```
THE SPARKS FOUNDATION
         DATA SCIENCE AND ANALYTICS TASK
         TASK - PREDICTION USING UNSUPERVISED ML¶
         Name- VEDANT GAIKWAD
         importing all required libraries
In [15]: # Importing the libraries
         import numpy as np
         import matplotlib.pyplot as plt
         import pandas as pd
         from sklearn import datasets
In [17]: iris = datasets.load_iris()
         iris_df = pd.DataFrame(iris.data, columns = iris.feature_names)
In [19]: iris_df
Out[19]:
              sepal length (cm) sepal width (cm) petal length (cm) petal width (cm)
           0
                        5.1
                                    3.5
                                                  1.4
                                                              0.2
           1
                        4.9
                                    3.0
                                                  1.4
                                                              0.2
                                                  1.3
                        4.7
                                    3.2
                                                              0.2
           3
                        4.6
                                    3.1
                                                  1.5
                                                              0.2
                                                  1.4
                        5.0
                                    3.6
                                                              0.2
                                                  ...
                        ...
                                     ...
                                                               ...
          145
                        6.7
                                    3.0
                                                  5.2
                                                              2.3
          146
                        6.3
                                    2.5
                                                  5.0
                                                              1.9
          147
                        6.5
                                    3.0
                                                  5.2
                                                              2.0
          148
                                    3.4
                                                  5.4
                                                              2.3
                        6.2
                                    3.0
                                                  5.1
          149
                        5.9
                                                              1.8
         150 rows × 4 columns
In [20]: #check the shape of dataset
         iris_df.shape
Out[20]: (150, 4)
In [21]: #Reading first 10 observation
         iris_df.head(10)
Out[21]:
            sepal length (cm) sepal width (cm) petal length (cm) petal width (cm)
         0
                                                1.4
                      5.1
                                   3.5
                                                            0.2
         1
                      4.9
                                   3.0
                                                1.4
                                                            0.2
          2
                      4.7
                                   3.2
                                                1.3
                                                             0.2
          3
                      4.6
                                   3.1
                                                1.5
                                                             0.2
                                   3.6
                                                1.4
                      5.0
                                                             0.2
          5
                      5.4
                                   3.9
                                                1.7
                                                            0.4
                                   3.4
                                                1.4
                      4.6
                                                             0.3
          7
                      5.0
                                   3.4
                                                1.5
                                                             0.2
                      4.4
                                   2.9
                                                1.4
                                                             0.2
          9
                      4.9
                                   3.1
                                                1.5
                                                             0.1
In [22]: #Reading last 10 observation
         iris_df.tail(10)
Out[22]:
              sepal length (cm) sepal width (cm) petal length (cm) petal width (cm)
          140
                        6.7
                                    3.1
                                                  5.6
                                                              2.4
          141
                        6.9
                                    3.1
                                                              2.3
                                                  5.1
          142
                        5.8
                                    2.7
                                                              1.9
          143
                                                  5.9
                        6.8
                                    3.2
                                                              2.3
                                                  5.7
          144
                        6.7
                                    3.3
                                                              2.5
          145
                        6.7
                                    3.0
                                                  5.2
                                                              2.3
          146
                        6.3
                                    2.5
                                                  5.0
                                                              1.9
          147
                        6.5
                                    3.0
                                                  5.2
                                                              2.0
          148
                        6.2
                                    3.4
                                                  5.4
                                                              2.3
          149
                        5.9
                                    3.0
                                                  5.1
                                                              1.8
In [23]: #cheking numerical data
         iris_df.describe()
Out[23]:
               sepal length (cm) sepal width (cm) petal length (cm) petal width (cm)
                   150.000000
                                                          150.000000
                                150.000000
                                              150.000000
          count
                     5.843333
                                                            1.199333
          mean
                                  3.057333
                                               3.758000
                     0.828066
                                  0.435866
                                               1.765298
                                                            0.762238
            std
           min
                     4.300000
                                  2.000000
                                               1.000000
                                                            0.100000
           25%
                     5.100000
                                  2.800000
                                               1.600000
                                                            0.300000
                     5.800000
                                  3.000000
                                               4.350000
                                                            1.300000
           50%
                                                            1.800000
           75%
                     6.400000
                                  3.300000
                                               5.100000
                     7.900000
                                  4.400000
                                                            2.500000
           max
                                               6.900000
         #checking null value in dataset
         iris_df.isnull().sum()
Out[24]: sepal length (cm)
         sepal width (cm)
                              0
         petal length (cm)
                              0
         petal width (cm)
                              0
         dtype: int64
In [25]: #plotting feature in line graph
         iris_df.plot(kind='line')
Out[25]: <matplotlib.axes._subplots.AxesSubplot at 0x1e9c2265188>
                sepal length (cm)
                                          120
In [26]: x=iris_df.iloc[:, [0,1,2,3]].values
         Using the Elbow Method to find the optimum number of clusters
In [28]: from sklearn.cluster import KMeans
In [29]: wcss = []
         for i in range(1, 11):
             kmeans = KMeans(n_clusters = i, init = 'k-means++',
                             max_iter = 300, n_init = 10, random_state = 0)
             kmeans.fit(x)
             wcss.append(kmeans.inertia_)
         # Plotting the results onto a line graph,
         # `allowing us to observe 'The elbow'
         plt.plot(range(1, 11), wcss)
         plt.title('The elbow method')
         plt.xlabel('Number of clusters')
         plt.ylabel('WCSS') # Within cluster sum of squares
         plt.show()
                            The elbow method
            700
            600
            500
          SS 400
300
            200
            100
                             Number of clusters
         Apply K-Means to dataset
In [30]: | # Applying kmeans to the dataset / Creating the kmeans classifier
         kmeans = KMeans(n_clusters = 3, init = 'k-means++',
                         max_iter = 300, n_init = 10, random_state = 0)
         y_kmeans = kmeans.fit_predict(x)
In [31]: #printing y_kmeans
         y_kmeans
2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 0, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
                2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                0, 0, 0, 2, 2, 0, 0, 0, 0, 2, 0, 2, 0, 2, 0, 0, 2, 2, 0, 0, 0, 0,
                0, 2, 0, 0, 0, 0, 2, 0, 0, 0, 2, 0, 0, 0, 2, 0, 0, 2])
         Visualizing the cluster
In [32]: # Visualising the clusters - On the first two columns
         plt.scatter(x[y_kmeans == 0, 0], x[y_kmeans == 0, 1],
                     s = 100, c = 'red', label = 'Iris-setosa')
         plt.scatter(x[y_kmeans == 1, 0], x[y_kmeans == 1, 1],
                     s = 100, c = 'blue', label = 'Iris-versicolour')
         plt.scatter(x[y_kmeans == 2, 0], x[y_kmeans == 2, 1],
                     s = 100, c = 'green', label = 'Iris-virginica')
         # Plotting the centroids of the clusters
         plt.scatter(kmeans.cluster_centers_[:, 0], kmeans.cluster_centers_[:,1],
                     s = 100, c = 'yellow', label = 'Centroids')
         plt.legend()
Out[32]: <matplotlib.legend.Legend at 0x1e9c2507b88>
          3.5
          2.5
          2.0
                                           Centroids
                                         7.0
                                              7.5
         CONCLUSION: The optimum number of cluster is 3 in the Iris dataset
         Thank You!
```

In []: