1.Data exploration and analysis

- Used both the dataset Life Expectancy-income and Literacy rate by gender.
- First dataset consists of 3attributes Districts, Life expectancy in years and per capita income in dollars.
- The 2nd dataset consists of 5attributes District, total literacy rate, female literacy rate , male literacy rate and years.

Data preparation

- df_hdi.isnull().values.any()
  False

- Use strip() function to remove extra space in District.
- Since the data type of per capita income attribute is object type(mixture of number and character) so i converted to float by using astype()function so that it will be easy for the coming process.
- Although there are all districts present in each dataset but there was variation in spelling so at the time of merging I replaced the district name in one of the dataset.

- Highest per capita income
  Manang USD 3166
- Lowest per capita income
  Bajhang USD 487
- Highest Literacy Rate
  Kathmandu 86.3
- Lowest Literacy Rate
  Rautahat 41.7

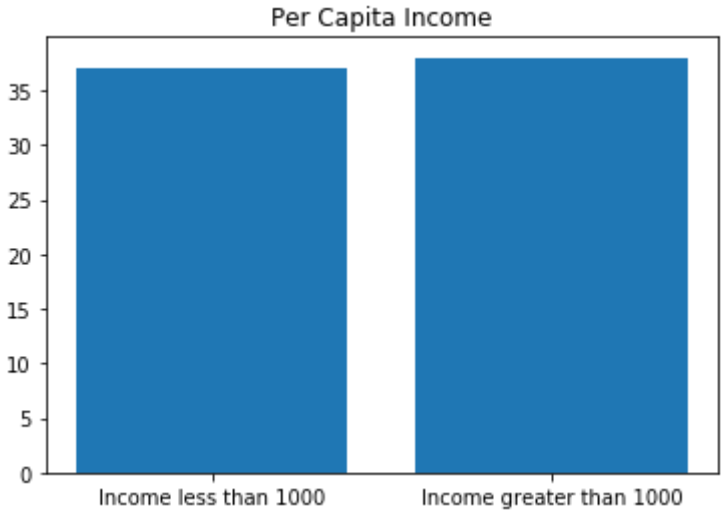


Fig: Here, the number of districts whose income less than 1000 is 37 and income greater than 1000 is 38.

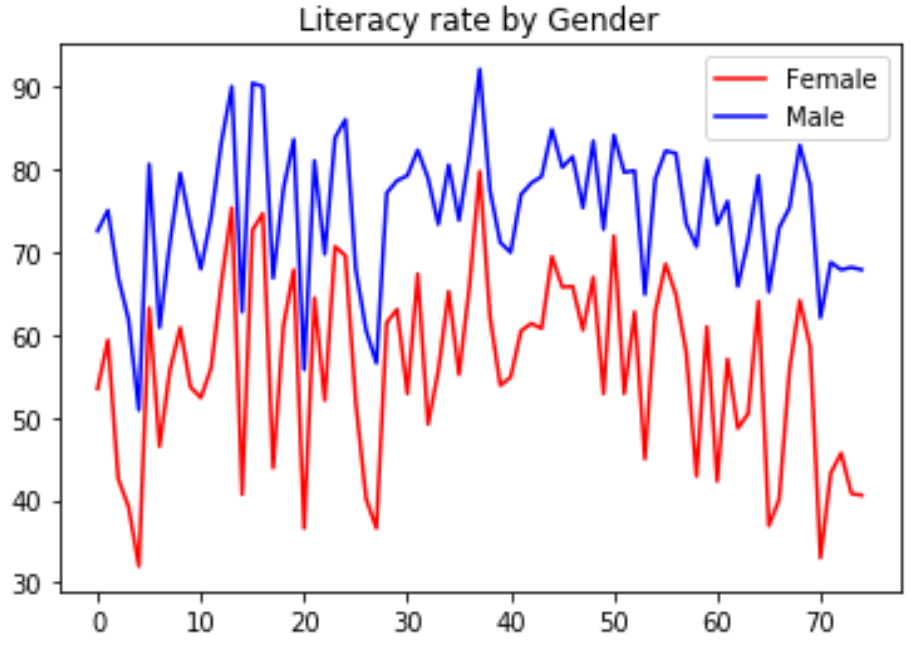


Fig2: Male literacy rate in each district is greater than Female.

