DMG Assignment 4
Clustering

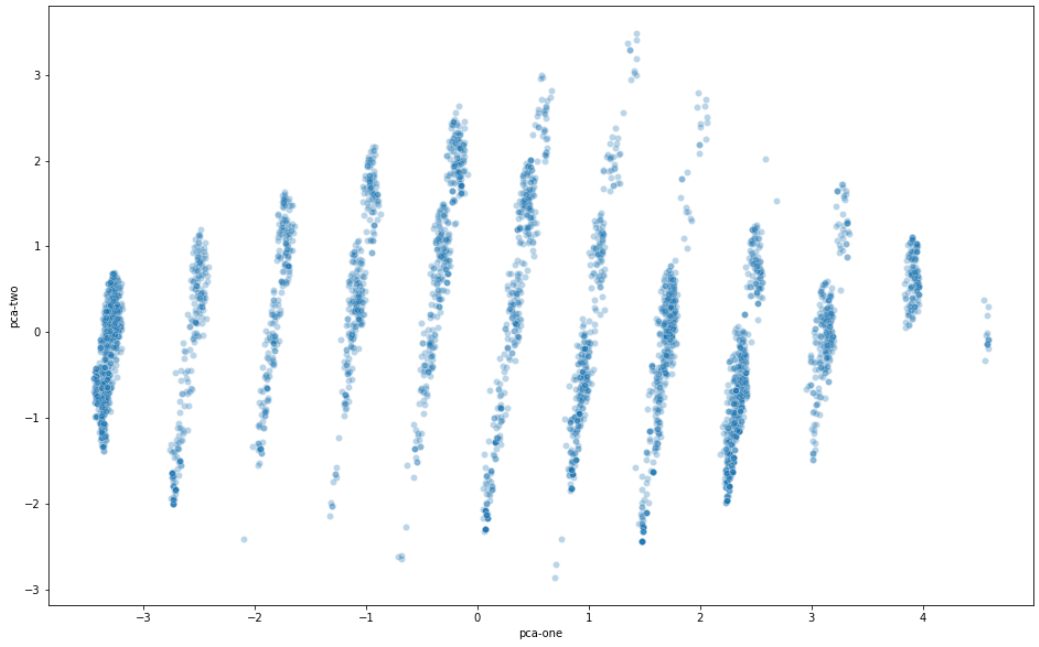
Submitted By :
Gaurav Lodhi (MT19063)
Nitindeep Singh (MT19069)

Preprocessing:- By using pandas library we read the given clustering.csv file into a data frame. We applied one-hot encoding to some ordinal attributes those are `'Elevation','Aspect','Slope','Wilderness','Soil_Type','Hillshade_9am','Hillshade_noon','Horizontal_Distance_To_Fire_Points'`

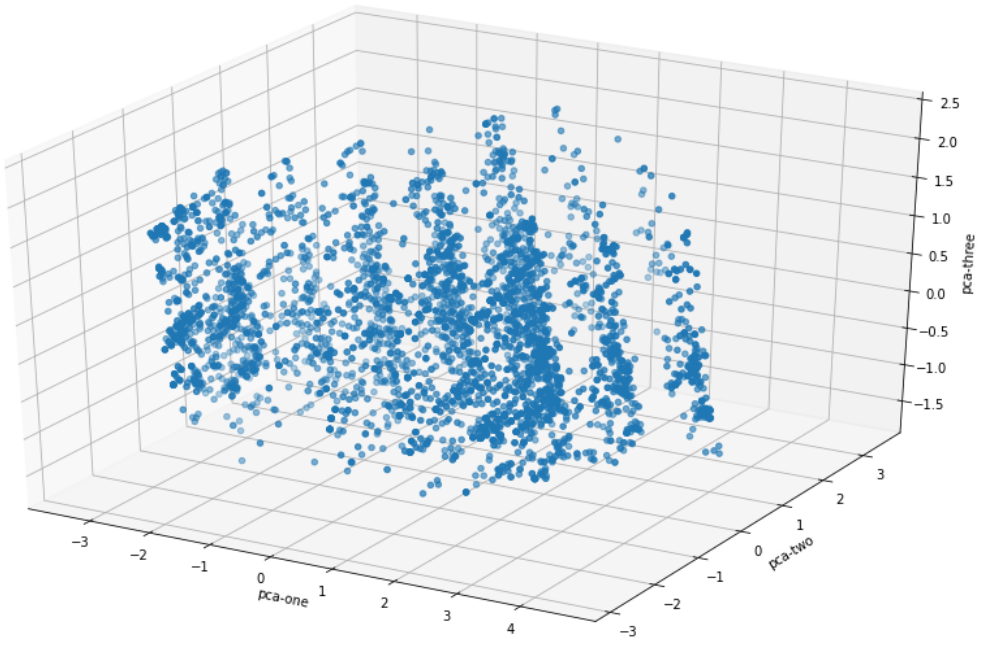
After applying the one-hot encoding I got a 4120x61 size data frame. Now I plotted all the data points using PCA for 2- Dimensional visualization and seaborn for 3-Dimensional visualization these are the following:

Skewness visualization:

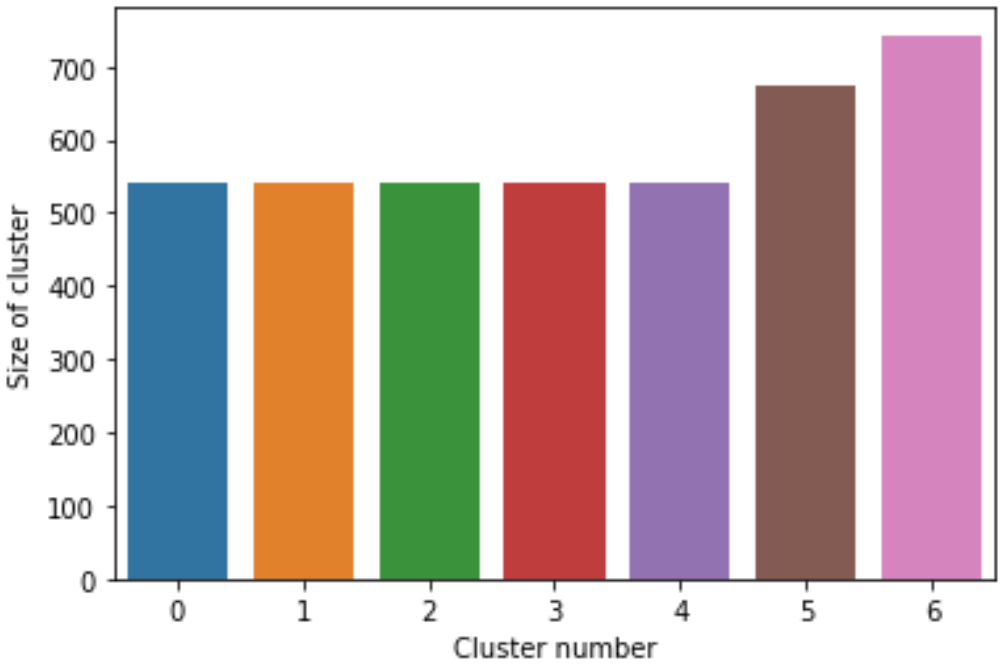
2-Dimensional visualization of the given data points:



