Data Mining

Table of Contents

[Business Report for the ads24x7 Digital Marketing Company 3](#_Toc120482640)

[Q1. Read the data and performing the basic analysis of the data 4](#_Toc120482641)

[Q2. How to do the Treatment of the missing values in the respective data fields? 4](#_Toc120482642)

[Q3. Check if there are any outliers. Do you think treating outliers is necessary for K-Means clustering? Based on your judgement decide whether to treat outliers and if yes, which method to employ. (As an analyst your judgement may be different from another analyst). 5](#_Toc120482643)

[Q4. Perform z-score scaling and discuss how it affects the speed of the algorithm. 6](#_Toc120482644)

[Q5. Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance. 7](#_Toc120482645)

[Q6.  Make Elbow plot (up to n=10) and identify optimum number of clusters for k-means algorithm. 8](#_Toc120482646)

[Q7.  What are the silhouette scores for up to 10 clusters and identify optimum number of clusters. 9](#_Toc120482647)

[Q8.  Profile the ads based on optimum number of clusters using silhouette score and your domain understanding [Hint: Group the data by clusters and take sum or mean to identify trends in Clicks, spend, revenue, CPM, CTR, & CPC based on Device Type. Make bar plots]. 9](#_Toc120482648)

[Q9. Conclude the project by providing summary of the learning. 12](#_Toc120482649)

[Business Report for the Indian Census report 13](#_Toc120482650)

[Q1. Read the data and performing the basic analysis of the data 14](#_Toc120482651)

[Q2. 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? (Example Questions). Pick 5 variables out of the given 24 variables below for EDA: No\_HH, TOT\_M, TOT\_F, M\_06, F\_06, M\_SC, F\_SC, M\_ST, F\_ST, M\_LIT, F\_LIT, M\_ILL, F\_ILL, TOT\_WORK\_M, TOT\_WORK\_F, MAINWORK\_M, MAINWORK\_F, MAIN\_CL\_M, MAIN\_CL\_F, MAIN\_AL\_M, MAIN\_AL\_F, MAIN\_HH\_M, MAIN\_HH\_F, MAIN\_OT\_M, MAIN\_OT\_F 16](#_Toc120482652)

[Q3. We choose not to treat outliers for this case. Do you think that treating outliers for this case is necessary? 16](#_Toc120482653)

[Q4. Scale the Data using z-score method. Does scaling have any impact on outliers? Compare boxplots before and after scaling and comment. 17](#_Toc120482654)

[Q5. Perform all the required steps for PCA (use sklearn only) Create the covariance Matrix Get eigen values and eigen vector. 18](#_Toc120482655)

# Business Report for the ads24x7 Digital Marketing Company

## Q1. Read the data and performing the basic analysis of the data

The data is having 23066 rows and 19 columns

|  |  |
| --- | --- |
| **Data Field** | **Data Type** |
| Timestamp | object |
| InventoryType | object |
| Ad - Length | int64 |
| Ad- Width | int64 |
| Ad Size | int64 |
| Ad Type | object |
| Platform | object |
| Device Type | object |
| Format | object |
| Available\_Impressions | int64 |
| Matched\_Queries | int64 |
| Impressions | int64 |
| Clicks | int64 |
| Spend | float64 |
| Fee | float64 |
| Revenue | float64 |
| CTR | float64 |
| CPM | float64 |
| CPC | float64 |

There are no duplicate data found in the dataset. But there are 4,736 Null values present in the CTR, CPM, CPC data field.

We are keeping only Impressions, Clicks, Spend, CTR, CPM, CPC and dropping rest of the columns as they are not necessary for our analysis.