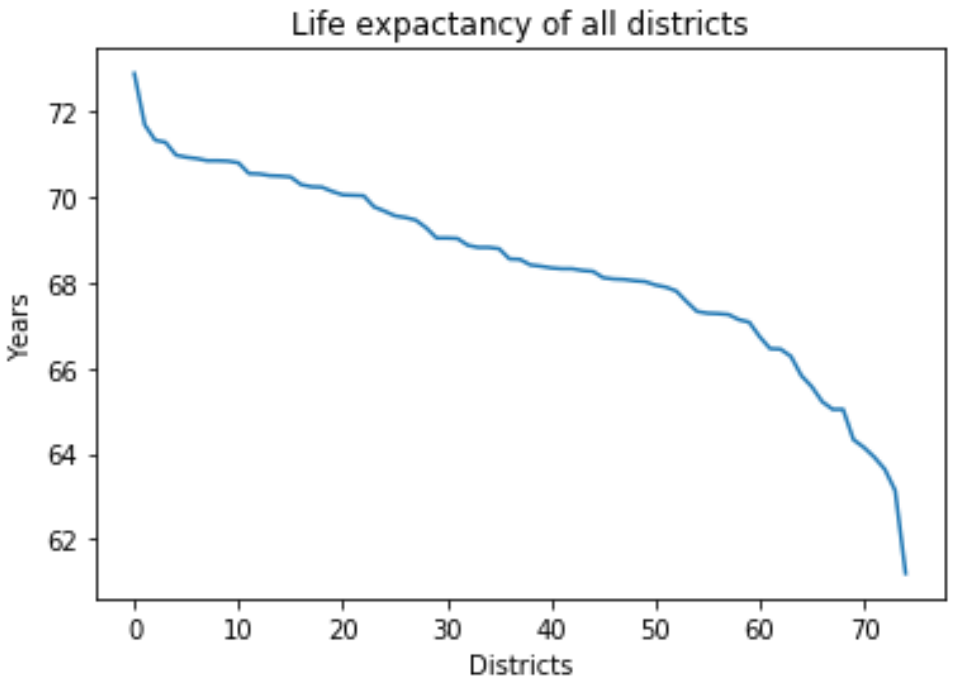


Fig3: life expectancy of all the District ranges from 61.20 to 72.90

Highest life expectancy : Ramechhap 72.90
Lowest life expectancy: Dolpa 61.20

Correlation Table

	Life expectancy	Per Capita	Total	Female	Male
Life expectancy	1.000	0.071	0.135	0.210	0.047
Per Capita	0.071	1.000	0.506	0.499	0.420
Total	0.135	0.506	1.000	0.987	0.971
Female	0.210	0.499	0.987	1.000	0.924
Male	0.047	0.420	0.971	0.924	1.000

2. Feature Selection and Preprocessing

- Total features used for clustering is 3 which are Life expectancy, per capita income and total literacy rate in each District.
- Years attribute has been removed as all the values are the same and didn't provide any information.
- As we are clustering on the basis of Human Development Index so i have used Total literacy rates instead of using male and female individually.
- **Normalization** rescales the values into a range of [0,1]. This might be useful in some cases where all parameters need to have the same positive scale. However, the outliers from the data set are lost. **Standardization** rescales data to have a mean (μ) of 0 and standard deviation (σ) of 1 (unit variance).
- Standard Scalar has been used for standardization of data.