3-Dimensional visualization of the given data points:



Then I plotted the given true cluster sizes of different clusters at n=7 (true cluster sizes):



Methodology: I used different clustering techniques to cluster the data into their clusters: Some of these are mentioned below:

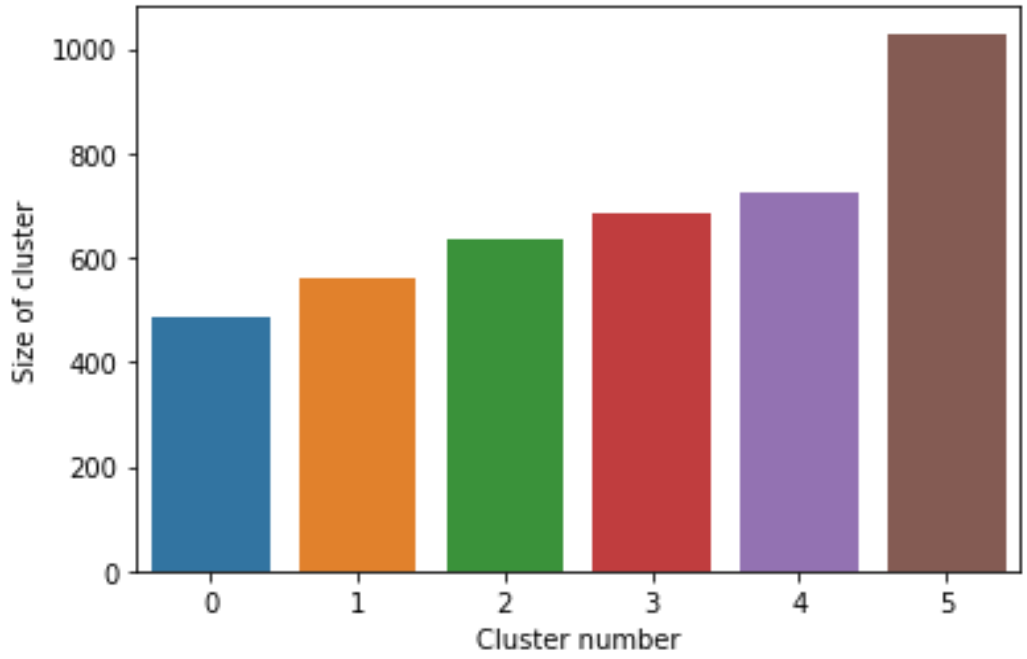
1. K-means clustering: We used sklearn library for k-means clustering implementation and we tune this clustering on the different parameters some parameter results is given below. We used this method for baseline.

   Note:- I am mentioning the bar plot for cluster sizes and and cluster visualization in 2-D and 3-D as following:

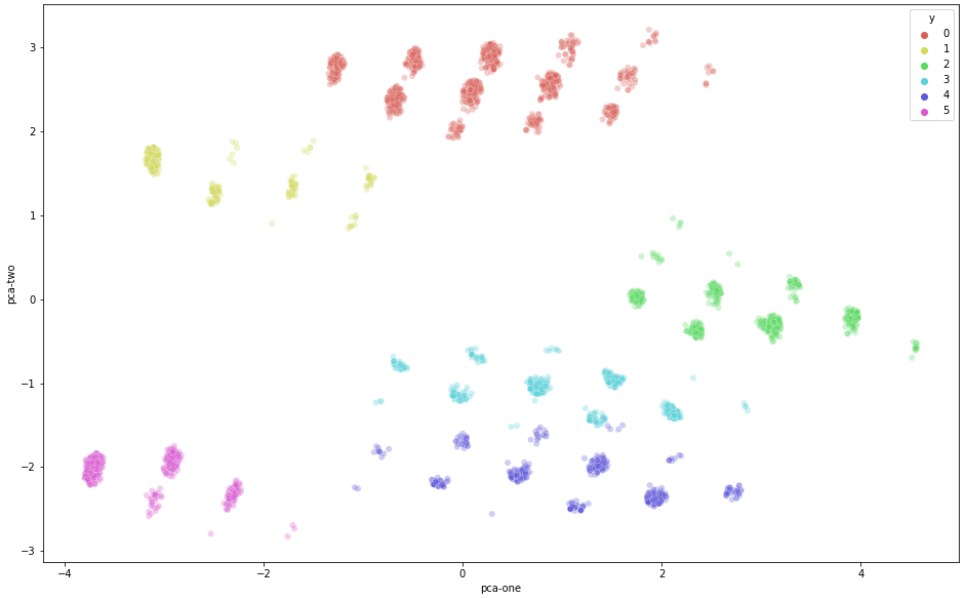
a. `n_clusters=6, random_state=0,init='random',max_iter=100`