## Q2. How to do the Treatment of the missing values in the respective data fields?

For CTR Missing values data we are using formula-based definition

We know that the formula of the CTR

= Total Measure clicks (Clicks)/Total Measure of Ad impressions (Impressions)\*100

For CPC Missing values data we are using formula-based definition

We know that the formula of the CPC

= Total Cost (Spend)/Total Number of clicks (Clicks)

For CPM Missing values data we are using formula-based definition

We know that the formula of the CPM

= Total campaign spends (Spend)/Total Measure of Ad impressions (Impressions)\*100

With the above 3 formulas we are able to treat the missing value.

## Q3. Check if there are any outliers. Do you think treating outliers is necessary for K-Means clustering? Based on your judgement decide whether to treat outliers and if yes, which method to employ. (As an analyst your judgement may be different from another analyst).

Chart, box and whisker chart

Description automatically generated

*Box plot representation for outliers*

As we can see the above plot have has some outliers present in it. But These outliers can be considered as they are not going affect the data.

No need to treat outlier in this case. In general case, of K means clustering we can treat the outlier if the outlier value is going out of limit, and it is extreme.

Q4. Perform z-score scaling and discuss how it affects the speed of the algorithm.  
  
The Z score Scaling is performed to scale all the value in the respective data field to be in between -1 to 1. This is done for us to easily find the correlation between the values and from the correlation we will be able to identify which elements are correlated respectively.

**Table

Description automatically generated**

Before Z score Scaling

**Text

Description automatically generated**

After Z score scaling

***A picture containing chart

Description automatically generated***

Correlation of data field

## Q5. Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance.

We have performed the dendrogram using ward and Euclidean distance

**Chart, histogram, box and whisker chart

Description automatically generated**

Dendrogram using WARD and Euclidean distance

## Q6.  Make Elbow plot (up to n=10) and identify optimum number of clusters for k-means algorithm.

**Chart, line chart

Description automatically generated**

Elbow plot

The Optimum number of clusters for K means algorithm are n=4.

## Q7.  What are the silhouette scores for up to 10 clusters and identify optimum number of clusters.

The Silhouette scores are 0.5385308525008989 and the optimum number of clusters are 4.

## Q8.  Profile the ads based on optimum number of clusters using silhouette score and your domain understanding [Hint: Group the data by clusters and take sum or mean to identify trends in Clicks, spend, revenue, CPM, CTR, & CPC based on Device Type. Make bar plots].

**We have profiled ads the based on the below criteria**

Chart, bar chart

Description automatically generated

Cluster vs Clicks based on Device type

Chart, bar chart

Description automatically generated

Cluster vs Spend based on Device type

Chart, bar chart

Description automatically generated

Cluster vs Revenue based on Device type

Chart, bar chart

Description automatically generated

Cluster vs CPM based on Device type

Chart, bar chart

Description automatically generated

Cluster vs CTR based on Device type

Chart, bar chart

Description automatically generated

Cluster vs CPC based on Device type

## Q9. Conclude the project by providing summary of the learning.

The above dataset can be divided into 4 clusters. Based upon the observation from the cluster. The **Cluster 2 provides more Revenue** Compared to other cluster. The number of clicks is **more for cluster 3.** The amount **Spent** by Ad agency is **more for Cluster 3.**