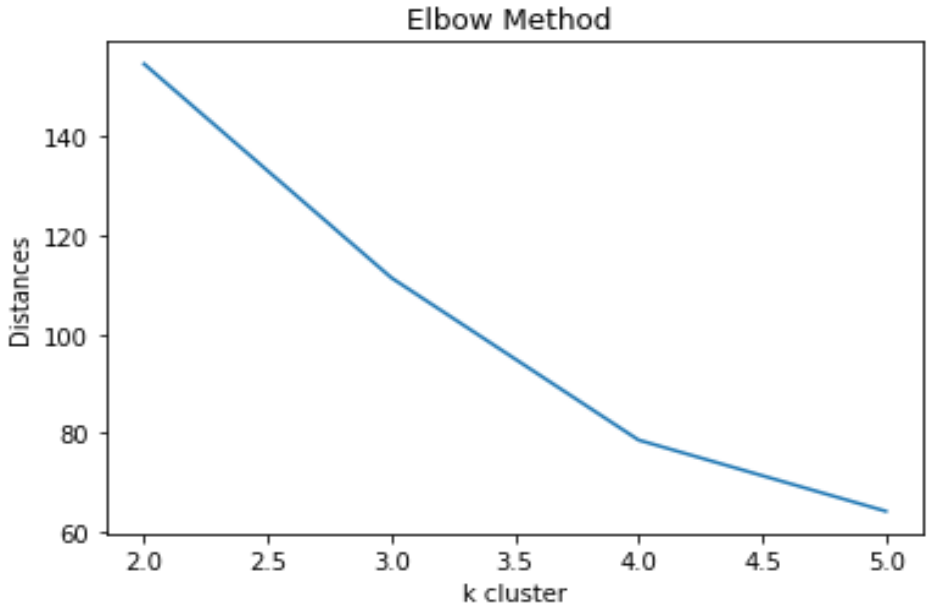
  z = (x - u) / s, where u is the mean of given datas, s is the standard deviation.

3. KMeans Algorithm for clustering

- **K-means clustering** is one of the simplest and popular unsupervised machine learning **algorithms.**
- The k-means algorithm is an algorithm to cluster n objects based on attributes into k partitions, where k < n.
- Choosing the k value with the help of both elbow-method and silhouette score
- Metric used 'Euclidean'

Elbow Method

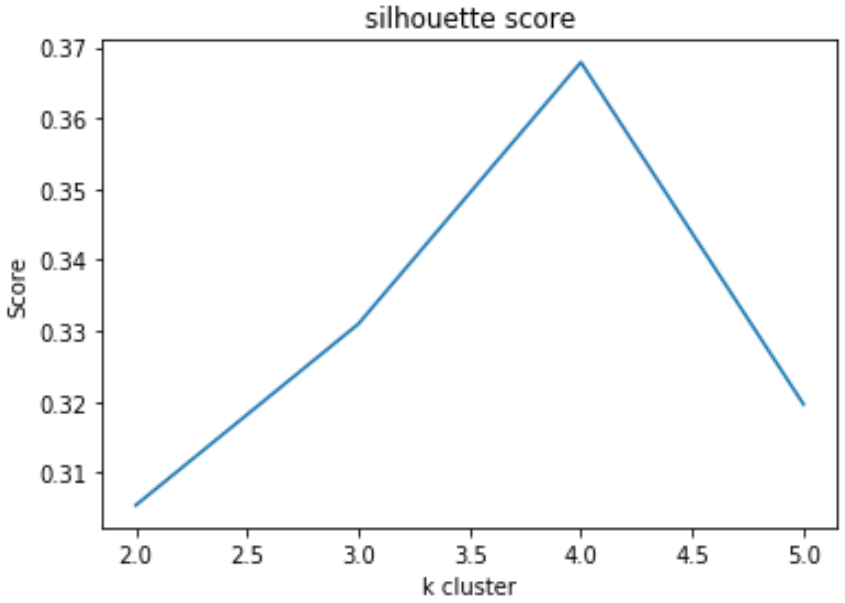
It calculates the distances of all the points from center within the cluster. (**Intradistance**)



Silhouette score

The silhouette value measures how similar a point is to its own cluster (cohesion) compared to other clusters (separation). The range of the Silhouette value is between +1 and -1. A high value is desirable and indicates that the point is placed in the correct cluster.