   `Cluster sizes: [684, 1029, 637, 559, 723, 488]`

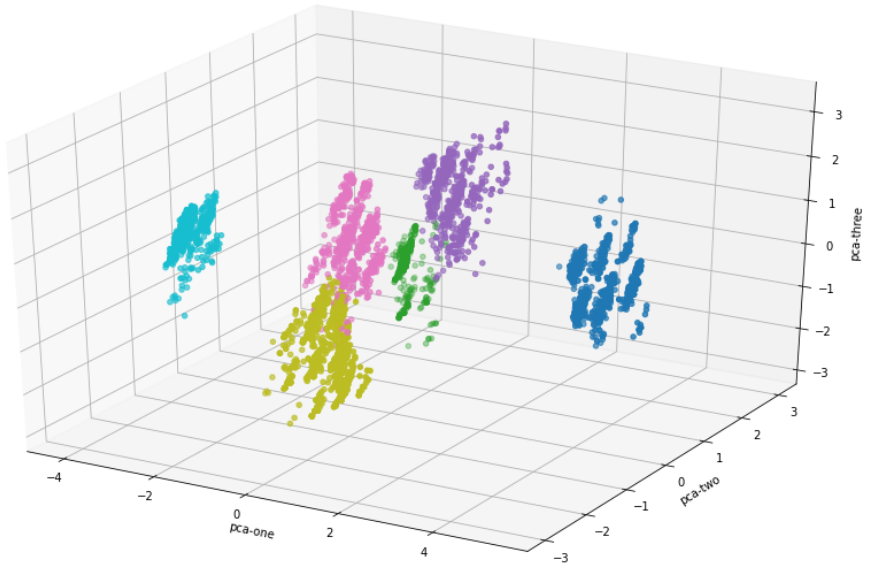
Bar graph for all the predicted cluster sizes for comparison with the true clusters:



2-Dimensional visualization of the clusters:



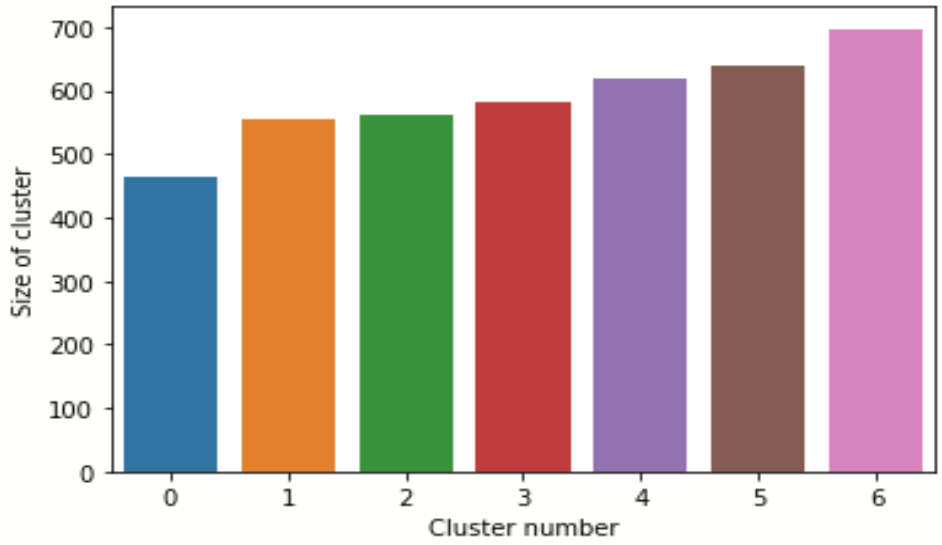
3-Dimensional visualization of the clusters:



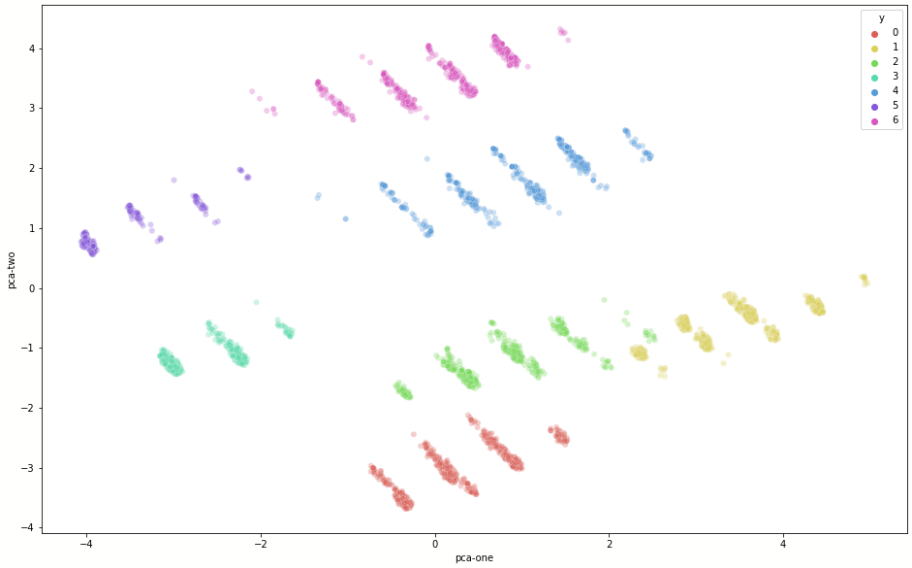
b. `n_clusters=7, random_state=0,init='random',max_iter=100`

`Cluster sizes: [640, 583, 617, 556, 563, 697, 464]`

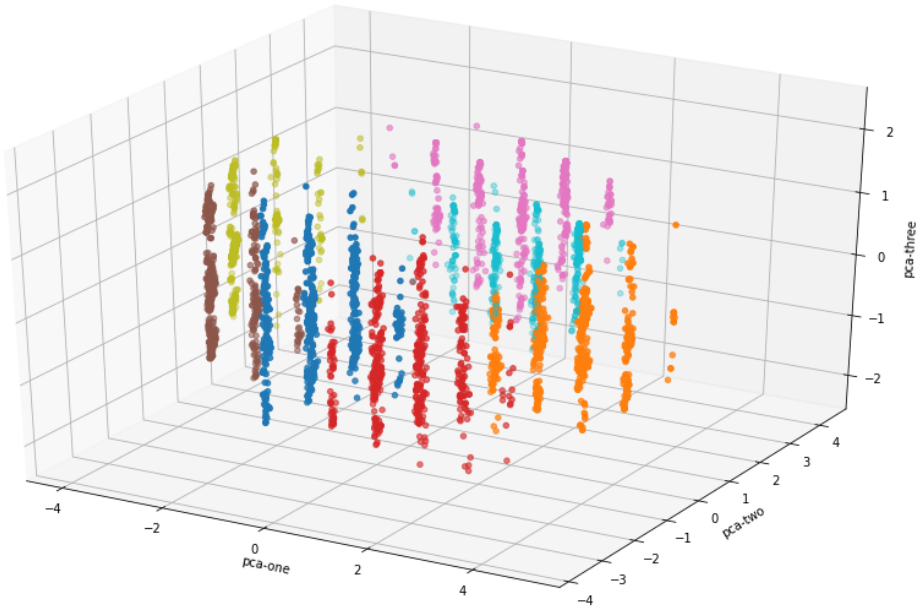
Bar graph for all the predicted cluster sizes for comparison :



2-Dimensional visualization of the clusters:



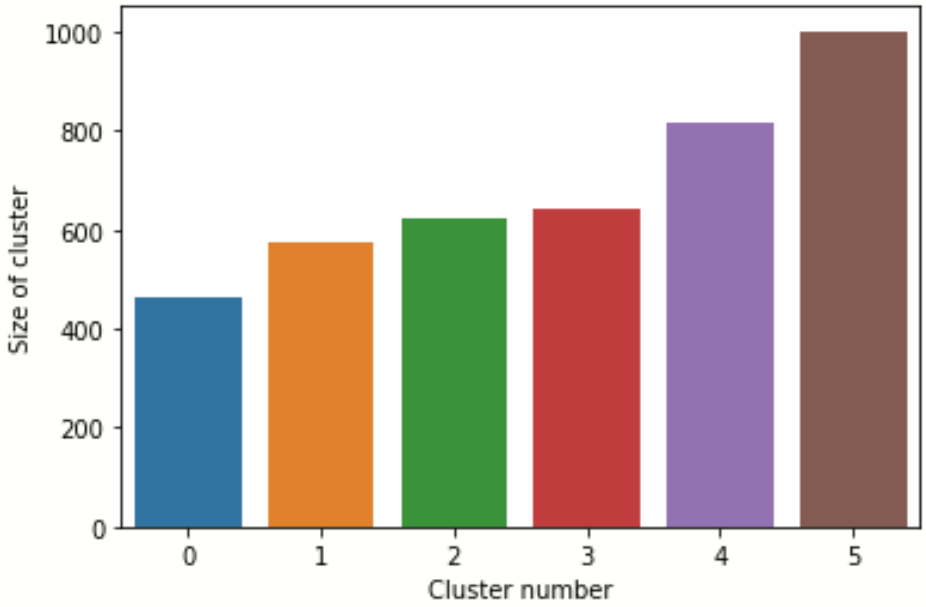
3-Dimensional visualization of the clusters:



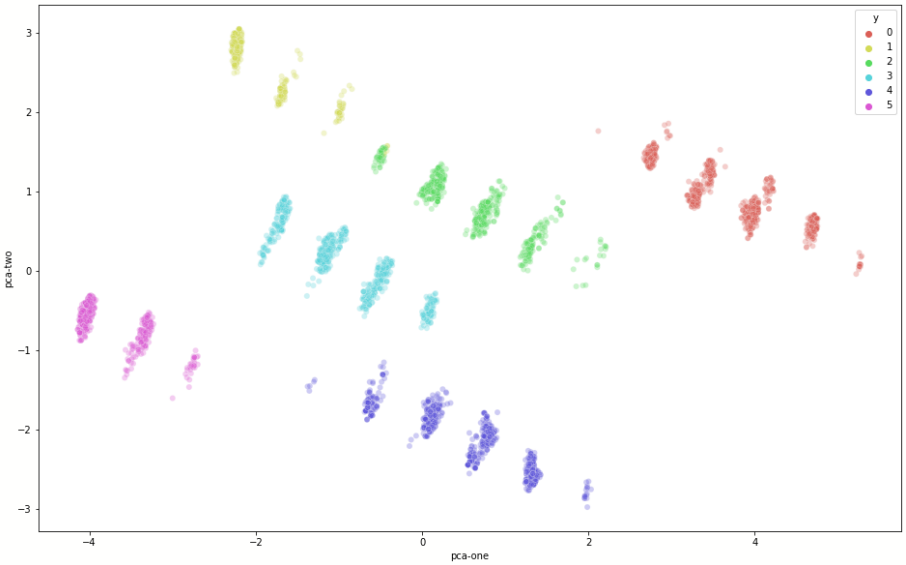
2. K-Means++ clustering algorithm: We used sklearn library for k-means++ clustering implementation and we tune this clustering on the different parameters some parameter results are given below. We used this method for making a more complex prediction with respect to the K-means clustering.

a. `n_clusters=6, random_state=0,init='k-means++',max_iter=100:-`
`Cluster sizes: [641, 621, 1002, 573, 819, 464]`

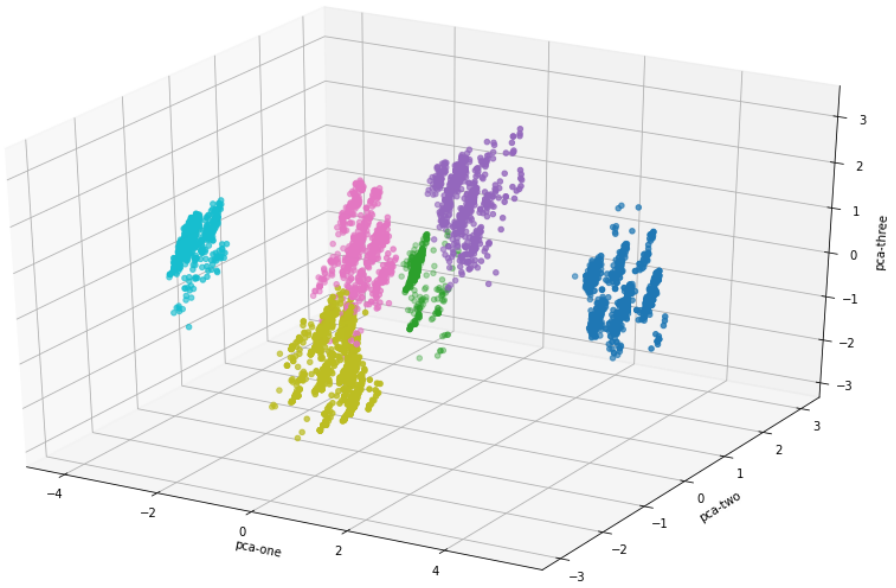
Bar graph for all the predicted cluster sizes for comparison with true clusters:



2-Dimensional visualization of the clusters:



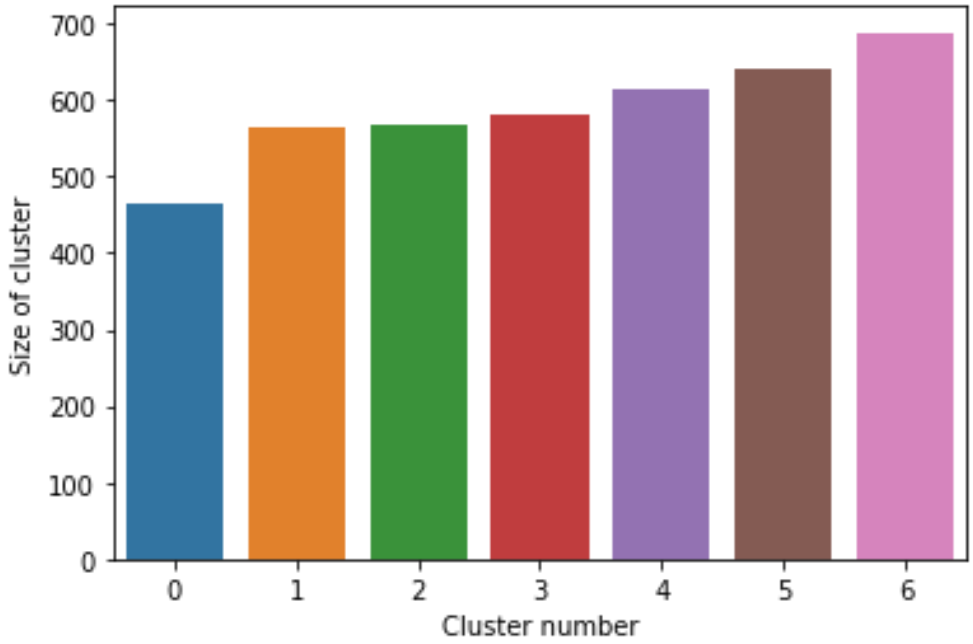
3-Dimensional visualization of the clusters:



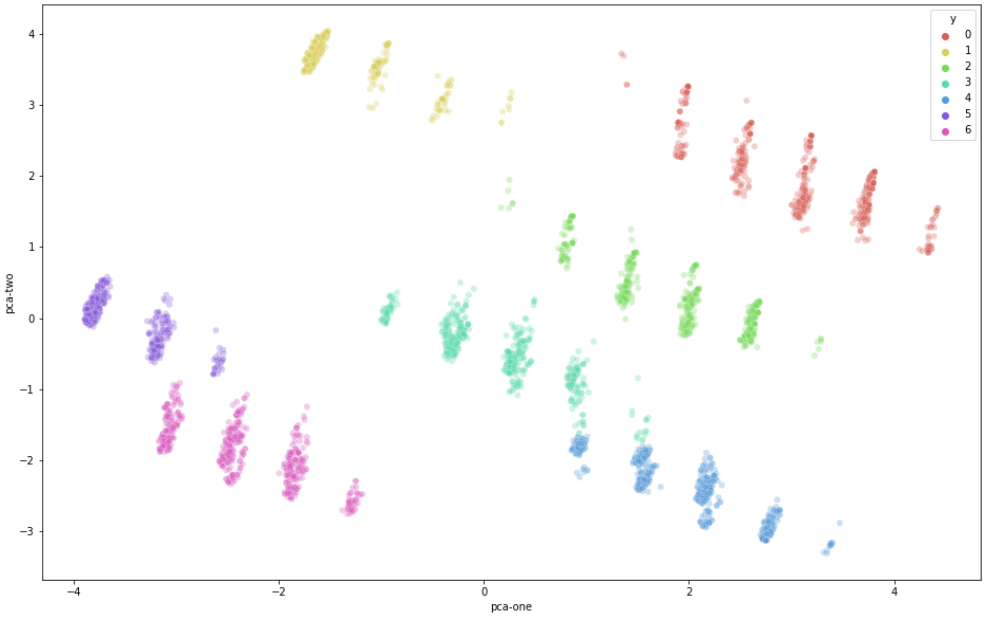
b. `n_clusters=7, random_state=0,init='k-means++',max_iter=100:-`

`Cluster sizes: [640, 582, 615, 566, 565, 688, 464]`

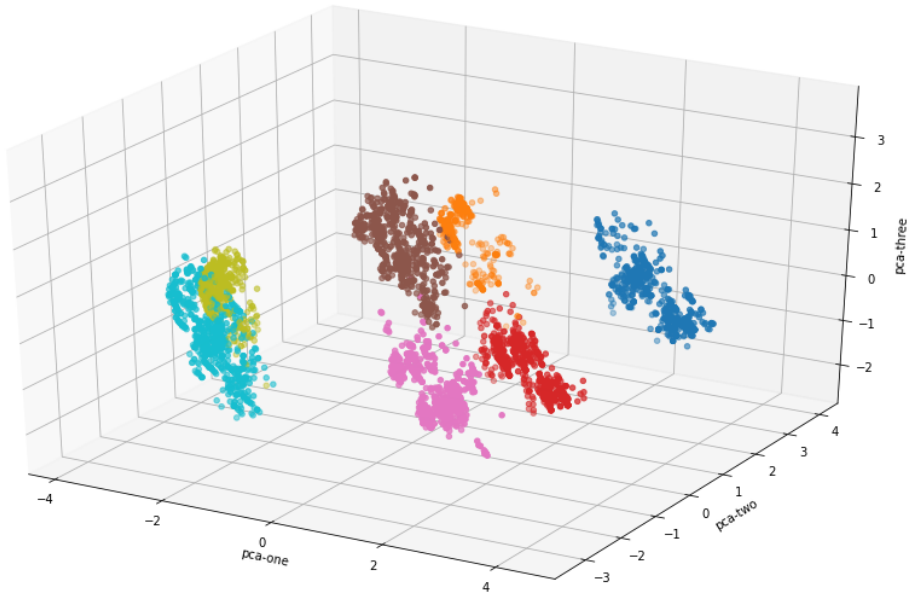
Bar graph for all the predicted cluster sizes For Comparison:



2-Dimensional visualization of the clusters:



3-Dimensional visualization of the clusters:

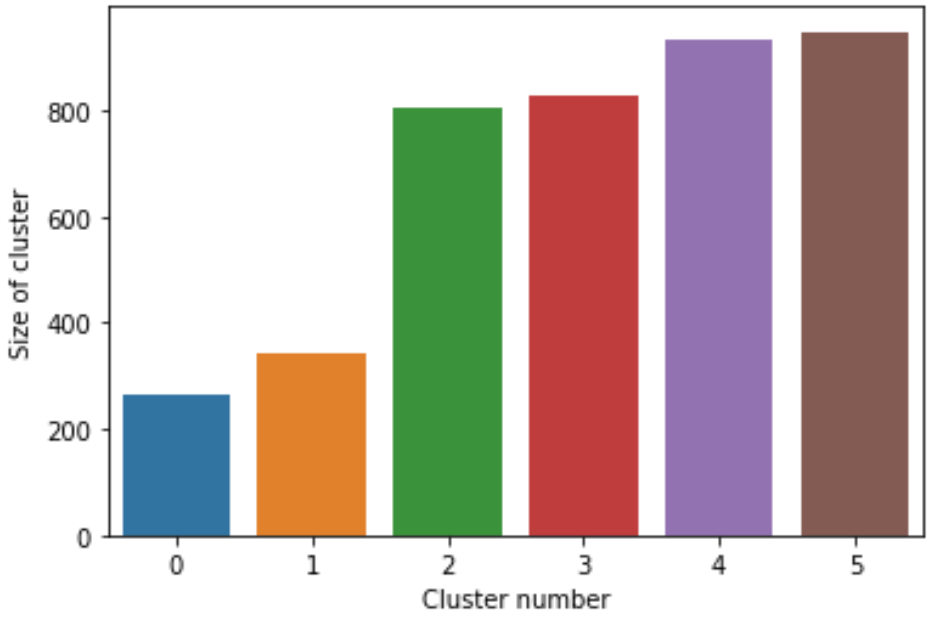


3. Birch Clustering: I used sklearn library for Birch clustering implementation for finding more optimal clustering and we tune this clustering on the different parameters some parameter results are given below.

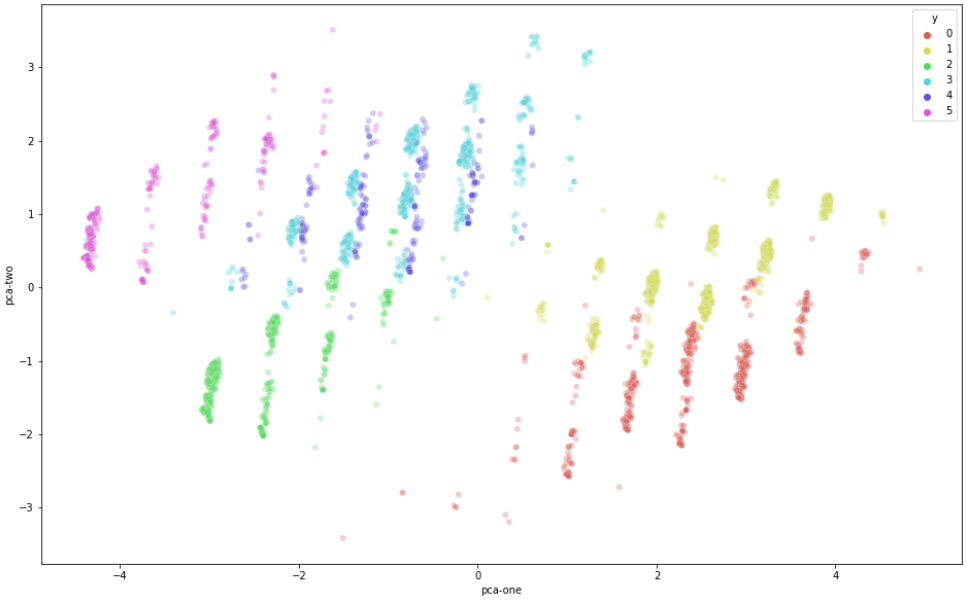
1. `n_clusters=6:`

   `Clusters size:[933, 341, 948, 829, 806, 263]`

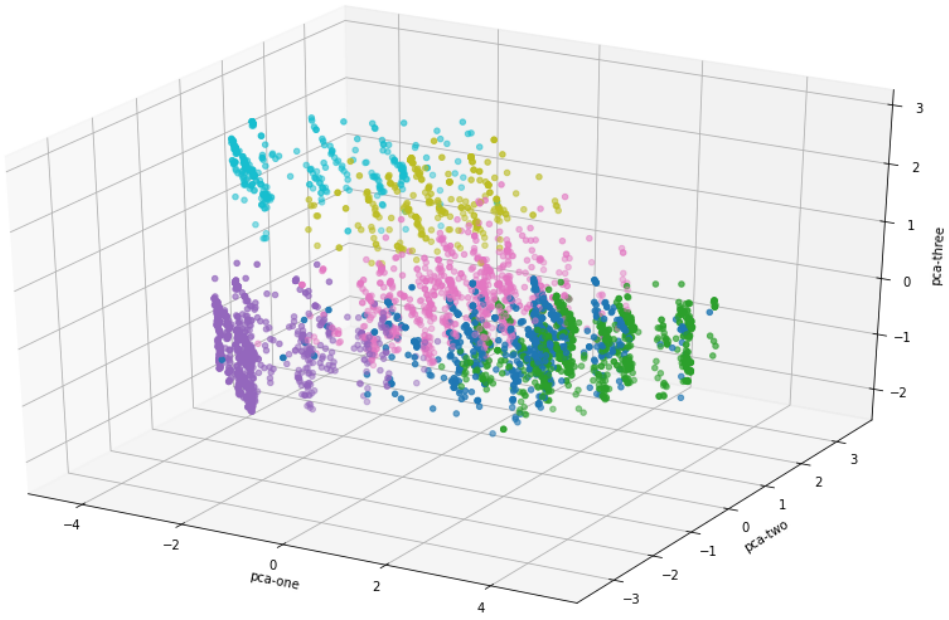
Bar graph for all the predicted cluster sizes for comparison with true labels:



2-Dimensional visualization of the clusters:



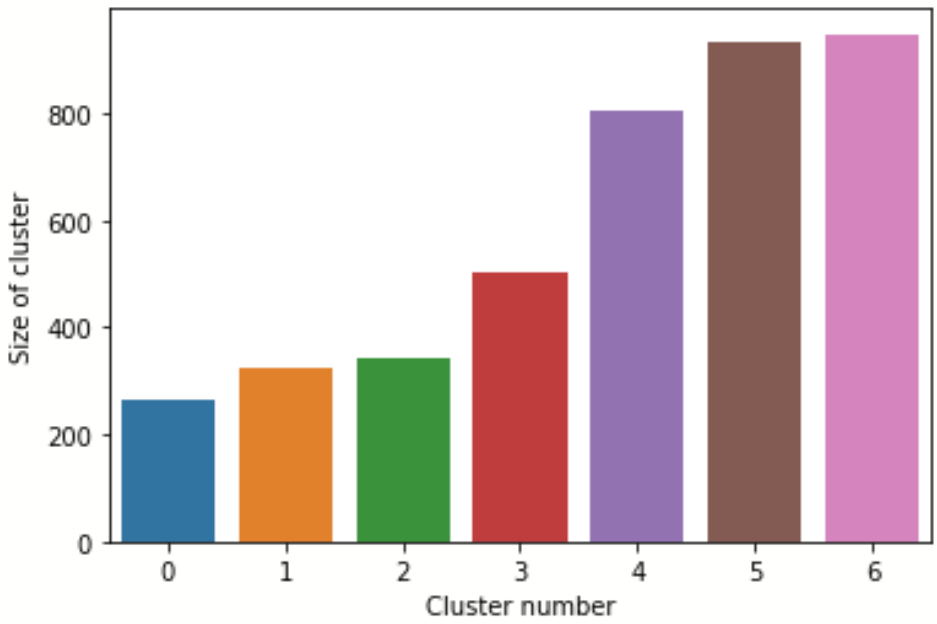
3-Dimensional visualization of the clusters:



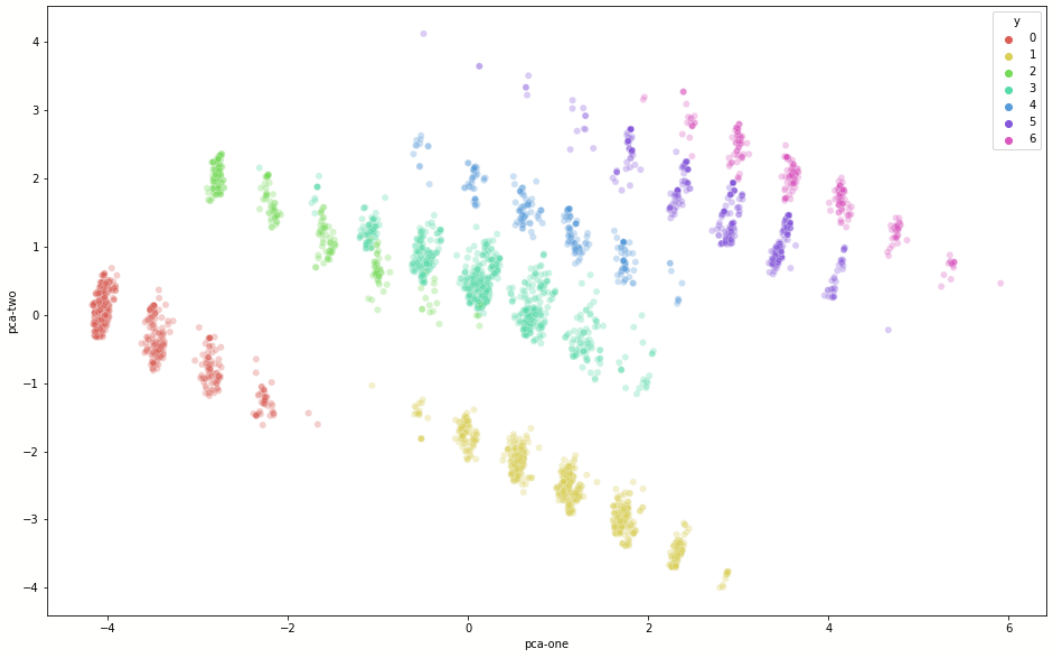
2. n_clusters=7:

Clusters Size: [933, 341, 948, 326, 806, 263, 503]

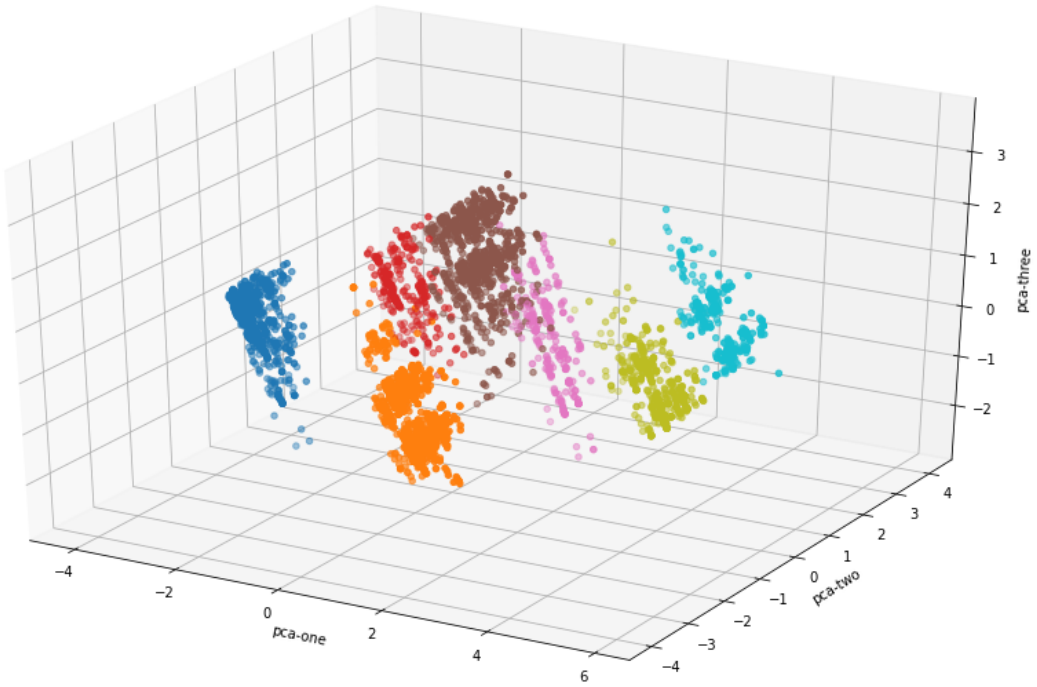
Bar graph for all the predicted cluster sizes for comparison with true labels:



2-Dimensional visualization of the clusters:



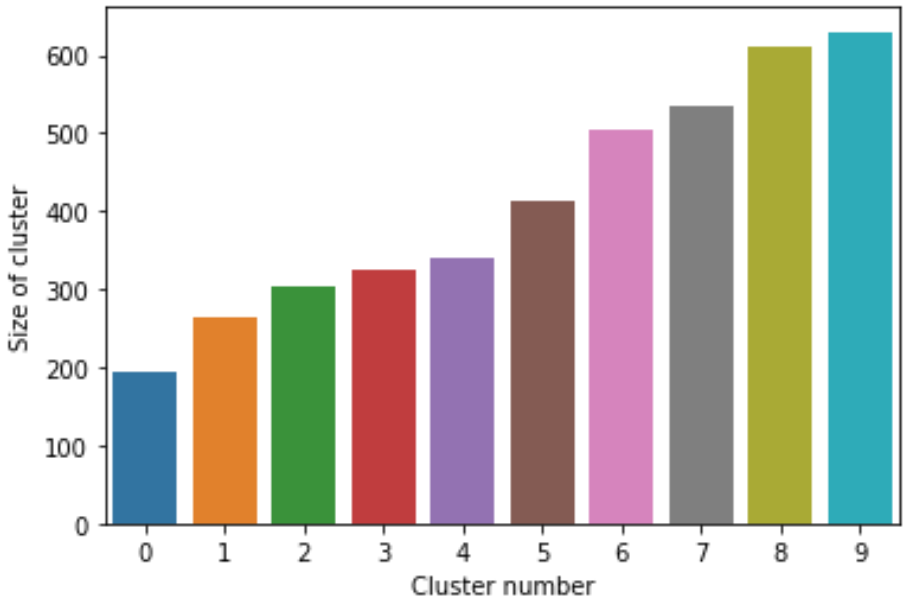
3-Dimensional visualization of the clusters:



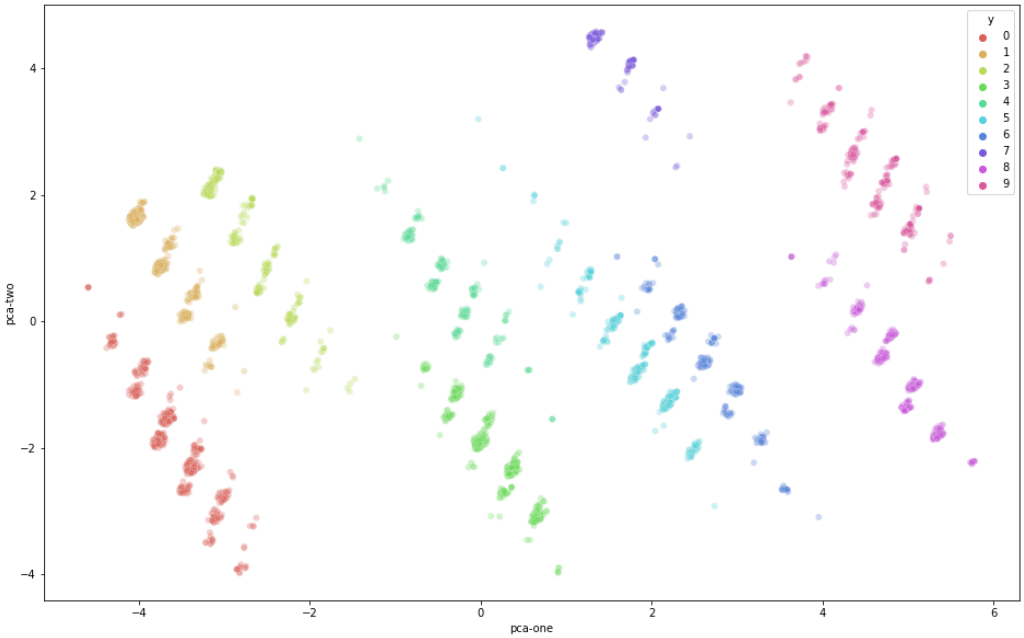
3. n_clusters=10

Clusters size:[630, 341, 535, 326, 194, 612, 263, 303, 503, 413]

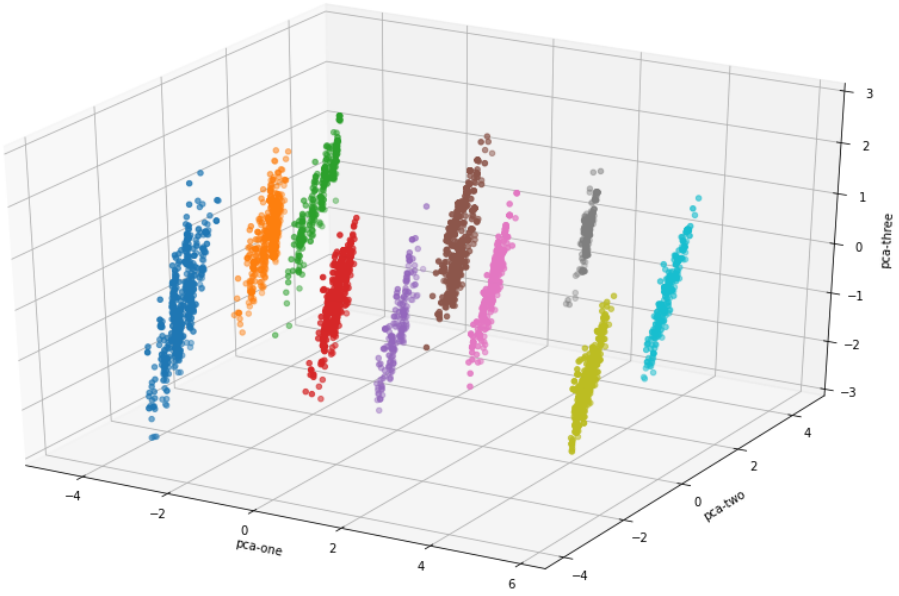
Bar graph for all the predicted cluster sizes for comparison with the true cluster labels:



2-Dimensional visualization of the clusters:



3-Dimensional visualization of the clusters:



Result: I am taking most accurate model to make result.csv file as K-mean++ with following parameters:

```
n_clusters=7, random_state=0,init='k-means++',max_iter=100
Cluster sizes: [640, 582, 615, 566, 565, 688, 464]
```

Learnings:

1. It made us familiar with many clustering algorithms.
2. We learned what are the different parameters and how to use them.
3. How to visualize clustering data.
4. How to visualize high dimensional data.
5. How to process ordinal attributes in the dataset.

Contribution:

Gaurav Lodhi: Preprocessing, K-means++ clustering, visualization

Nitindeep Singh: K-means clustering, Birch clustering

Note:

1. We printed all the centroids in the program itself.
2. We mentioned the plot of true clusters in the beginning and we plotted the bar graph for all the clusters to compare the clusters. So please consider all the different bar graphs for comparing purposes.