPM is highest. This could be negative

### CPC is the cost per click. That is, you pay for each click, which is also low.

### CTR is Click through rate is the highest

### Cluster 2:

### length the is highest and Ad size is also the highest

### Money spent on Ads is low

### Revenue is low

### Clicks are moderate (second highest), that is people are showing interest in these ads and clicking more.

### The impression rate is low.

### CPM is the cost per 1000 impressions. The money spent impressions is calculated. Here CPM is moderate.

### CPC is the cost per click. That is, you pay for each click, which is very low.

### CTR is Click through rate is moderate

### Cluster 3:

### Length is moderate and Ad size is also moderate

### Money spent on Ads is moderate

### Revenue is moderate

### Clicks are the lowest, that is people are not showing interest in these ads.

### The impression rate is moderate.

### CPM is the cost per 1000 impressions. The money spent on impressions is calculated. Here CPM is very low.

### CPC is the cost per click. That is, you pay for each click, which is moderate.

### CTR is Click through rate is very low.

### Cluster 4:

### length is lowest and Ad size is moderate.

### Money spent on Ads is high

### Revenue is high

### Clicks are highest, that is people are clicking more that is the ads are interesting.

### The impression rate is low.

### CPM is the cost per 1000 impressions. The money spent on impressions is calculated. Here CPM is high.

### CPC is the cost per click. That is, you pay for each click, which is also very low.

### CTR is Click through rate is high.

### Cluster 5:

### Ad length is high and Ad size is high.

### Money spent on Ads is the highest

### Revenue is the highest

### Clicks are moderate, that is people are clicking is the ads are a little interesting.

### The impression rate is the highest.

### CPM is the cost per 1000 impressions. The money spent on impressions is calculated. Here CPM is the lowest.

### CPC is the cost per click. That is, you pay for each click, which is also highest.

### CTR is Click through rate is the lowest

### From the above conclusion about each cluster, we could get an idea about the clusters.

### According to spend and revenue, if we compare them: This is the difference between spending and revenue(income)

### cluster 1 = 129.90737000000001

### cluster 2 = 376.70950099999993

### cluster 3 = 551.1767420000001

### cluster 4 = 2113.3316259999992

### cluster 5 = 2271.6493710000004

**PRINCIPAL COMPONENT ANALYSIS**

Primary census abstracts 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.

**Data Dictionary:**

|  |  |
| --- | --- |
| 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 |

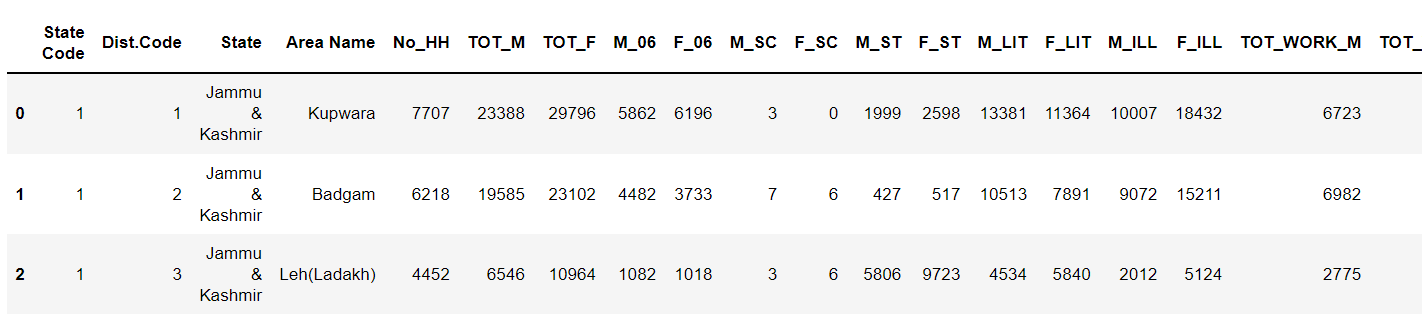
1. **Read the data and perform basic checks like checking head, info, summary, nulls, and duplicates, etc.**

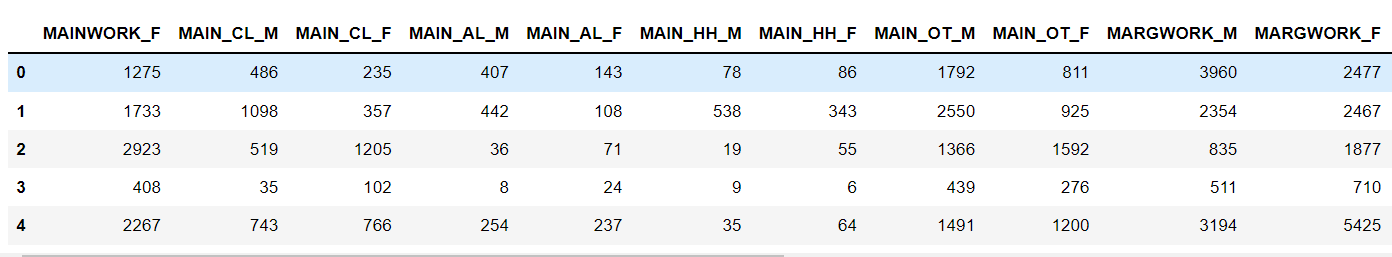
First, we upload data into and import the required packages and set the working directory.