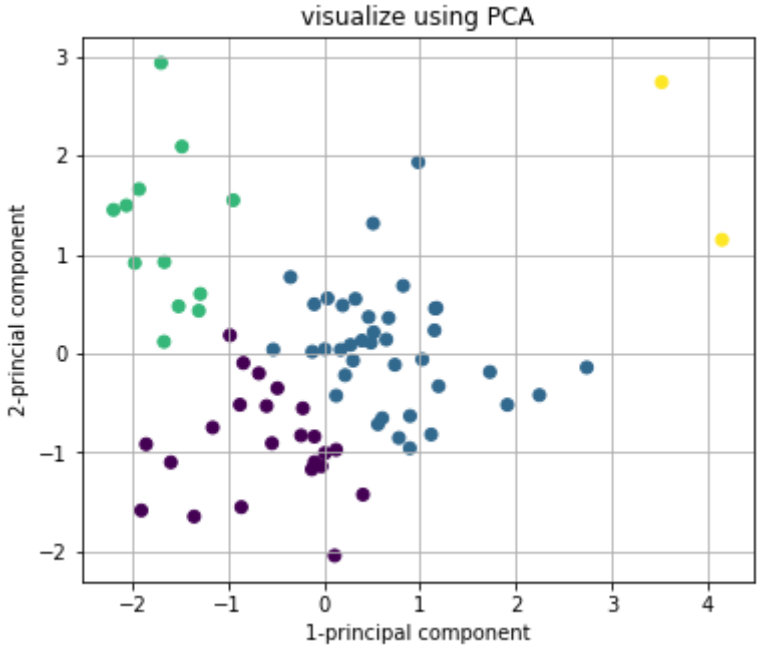
(b - a) / max(a, b), a is the mean intra cluster distance and b is the mean nearest cluster distance.



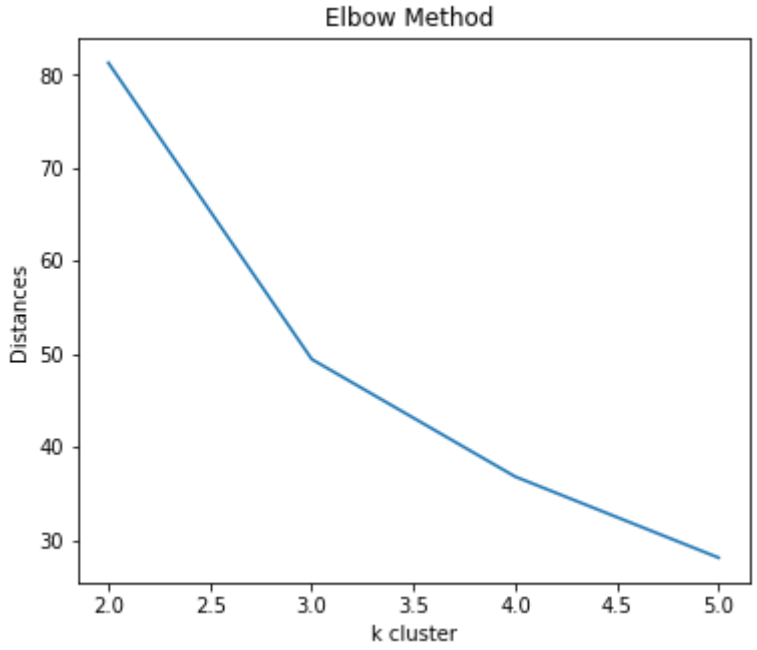
By observing both the methods we can choose k = 4

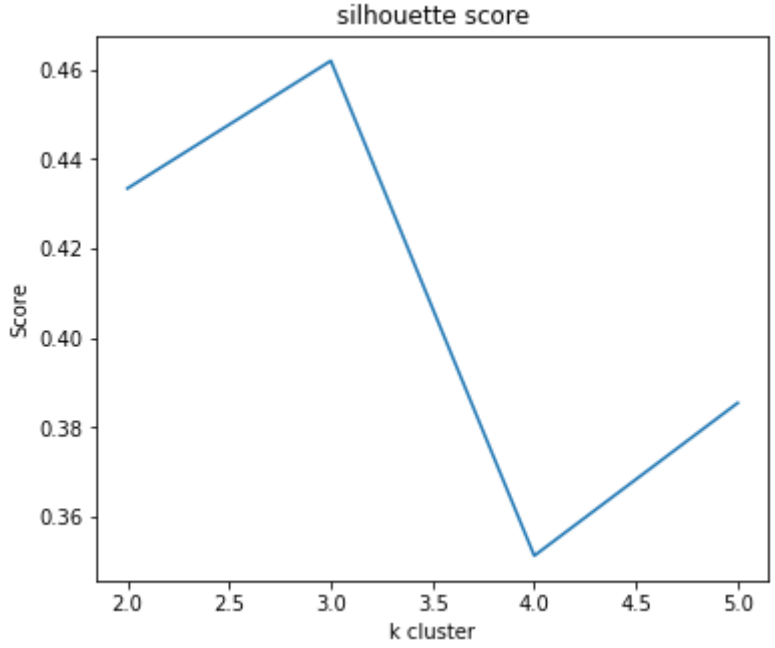
PCA to visualize the cluster in 2-D

Principal component analysis (PCA) is a technique for reducing the dimensionality of the datasets, preserving much of the information.

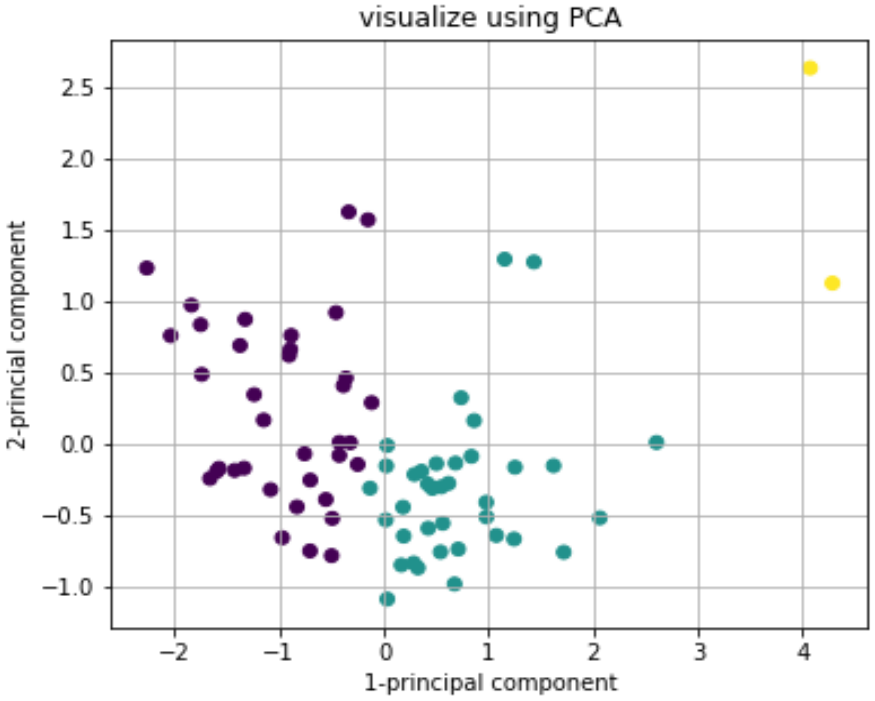


->When removing the life expectancy dimension





By choosing value k=3



DBSCAN

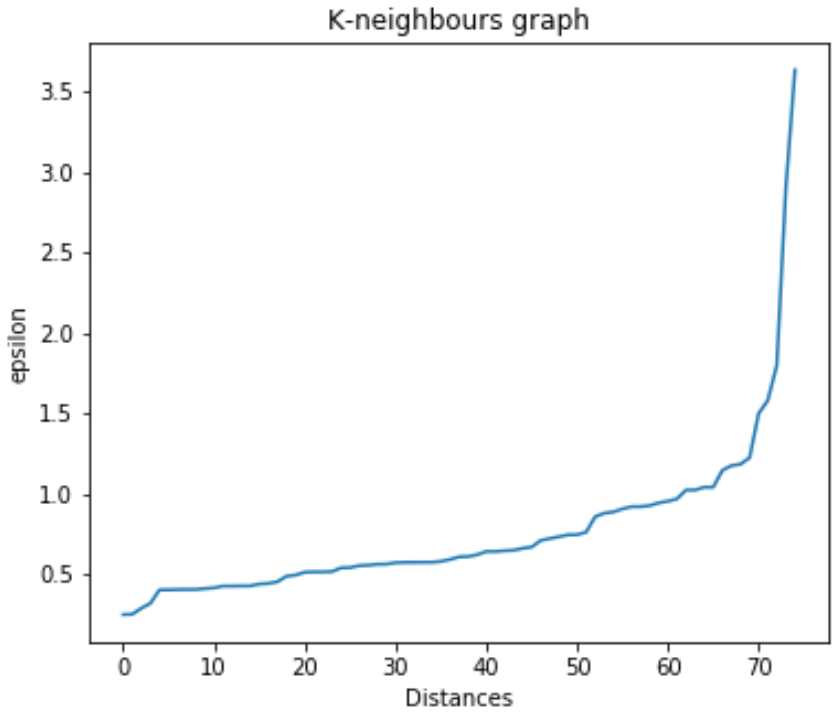
Discovers clusters of arbitrary shape in spatial databases with noise.

DBSCAN algorithm requires two parameters

a) **Eps**
   It defines the neighborhood around a data point

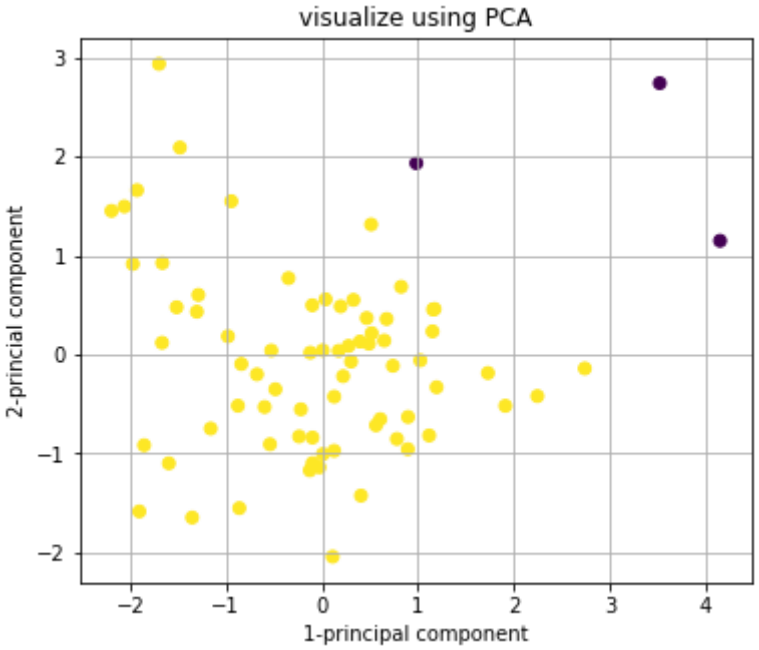
b) **MinPts**: Minimum number of neighbors (data points) within eps radius

k-distance graph to find epsilon values.
N-neighbours distance upto: 4
Distance of 3rd neighbour