**The more the advertising agency amount spend on the Advertisement the more they produce Revenue from that ad.**

# Business Report for the Indian Census report

## Q1. Read the data and performing the basic analysis of the data

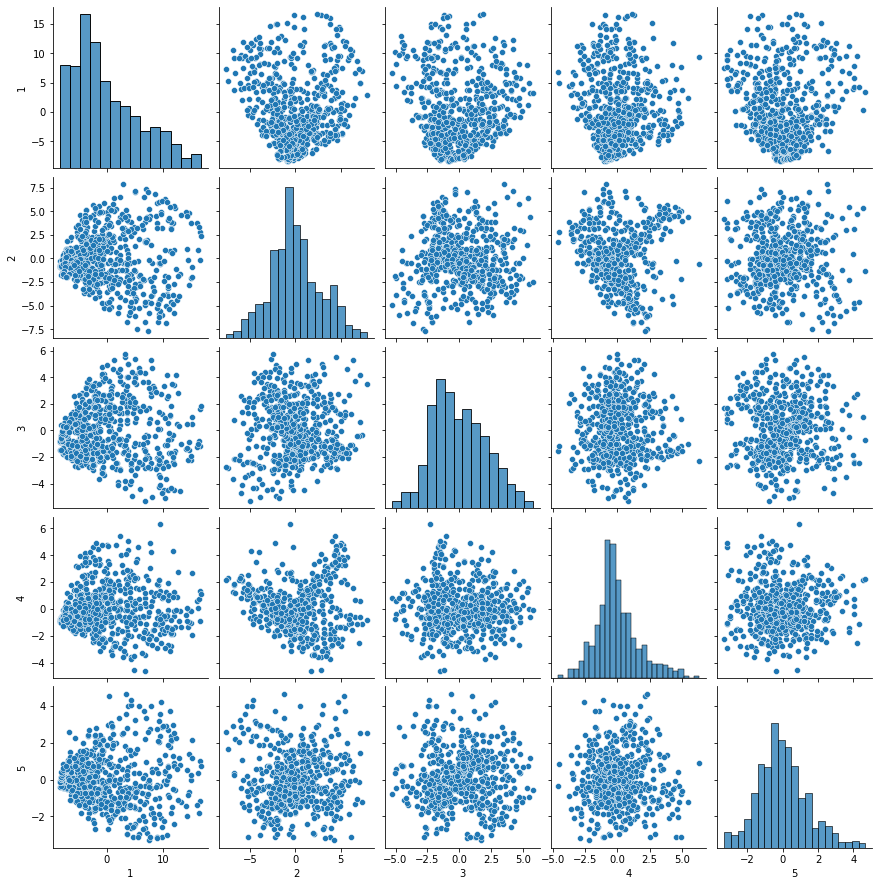
The data Consists of 61 columns and 5 rows. There are no duplicate data and missing data found in the dataset

|  |  |
| --- | --- |
| **Column** | **Dtype** |
| State Code | int64 |
| Dist.Code | int64 |
| State | object |
| Area Name | object |
| No\_HH | int64 |
| TOT\_M | int64 |
| TOT\_F | int64 |
| M\_06 | int64 |
| F\_06 | int64 |
| M\_SC | int64 |
| F\_SC | int64 |
| M\_ST | int64 |
| F\_ST | int64 |
| M\_LIT | int64 |
| F\_LIT | int64 |
| M\_ILL | int64 |
| F\_ILL | int64 |
| TOT\_WORK\_M | int64 |
| TOT\_WORK\_F | int64 |
| MAINWORK\_M | int64 |
| MAINWORK\_F | int64 |
| MAIN\_CL\_M | int64 |
| MAIN\_CL\_F | int64 |
| MAIN\_AL\_M | int64 |
| MAIN\_AL\_F | int64 |
| MAIN\_HH\_M | int64 |
| MAIN\_HH\_F | int64 |
| MAIN\_OT\_M | int64 |
| MAIN\_OT\_F | int64 |
| MARGWORK\_M | int64 |
| MARGWORK\_F | int64 |
| MARG\_CL\_M | int64 |
| MARG\_CL\_F | int64 |
| MARG\_AL\_M | int64 |
| MARG\_AL\_F | int64 |
| MARG\_HH\_M | int64 |
| MARG\_HH\_F | int64 |
| MARG\_OT\_M | int64 |
| MARG\_OT\_F | int64 |
| MARGWORK\_3\_6\_M | int64 |
| MARGWORK\_3\_6\_F | int64 |
| MARG\_CL\_3\_6\_M | int64 |
| MARG\_CL\_3\_6\_F | int64 |
| MARG\_AL\_3\_6\_M | int64 |
| MARG\_AL\_3\_6\_F | int64 |
| MARG\_HH\_3\_6\_M | int64 |
| MARG\_HH\_3\_6\_F | int64 |
| MARG\_OT\_3\_6\_M | int64 |
| MARG\_OT\_3\_6\_F | int64 |
| MARGWORK\_0\_3\_M | int64 |
| MARGWORK\_0\_3\_F | int64 |
| MARG\_CL\_0\_3\_M | int64 |
| MARG\_CL\_0\_3\_F | int64 |
| MARG\_AL\_0\_3\_M | int64 |
| MARG\_AL\_0\_3\_F | int64 |
| MARG\_HH\_0\_3\_M | int64 |
| MARG\_HH\_0\_3\_F | int64 |
| MARG\_OT\_0\_3\_M | int64 |
| MARG\_OT\_0\_3\_F | int64 |
| NON\_WORK\_M | int64 |
| NON\_WORK\_F | int64 |

The state and Area name are object data type. In PCA the Object type cannot be taken into the account.

## Q2. 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? (Example Questions). Pick 5 variables out of the given 24 variables below for EDA: No\_HH, TOT\_M, TOT\_F, M\_06, F\_06, M\_SC, F\_SC, M\_ST, F\_ST, M\_LIT, F\_LIT, M\_ILL, F\_ILL, TOT\_WORK\_M, TOT\_WORK\_F, MAINWORK\_M, MAINWORK\_F, MAIN\_CL\_M, MAIN\_CL\_F, MAIN\_AL\_M, MAIN\_AL\_F, MAIN\_HH\_M, MAIN\_HH\_F, MAIN\_OT\_M, MAIN\_OT\_F

## General pairplot for the whole dataset



## Q3. We choose not to treat outliers for this case. Do you think that treating outliers for this case is necessary?

A picture containing calendar

Description automatically generated

Box plot for all the data fields

As we can see the above plot have has some outliers present in it. But These outliers can be considered as they are not going affect the data. No need to treat outlier in this case.

The General scenario for the outliers

Correlation is affected averages

Averages are affected by the outliers. Standard scaler and min max scaler are affected by the outliers. So during the scaling the outliers can get affected.

## Q4. Scale the Data using z-score method. Does scaling have any impact on outliers? Compare boxplots before and after scaling and comment.

Yes we have scaled the data using Z score method.

A picture containing calendar

Description automatically generated

*Box plot before scaling*

A picture containing text, electronics, keyboard

Description automatically generated

*Box plot after scaling*

There is not much difference between before and after scaling box plot.

## Q5. Perform all the required steps for PCA (use sklearn only) Create the covariance Matrix Get eigen values and eigen vector.

**We have created the covariance matrix got the eigen values and eigen vector**

array ([53.9, 67.8, 75.5, 82. , 85.8, 89.1, 91.4, 92.9, 94.1, 95.1, 95.9,

96.6, 97.2, 97.7, 98.2, 98.5, 98.7, 98.9, 99.1, 99.3, 99.4, 99.5,

99.6, 99.7, 99.8, 99.8, 99.8, 99.8, 99.8, 99.8, 99.8, 99.8, 99.8,

99.8, 99.8, 99.8, 99.8, 99.8, 99.8, 99.8, 99.8, 99.8, 99.8, 99.8,

99.8, 99.8, 99.8, 99.8, 99.8, 99.8, 99.8, 99.8, 99.8, 99.8, 99.8,

99.8, 99.8, 99.8, 99.8])

The Detailed steps ae from eigen vector and eigen value we get the variance ratio. From that variance we are getting the above matrix.

Q6.  Identify the optimum number of PCs (for this project, take at least 90% explained variance). Show Scree plot.

Chart, line chart

Description automatically generated

*Scree plot*

The optimum number of PCs are 6. Because the value is ~90% after that.

Q7. 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 PCs are given. The principal component in terms actual variable reduce the dimensionality .

Q8. Write linear equation for first PC.

The Linear equation for first PC

PC1 = a1X1+a2X2+a3X3+…. +anXn