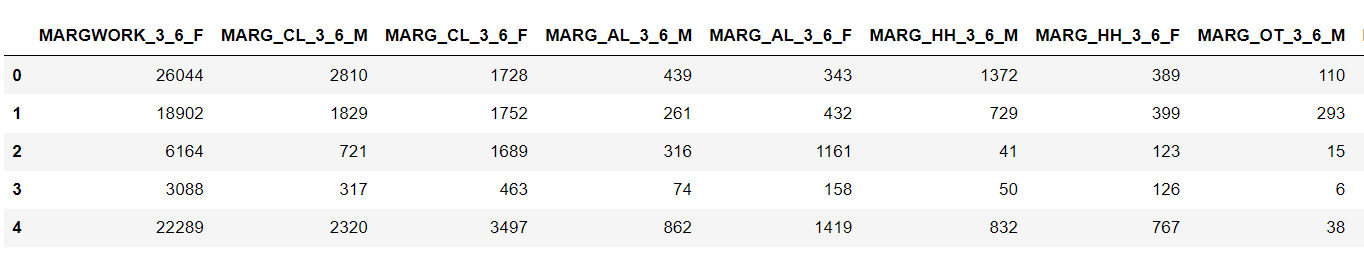
The data set has 640 rows and 16 columns, which can be identified by using the shape attribute.

Before starting, we should have a glance at the data in the data set by using head() function.

**Table 8: First five rows and few columns of census data**

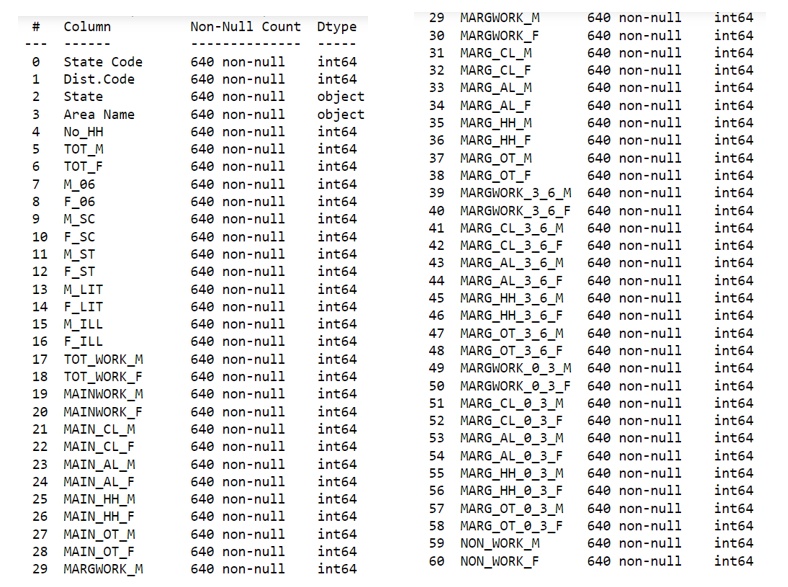






Let’s check the basic information of the dataset using the info() function which shows the data types of columns, the number of non-null values in the column.

**Table 9: Basic information on census data**

****

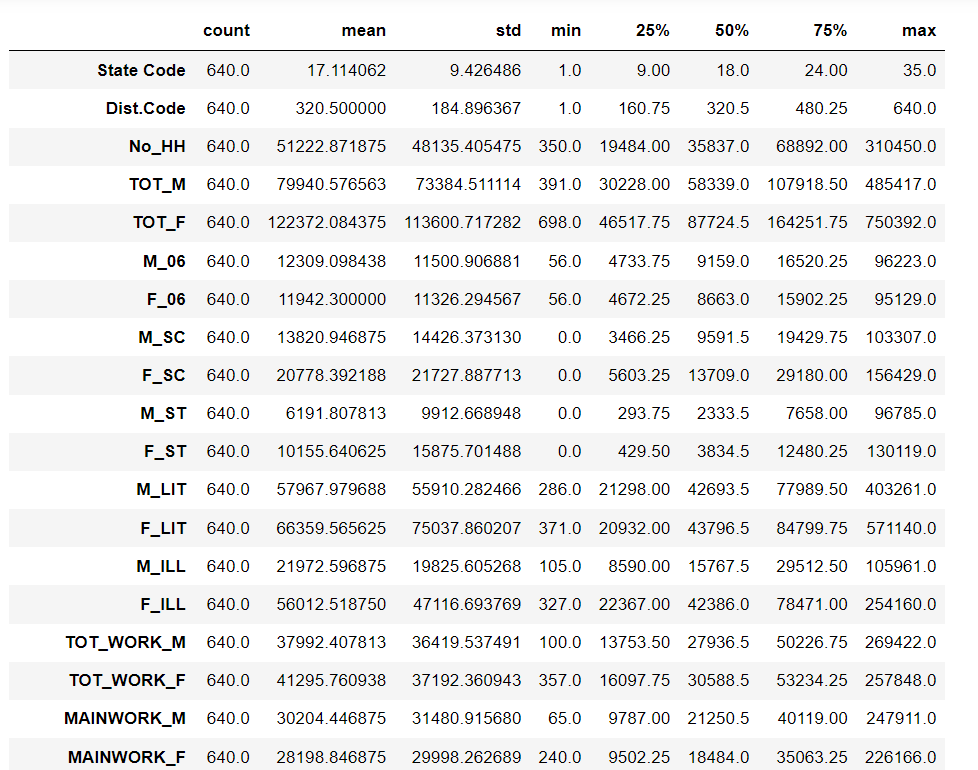
From the above table, we see there are no null-values in the data set and all are numerical values except Area Name and State are categorical data types.

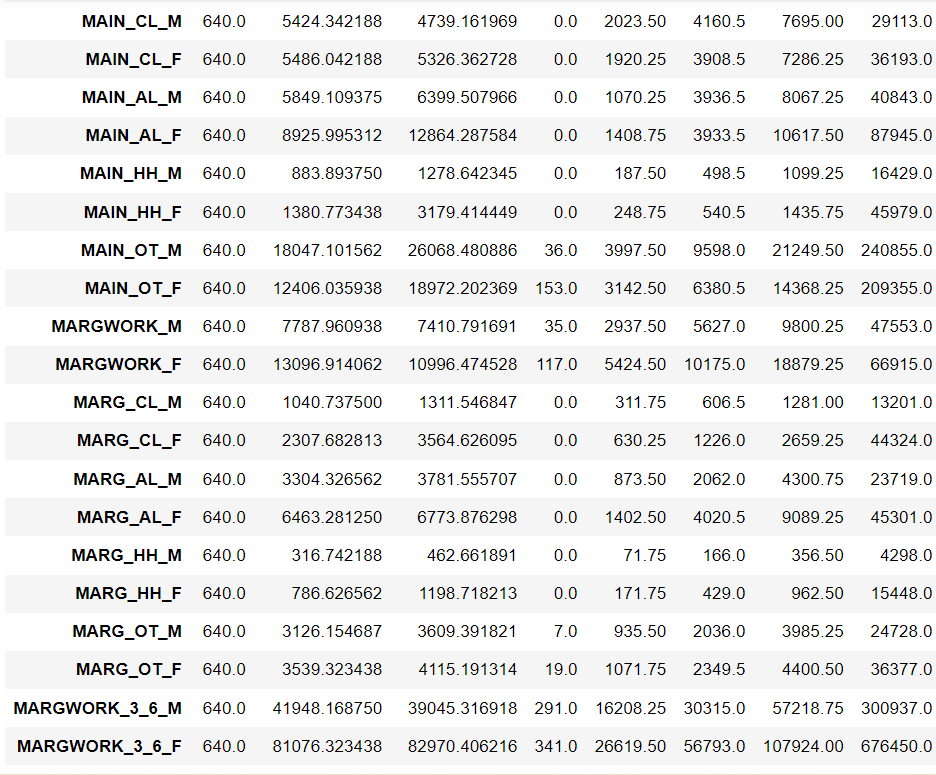
As they are not more important to us to perform PCA. We drop a few columns which are not required while performing PCA

I have dropped State code, Dist. Code, State, Area name and stored in a new data frame for further process.

Let’s check the basic statistics of the data by using describe() function.

**Table 10: Statistics of census data**





### 

### Perform detailed Exploratory analysis by creating certain questions Which state has highest gender ratio and which has the lowest?

Dadara & Nagar Havelli state has minimum population of male and female.

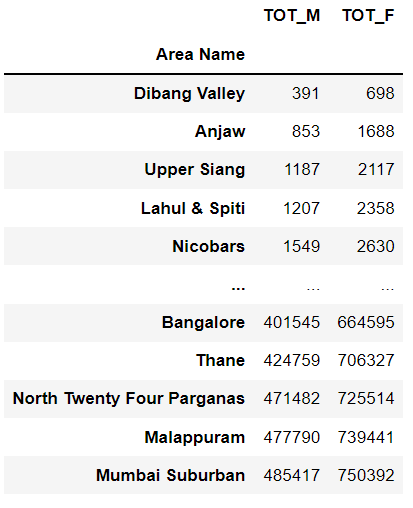
Uttar Pradesh state has maximum population of male and female.

**Table 11: Gender ratio with resp. to State**

### 

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

**Table 12: Area vs Gender ration**



Dibang Valley has lowest Male and Female population.

Mumbai Suburban has the highest Male and Female population.

### (iii) Which state has the highest & lowest ST population?

**Table 13: State vs ST population**

### 

Punjab, Puducherry, NCT of Delhi, Chandigarh, Haryana have zero or lowest Scheduled Tribe population

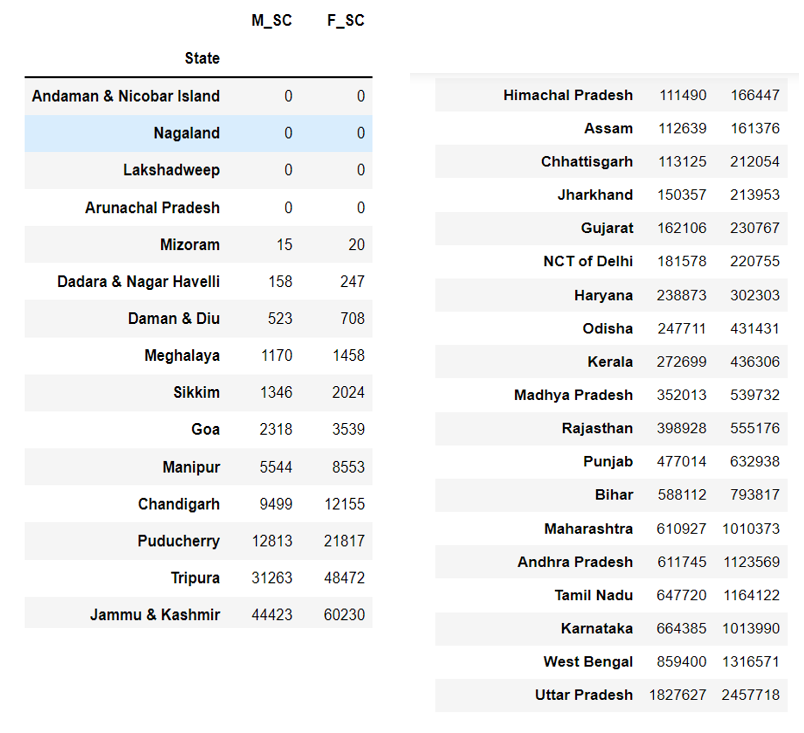
Madhya Pradesh has highest Scheduled tribe population.

### (iv) Which state has the highest & lowest SC population?

Andaman & Nicobar Island, Arunachal Pradesh, Lakshadweep, Nagaland has minimum Scheduled Castes population.

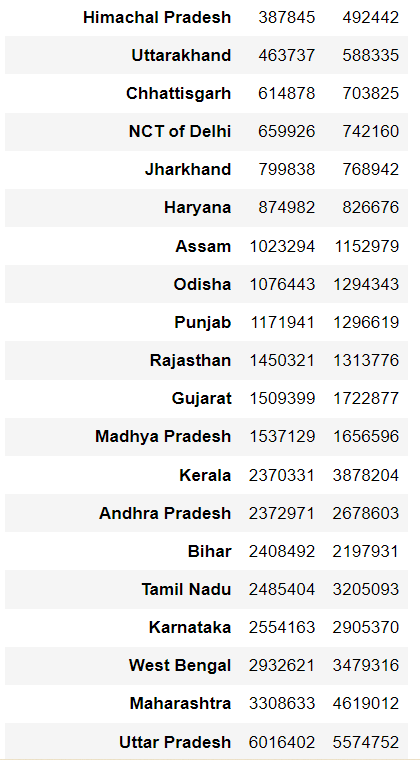
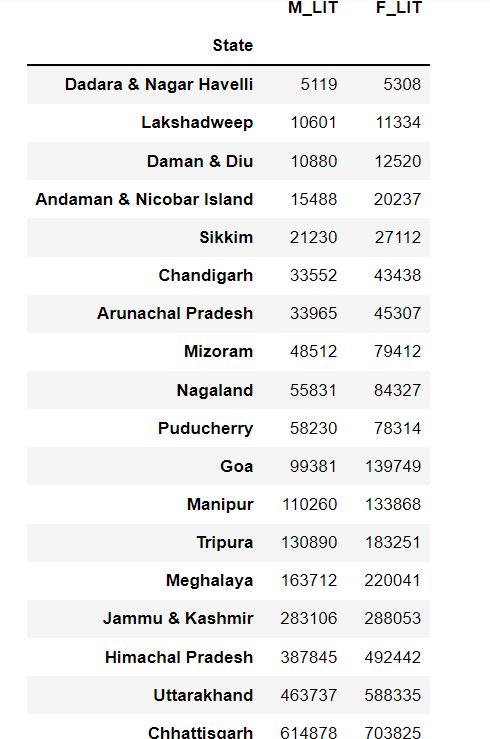
Uttar Pradesh has highest Scheduled Caste population.

**Table 14: State vs SC population**



**(v) Which state has the highest & lowest literates’ population?**

**Table 15: State vs Literates**

****

Dadara & Nagar Havelli has lowest number of literates

Uttar Pradesh has highest literates.

1. **We choose not to treat outliers for this case. Do you think that treating outliers for this case is necessary?**

As the data set is about the census, that is about the population it is not necessary to remove the outliers.

Hence, we are not treating the outliers here.

### Scale the Data using z-score method. Does scaling have any impact on outliers? Compare boxplots before and after scaling and comment.

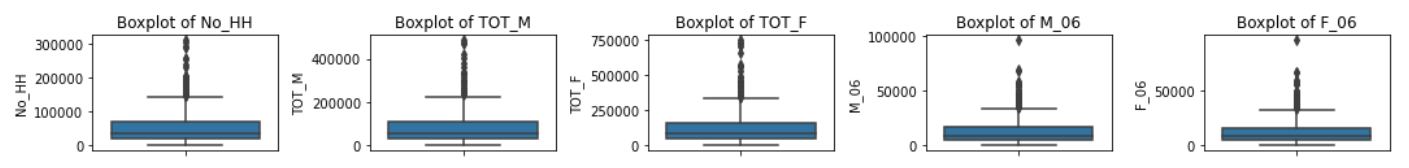
Algorithms that use a weighted sum of the input, like linear regression, and algorithms that use distance measures, like k-nearest neighbors perform better when numerical input variables are scaled to a standard range.

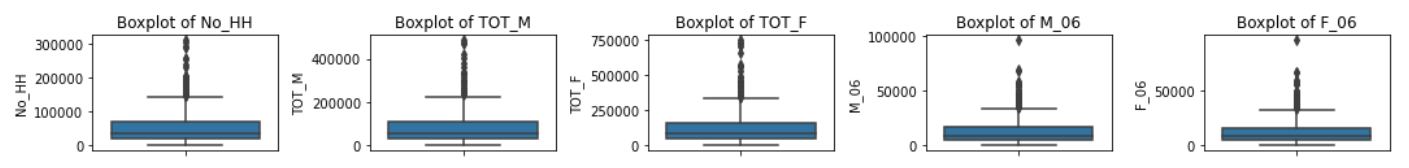
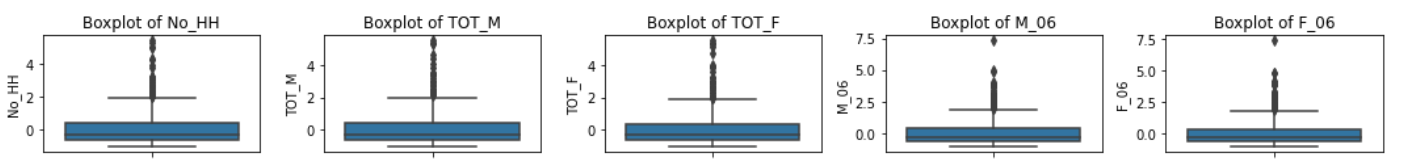
Standardizing or Z-Score is a popular scaling technique that subtracts the mean from values and divides by the standard deviation.

Standardization can become skewed or biased if the input variable contains outlier values.

But here as we are not treating outliers, let’s check the outliers after scaling.

For reference, let’s check for first 5 variables as there are 57 variables.

**Figure 6: Boxplot of 5 variables of census data before scaling**

**Figure 7: Boxplot of 5 variables of census data after scaling**

As we can see there is no affect on outliers due to scaling.

### Perform all the required steps for PCA (use sklearn only) Create the covariance Matrix Get eigen values and eigen vector

**BARTLETT'S TEST FOR SPHERICITY**

**check for correlation**

This helps to find if there is any redundancy present between the variables.

To calculate this, we should import calculate\_bartlett\_sphericity module from factor\_analyzer.

For this we get, chi-square value and p value.

The p-value must be less than significant value 𝛼α

Here 𝛼α = 0.05

H0: There are no significant correlations

Ha: There are significant correlations

We Reject H0 if p-value < 0.05

**The P-value is less than 0.05. Therefore, we reject null hypothesis.**

**By this, we can understand that there is a significant correlation between variables.**

**KAISER-MEYER-OLKIN TEST**

**check the adequacy of the sample size** by **KMO**

This helps in finding the Measure of Sample Adequacy (MSA)

The KMO value must be greater than 0.7.

The kmo\_model represent the KMO value for complete variables, kmo\_all gives the values for individual variables.

The value of kmo\_model is 0.8 which indicates that the component analysis will be useful for these variables.

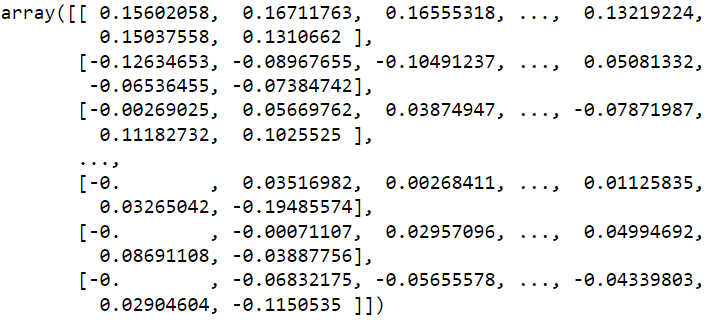
### Apply PCA

Principal components are new variables that are constructed as linear combinations of the initial variables. These combinations are done in such a way that these new variables are uncorrelated and most of the information within the initial variables is stored in the first components.

We import few required packages and apply PCA to the scaled data set.

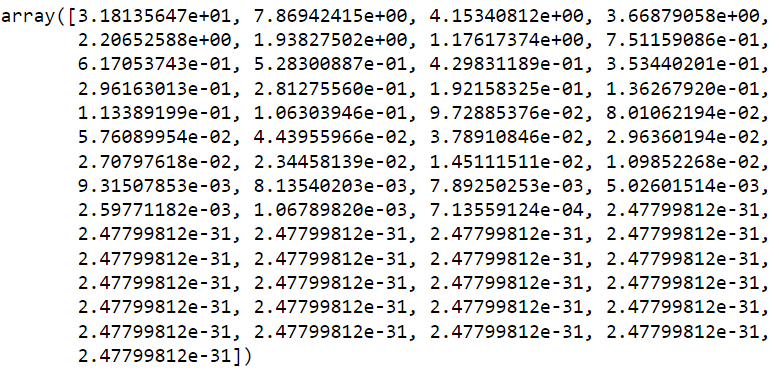
We get Eigen Vectors by using components\_ attribute.

**Table 16: Eigen Vector for scaled census data**

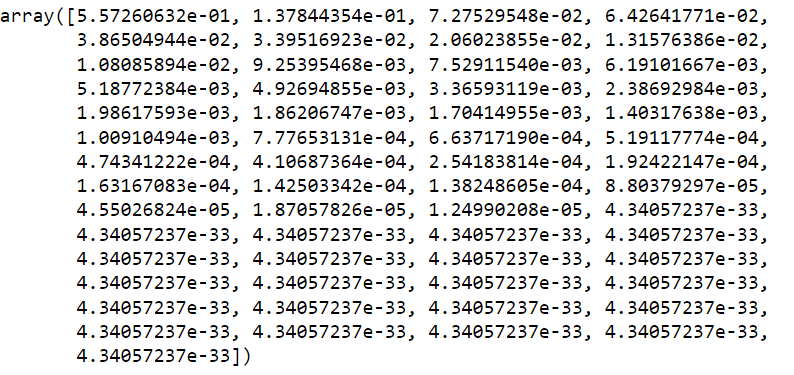


We get Eigen Values by explained\_variance\_ attribute.

**Table 17: Eigen Values for scaled census data**

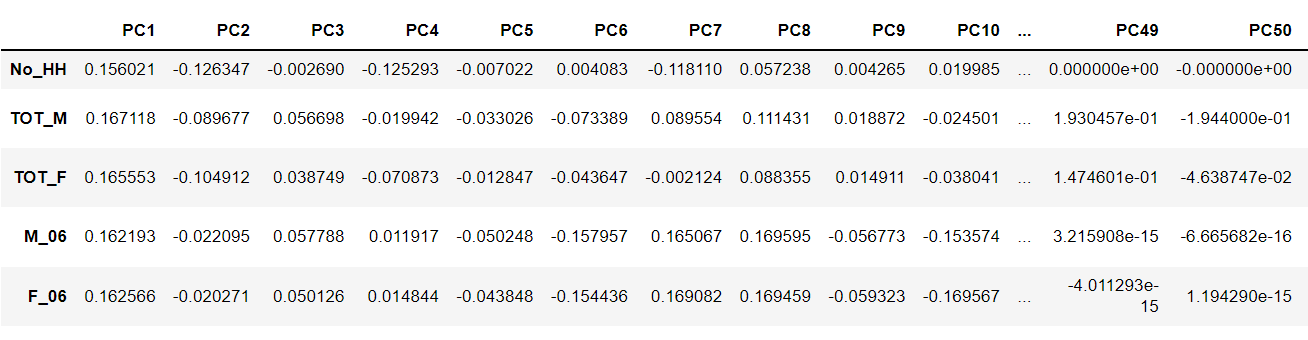


**Table 18: Eigen Value for each component**



We get Eigen value for each component by dividing Eigen value to sum of Eigen values.

**Table 19: PCA applied census data**



### Identify the optimum number of PCs (for this project, take at least 90% explained variance). Show Scree plot.

**Figure 8: Scree Plot for census data**

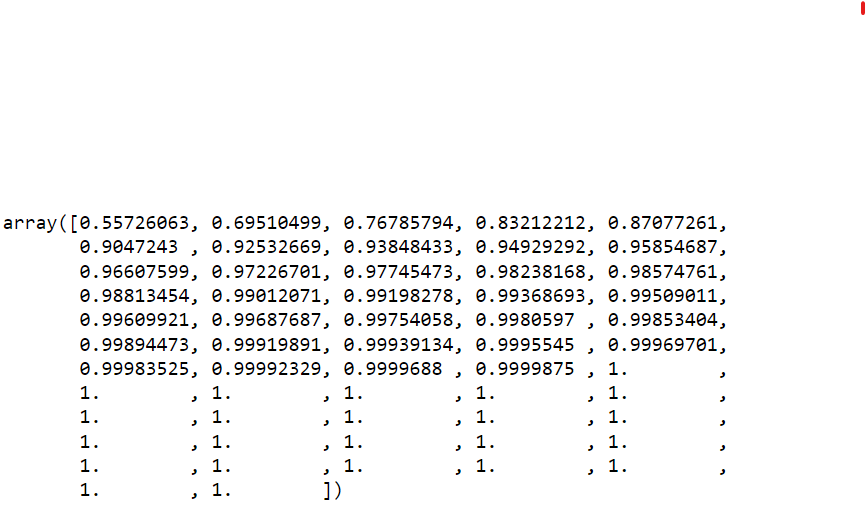
### 

By observing, the scree plot maybe we could consider 8 PCs as the variance is low after 8 PC’s.

**Cumulative explained variance ratio:**

This gives the cumulative of explained variance ratio or Eigen values.

**Table 20: Cumulative Explained Ratio of census data**



After applying PCA to the data lets check the head of the data

**Table 21: PCA of census data**

### 

### Compare PCs with Actual Columns and identify which is explaining most variance. Write inferences about all the principal components in terms of actual variables

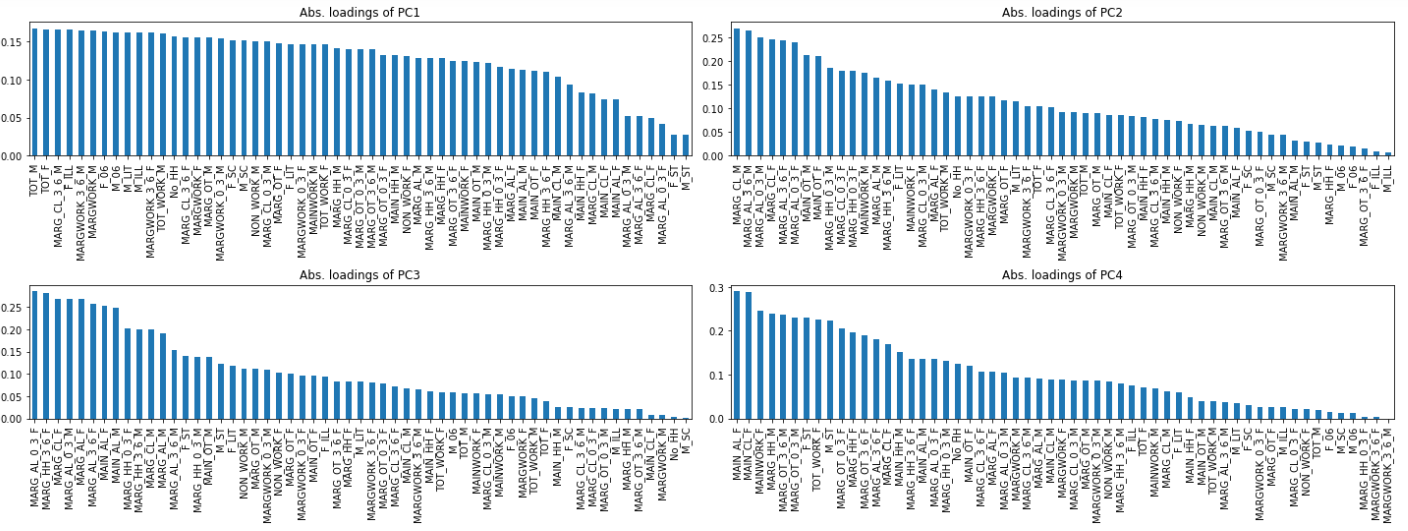
The Principal Components are the linear combination of all the scaled data variables.

Let's check the features that matter for each pc

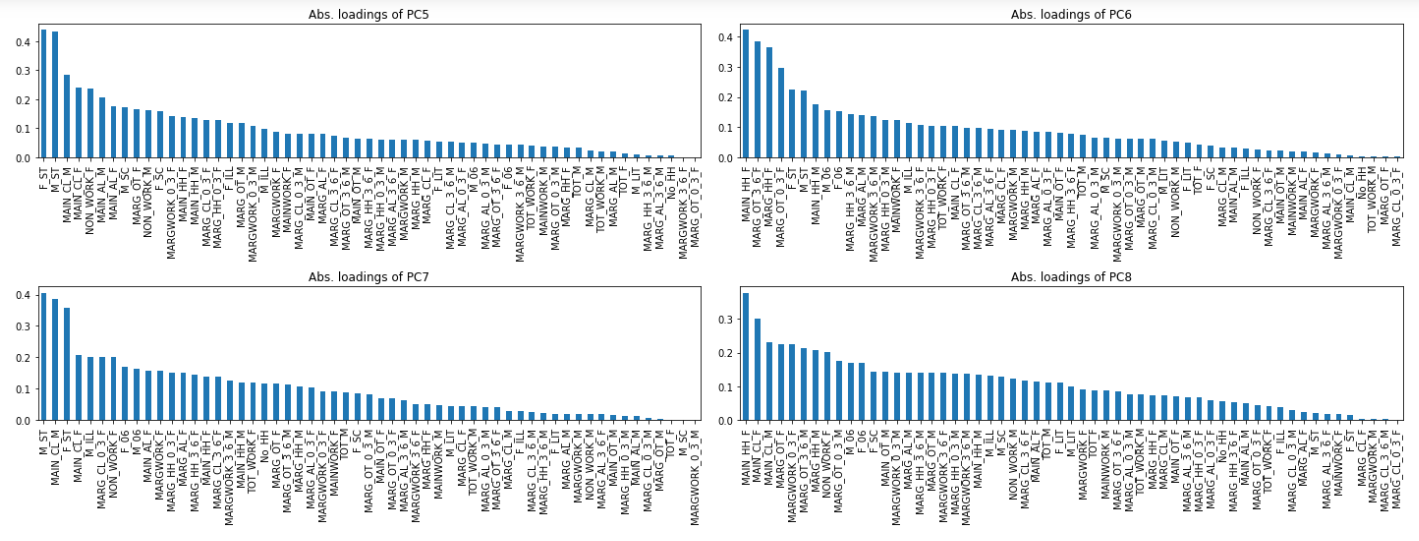
Structure of Principal Components and PC Scores

To get the PC scores, we must do dot product of components and scaled data.

**Figure 9: Loadings of PC1, PC2, PC3, PC4**

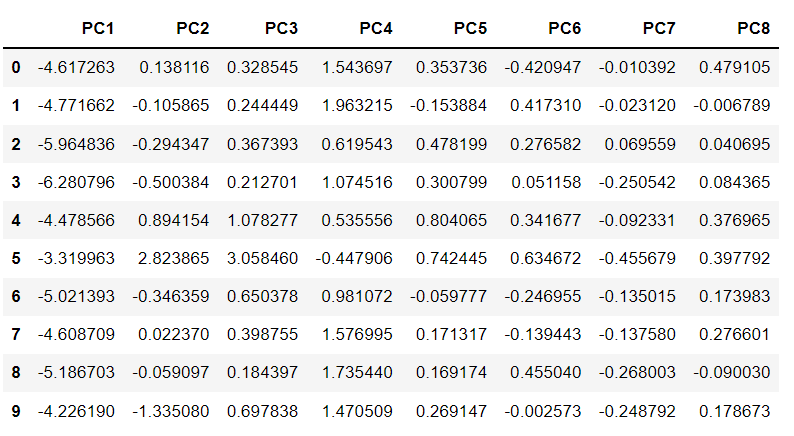


**Figure 10: Loadings of PC5, PC6, PC7, PC8**



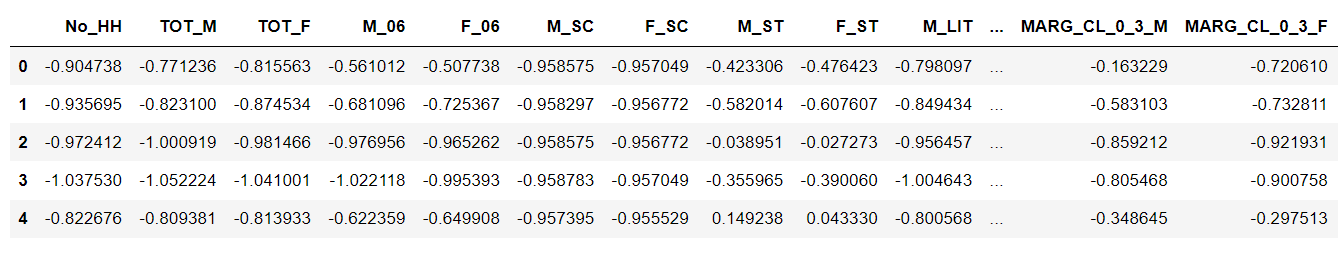
When we compare the above figure with loadings of all PC’s, shows how much that variable matter to the respective component.

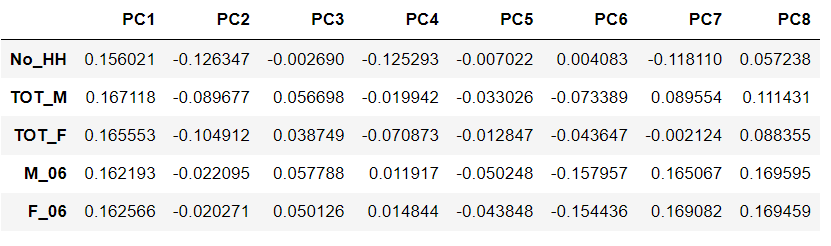
**Table 22: Final PCA applied census data**



This final PCA is obtained by the dot product of the scaled data and PCA scores.

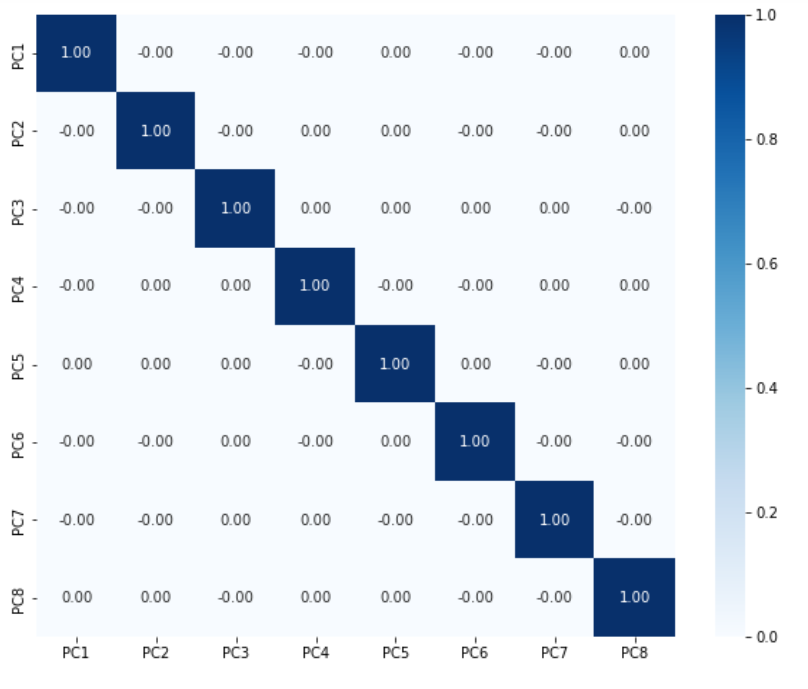
**Table 23: Scaled data of census data**





The dot product of PC1 and scaled data of the first column give the PCA score of the first component.

**Figure 11: Final correlation check after PCA**



The correlation which was present at beginning between the variables is

Now limited and it is changed to identity Matrix.

1. **Write the linear equation for the first PC.**

-0.9047375442726222 X 0.15602057858567936+

-0.7712364536733182 X 0.1671176348853346+

-0.8155625820369515 X 0.16555317909064893+

-0.5610115106758324 X 0.16219294820465552+

-0.5077383897071448 X 0.16256639565734834+

-0.9585745972915195 X 0.15135784909060584+

-0.9570485925256337 X 0.15156650019208875+

-0.42330550019368 X 0.02723419457100421+

-0.47642316565413956 X 0.028183315015872686+

-0.7980974549077947 X 0.16199283733629158+

-0.7334774895508266 X 0.14687268030140285+

-0.604014646963687 X 0.16174944463471633+

-0.7982290346008761 X 0.16524818736833372+

-0.8592603821567523 X 0.1598719881620129+

-1.01023773336172 X 0.14593580377247625+

-0.8723669584689564 X 0.14620072976305987+

-0.8982155430715099 X 0.1239702835727365+

-1.0428436137704937 X 0.1031271588301987+

-0.9866300225342314 X 0.07453978555483678+

-0.8510601190840429 X 0.11335571218156724+

-0.6832764753976269 X 0.0738821590315588+

-0.6307659813495755 X 0.13157258402275604+

-0.40755497183694095 X 0.08338263967435766+

-0.6240415708671648 X 0.12352624192253085+

-0.6116372239374073 X 0.1110212639132013+

-0.5169427506711185 X 0.16461547856011013+

-0.9665116377825472 X 0.15539561810834135+

-0.32180883960166345 X 0.08238854140704549+

-0.48505343961666236 X 0.04919539567873828+

-0.331426055233964 X 0.12859856294668567+

-0.8601921541797595 X 0.11430507278919892+

-0.3779841570524406 X 0.14085322696185143+

-0.4530263917933537 X 0.12766959801475364+

-0.548763563204698 X 0.15526287162311606+

-0.614625185724889 X 0.1472865835652339+

-0.6480404412222549 X 0.16497194993714456+

-0.6637953265683431 X 0.1612534325753136+

-0.5959978069219448 X 0.1655016110258063+

-1.0178478914467044 X 0.15564704914483396+

-0.3877071892324557 X 0.09301420640192852+

-0.5638538557286629 X 0.05153586397015223+

-0.448658286439496 X 0.12857611642867825+

-0.896722522644112 X 0.11064584323696924+

-0.37763475967454396 X 0.1395927625215884+

-0.4313073667481123 X 0.12454590917258754+

-0.5691510732157568 X 0.15429378578916045+

-0.6124513526744523 X 0.14628565406214422+

-0.1632290610313297 X 0.1501257061026207+

-0.7206100966969275 X 0.14015704689010394+

-0.156494116071813 X 0.05254178285396346+

-0.2875244263165451 X 0.04178595301201035+

0.15657747746968836 X 0.12184035387925027+

-0.6574115856750813 X 0.11601141016824112+

-0.3652581632815362 X 0.13986877411042814 +

-0.49997673757504263 X 0.13219224458196538 +

-0.4130525109528059 X 0.15037557804411297+

-0.5396143895404282 X 0.1310662031320733+

By performing this multiplication and addition we get the PC1. The above is the linear equation of PC1.