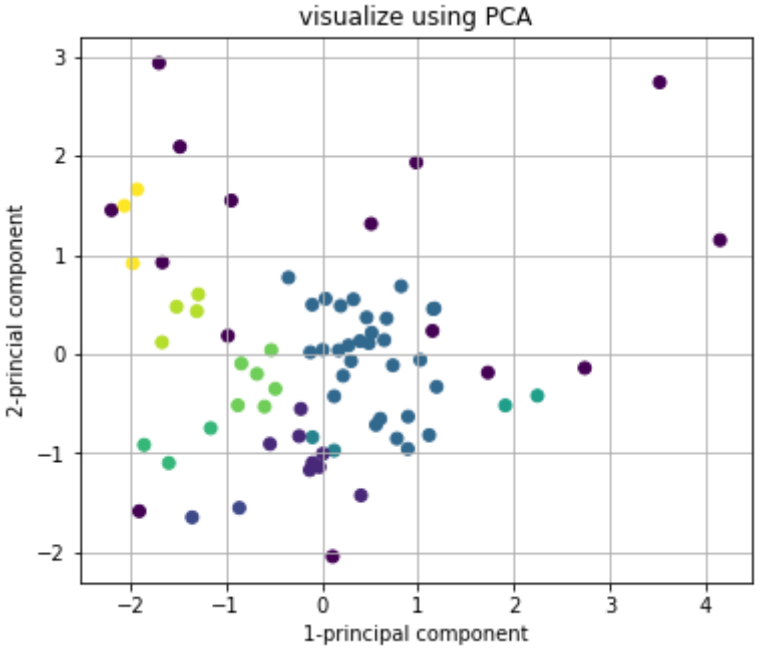


From the above figure we cannot do clear observation although it seems to be elbow at esp 1.2 but the clustering was bad.

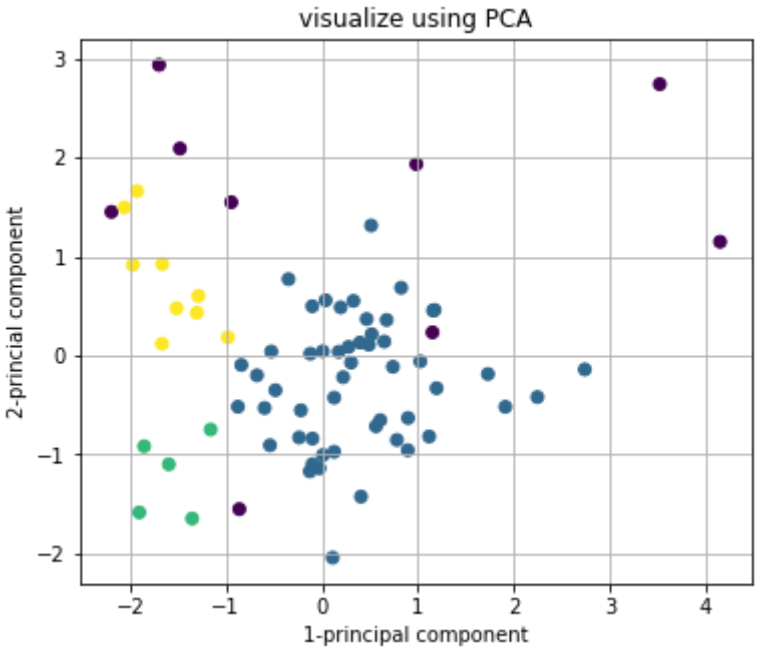
❖ When **eps** = 1.2 and **min_samples** = 4



When **eps** = 0.6 and **min_samples** = 2



When **eps** = 0.8 and **min_samples** = 4



Although when eps = 0.8 and min_samples =4, outliers was less but the cluster was not so good and it's hard to choose the value of eps so i choose k means algorithm for clustering the 75District of Nepal by HDI.