Species Segmentation with Cluster Analysis

Data Science and Business Analytics intern -The spark foundation

By- Srishti Chopra

Import the relevant libraries

In [1]:
```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
sns.set()
from sklearn.cluster import KMeans
```

Load the data

Load data from the csv file: '*iris_dataset.csv*'.

In [2]:
```python
# Load the data
data = pd.read_csv('iris_dataset.csv')
data
```

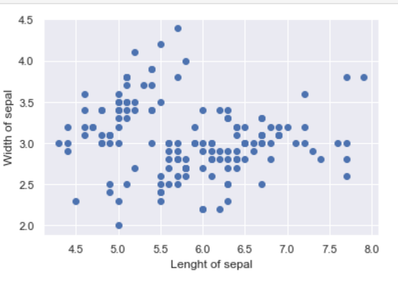
Out[2]:

	sepal_length	sepal_width	petal_length	petal_width
0	5.1	3.5	1.4	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
4	5.0	3.6	1.4	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	6.2	3.4	5.4	2.3
149	5.9	3.0	5.1	1.8

150 rows × 4 columns

Plot the data

In [3]:
```python
# Create a scatter plot based on two corresponding features
plt.scatter(data['sepal_length'],data['sepal_width'])
plt.xlabel('Lenght of sepal')
plt.ylabel('Width of sepal')
plt.show()
```



Standardize the variables

In [4]:
```python
# import some preprocessing module
from sklearn import preprocessing

# scale the data for better results
x_scaled = preprocessing.scale(data)
x_scaled
```

Out[4]:
```
array([[-9.00681170e-01,  1.03205722e+00, -1.34127240e+00,
        -1.31297673e+00],
       [-1.14301691e+00, -1.24957601e-01, -1.34127240e+00,
        -1.31297673e+00],
       [-1.38535265e+00,  3.37848329e-01, -1.39813811e+00,
        -1.31297673e+00],
       [-1.50652052e+00,  1.06445364e-01, -1.28440670e+00,
        -1.31297673e+00],
       [-1.02184904e+00,  1.26346019e+00, -1.34127240e+00,
        -1.31297673e+00],
       [-5.37177559e-01,  1.95766909e+00, -1.17067529e+00,
        -1.05003079e+00],
       [-1.50652052e+00,  8.00654259e-01, -1.34127240e+00,
        -1.18150376e+00],
       [-1.02184904e+00,  8.00654259e-01, -1.28440670e+00,
        -1.31297673e+00],
       [-1.74885626e+00, -3.56360566e-01, -1.34127240e+00,
        -1.31297673e+00],
       [-1.14301691e+00,  1.06445364e-01, -1.28440670e+00,
        -1.44444970e+00],
       [-5.37177559e-01,  1.49486315e+00, -1.28440670e+00,
        -1.31297673e+00],
       [-1.26418478e+00,  8.00654259e-01, -1.22754100e+00,
        -1.31297673e+00],
       [-1.26418478e+00, -1.24957601e-01, -1.34127240e+00,
        -1.44444970e+00],
       [-1.87002413e+00, -1.24957601e-01, -1.51186952e+00,
        -1.44444970e+00],
       [-5.25060772e-02,  2.18907205e+00, -1.45500381e+00,
        -1.31297673e+00],
       [-1.73673948e-01,  3.11468391e+00, -1.28440670e+00,
        -1.05003079e+00],
       [-5.37177559e-01,  1.95766909e+00, -1.39813811e+00,
        -1.05003079e+00],
       [-9.00681170e-01,  1.03205722e+00, -1.34127240e+00,
        -1.18150376e+00],
       [-1.73673948e-01,  1.72626612e+00, -1.17067529e+00,
        -1.18150376e+00],
       [-9.00681170e-01,  1.72626612e+00, -1.28440670e+00,
        -1.18150376e+00],
       [-5.37177559e-01,  8.00654259e-01, -1.17067529e+00,
        -1.31297673e+00],
       [-9.00681170e-01,  1.49486315e+00, -1.28440670e+00,
        -1.05003079e+00],
       [-1.50652052e+00,  1.26346019e+00, -1.56873522e+00,
        -1.31297673e+00],
       [-9.00681170e-01,  5.69251294e-01, -1.17067529e+00,
        -9.18557817e-01],
       [-1.26418478e+00,  8.00654259e-01, -1.05694388e+00,
        -1.31297673e+00],
       [-1.02184904e+00, -1.24957601e-01, -1.22754100e+00,
        -1.31297673e+00],
       [-1.02184904e+00,  8.00654259e-01, -1.22754100e+00,
        -1.05003079e+00],
       [-7.79513300e-01,  1.03205722e+00, -1.28440670e+00,
        -1.31297673e+00],
       [-7.79513300e-01,  8.00654259e-01, -1.34127240e+00,
        -1.31297673e+00],
       [-1.38535265e+00,  3.37848329e-01, -1.22754100e+00,
        -1.31297673e+00],
       [-1.26418478e+00,  1.06445364e-01, -1.22754100e+00,
        -1.31297673e+00],
       [-5.37177559e-01,  8.00654259e-01, -1.28440670e+00,
        -1.05003079e+00],
       [-7.79513300e-01,  2.42047502e+00, -1.28440670e+00,
        -1.44444970e+00],
       [-4.16009689e-01,  2.65187798e+00, -1.34127240e+00,
        -1.31297673e+00],
       [-1.14301691e+00,  1.06445364e-01, -1.28440670e+00,
        -1.44444970e+00],
       [-1.02184904e+00,  3.37848329e-01, -1.45500381e+00,
        -1.31297673e+00],
       [-4.16009689e-01,  1.03205722e+00, -1.39813811e+00,
        -1.31297673e+00],
       [-1.14301691e+00,  1.06445364e-01, -1.28440670e+00,
        -1.44444970e+00],
       [-1.74885626e+00, -1.24957601e-01, -1.39813811e+00,
        -1.31297673e+00],
       [-9.00681170e-01,  8.00654259e-01, -1.28440670e+00,
        -1.31297673e+00],
       [-1.02184904e+00,  1.03205722e+00, -1.39813811e+00,
        -1.18150376e+00],
       [-1.62768839e+00, -1.74477836e+00, -1.39813811e+00,
        -1.18150376e+00],
       [-1.74885626e+00,  3.37848329e-01, -1.39813811e+00,
        -1.31297673e+00],
       [-1.02184904e+00,  1.03205722e+00, -1.22754100e+00,
        -7.87084847e-01],
       [-9.00681170e-01,  1.72626612e+00, -1.05694388e+00,
        -1.05003079e+00],
       [-1.26418478e+00, -1.24957601e-01, -1.34127240e+00,
        -1.18150376e+00],
       [-9.00681170e-01,  1.72626612e+00, -1.22754100e+00,
        -1.31297673e+00],
       [-1.50652052e+00,  3.37848329e-01, -1.34127240e+00,
        -1.31297673e+00],
       [-6.58345429e-01,  1.49486315e+00, -1.28440670e+00,
        -1.31297673e+00],
       [-1.02184904e+00,  5.69251294e-01, -1.34127240e+00,
        -1.31297673e+00],
       [ 1.40150837e+00,  3.37848329e-01,  5.35295827e-01,
         2.64698913e-01],
       [ 6.74501145e-01,  3.37848329e-01,  4.21564419e-01,
         3.96171883e-01],
       [ 1.28034050e+00,  1.06445364e-01,  6.49027235e-01,
         3.96171883e-01],
       [-4.16009689e-01, -1.74477836e+00,  1.37235899e-01,
         1.33225943e-01],
       [ 7.95669016e-01, -5.87763531e-01,  4.78430123e-01,
         3.96171883e-01],
       [-1.73673948e-01, -5.87763531e-01,  4.21564419e-01,
         1.33225943e-01],
       [ 5.53333275e-01,  5.69251294e-01,  5.35295827e-01,
         5.27644853e-01],
       [-1.14301691e+00, -1.51337539e+00, -2.60824029e-01,
        -2.61192967e-01],
       [ 9.16836886e-01, -3.56360566e-01,  4.78430123e-01,
         1.33225943e-01],
       [-7.79513300e-01, -8.19166497e-01,  8.03701950e-02,
         2.64698913e-01],
       [-1.02184904e+00, -2.43898725e+00, -1.47092621e-01,
        -2.61192967e-01],
       [ 6.86617933e-02, -1.24957601e-01,  2.50967307e-01,
         3.96171883e-01],
       [ 1.89829664e-01, -1.97618132e+00,  1.37235899e-01,
        -2.61192967e-01],
       [ 3.10997534e-01, -3.56360566e-01,  5.35295827e-01,
         2.64698913e-01],
       [-2.94841818e-01, -3.56360566e-01, -9.02269170e-02,
         1.33225943e-01],
       [ 1.03800476e+00,  1.06445364e-01,  3.64698715e-01,
         2.64698913e-01],
       [-2.94841818e-01, -1.24957601e-01,  4.21564419e-01,
         3.96171883e-01],
       [-5.25060772e-02, -8.19166497e-01,  1.94101603e-01,
        -2.61192967e-01],
       [ 4.32165405e-01, -1.97618132e+00,  4.21564419e-01,
         3.96171883e-01],
       [-2.94841818e-01, -1.28197243e+00,  8.03701950e-02,
        -1.29719997e-01],
       [ 6.86617933e-02,  3.37848329e-01,  5.92161531e-01,
         7.90590793e-01],
       [ 3.10997534e-01, -5.87763531e-01,  1.37235899e-01,
         1.33225943e-01],
       [ 5.53333275e-01, -1.28197243e+00,  6.49027235e-01,
         3.96171883e-01],
       [ 3.10997534e-01, -5.87763531e-01,  5.35295827e-01,
         1.75297293e-03],
       [ 6.74501145e-01, -3.56360566e-01,  3.07833011e-01,
         1.33225943e-01],
       [ 9.16836886e-01, -1.24957601e-01,  3.64698715e-01,
         2.64698913e-01],
       [ 1.15917263e+00, -5.87763531e-01,  5.92161531e-01,
         2.64698913e-01],
       [ 1.03800476e+00, -1.24957601e-01,  7.05892939e-01,
         6.59117823e-01],
       [ 1.89829664e-01, -3.56360566e-01,  4.21564419e-01,
         3.96171883e-01],
       [-1.73673948e-01, -1.05056946e+00, -1.47092621e-01,
        -2.61192967e-01],
       [-4.16009689e-01, -1.51337539e+00,  2.35044910e-02,
        -1.29719997e-01],
       [-4.16009689e-01, -1.51337539e+00, -3.33612130e-02,
        -2.61192967e-01],
       [-5.25060772e-02, -8.19166497e-01,  8.03701950e-02,
         1.75297293e-03],
       [ 1.89829664e-01, -8.19166497e-01,  7.62758643e-01,
         5.27644853e-01],
       [-5.37177559e-01, -1.24957601e-01,  4.21564419e-01,
         3.96171883e-01],
       [ 1.89829664e-01,  8.00654259e-01,  4.21564419e-01,
         5.27644853e-01],
       [ 1.03800476e+00,  1.06445364e-01,  5.35295827e-01,
         3.96171883e-01],
       [ 5.53333275e-01, -1.74477836e+00,  3.64698715e-01,
         1.33225943e-01],
       [-2.94841818e-01, -1.24957601e-01,  1.94101603e-01,
         1.33225943e-01],
       [-4.16009689e-01, -1.28197243e+00,  1.37235899e-01,
         1.33225943e-01],
       [-4.16009689e-01, -1.05056946e+00,  3.64698715e-01,
         1.75297293e-03],
       [ 3.10997534e-01, -1.24957601e-01,  4.78430123e-01,
         2.64698913e-01],
       [-5.25060772e-02, -1.05056946e+00,  1.37235899e-01,
         1.75297293e-03],
       [-1.02184904e+00, -1.74477836e+00, -2.60824029e-01,
        -2.61192967e-01],
       [-2.94841818e-01, -8.19166497e-01,  2.50967307e-01,
         1.33225943e-01],
       [-1.73673948e-01, -1.24957601e-01,  2.50967307e-01,
         1.75297293e-03],
       [-1.73673948e-01, -3.56360566e-01,  2.50967307e-01,
         1.33225943e-01],
       [ 4.32165405e-01, -3.56360566e-01,  3.07833011e-01,
         1.33225943e-01],
       [-9.00681170e-01, -1.28197243e+00, -4.31421141e-01,
        -1.29719997e-01],
       [-1.73673948e-01, -5.87763531e-01,  1.94101603e-01,
         1.33225943e-01],
       [ 5.53333275e-01,  5.69251294e-01,  1.27454998e+00,
         1.71090158e+00],
       [-5.25060772e-02, -8.19166497e-01,  7.62758643e-01,
         9.22063763e-01],
       [ 1.52267624e+00, -1.24957601e-01,  1.21768427e+00,
         1.18500970e+00],
       [ 5.53333275e-01, -3.56360566e-01,  1.04708716e+00,
         7.90590793e-01],
       [ 7.95669016e-01, -1.24957601e-01,  1.16081857e+00,
         1.31648267e+00],
       [ 2.12851559e+00, -1.24957601e-01,  1.61574420e+00,
         1.18500970e+00],
       [-1.14301691e+00, -1.28197243e+00,  4.21564419e-01,
         6.59117823e-01],
       [ 1.76501198e+00, -3.56360566e-01,  1.44514709e+00,
         7.90590793e-01],
       [ 1.03800476e+00, -1.28197243e+00,  1.16081857e+00,
         7.90590793e-01],
       [ 1.64384411e+00,  1.26346019e+00,  1.33141568e+00,
         1.71090158e+00],
       [ 7.95669016e-01,  3.37848329e-01,  7.62758643e-01,
         1.05353673e+00],
       [ 6.74501145e-01, -8.19166497e-01,  8.76490051e-01,
         9.22063763e-01],
       [ 1.15917263e+00, -1.24957601e-01,  9.90221459e-01,
         1.18500970e+00],
       [-1.73673948e-01, -1.28197243e+00,  7.05892939e-01,
         1.05353673e+00],
       [-5.25060772e-02, -5.87763531e-01,  7.62758643e-01,
         1.57942861e+00],
       [ 6.74501145e-01,  3.37848329e-01,  8.76490051e-01,
         1.44795564e+00],
       [ 7.95669016e-01, -1.24957601e-01,  9.90221459e-01,
         7.90590793e-01],
       [ 2.24968346e+00,  1.72626612e+00,  1.67260991e+00,
         1.31648267e+00],
       [ 2.24968346e+00, -1.05056946e+00,  1.78634131e+00,
         1.44795564e+00],
       [ 1.89829664e-01, -1.97618132e+00,  7.05892939e-01,
         3.96171883e-01],
       [ 1.28034050e+00,  3.37848329e-01,  1.10395287e+00,
         1.44795564e+00],
       [-2.94841818e-01, -5.87763531e-01,  6.49027235e-01,
         1.05353673e+00],
       [ 2.24968346e+00, -5.87763531e-01,  1.67260991e+00,
         1.05353673e+00],
       [ 5.53333275e-01, -8.19166497e-01,  6.49027235e-01,
         7.90590793e-01],
       [ 1.03800476e+00,  5.69251294e-01,  1.10395287e+00,
         1.18500970e+00],
       [ 1.64384411e+00,  3.37848329e-01,  1.27454998e+00,
         7.90590793e-01],
       [ 4.32165405e-01, -5.87763531e-01,  5.92161531e-01,
         7.90590793e-01],
       [ 3.10997534e-01, -1.24957601e-01,  6.49027235e-01,
         7.90590793e-01],
       [ 6.74501145e-01, -5.87763531e-01,  1.04708716e+00,
         1.18500970e+00],
       [ 1.64384411e+00, -1.24957601e-01,  1.16081857e+00,
         5.27644853e-01],
       [ 1.88617985e+00, -5.87763531e-01,  1.33141568e+00,
         9.22063763e-01],
       [ 2.49201920e+00,  1.72626612e+00,  1.50201279e+00,
         1.05353673e+00],
       [ 6.74501145e-01, -5.87763531e-01,  1.04708716e+00,
         1.31648267e+00],
       [ 5.53333275e-01, -5.87763531e-01,  7.62758643e-01,
         3.96171883e-01],
       [ 3.10997534e-01, -1.05056946e+00,  1.04708716e+00,
         2.64698913e-01],
       [ 2.24968346e+00, -1.24957601e-01,  1.33141568e+00,
         1.44795564e+00],
       [ 5.53333275e-01,  8.00654259e-01,  1.04708716e+00,
         1.57942861e+00],
       [ 6.74501145e-01,  1.06445364e-01,  9.90221459e-01,
         7.90590793e-01],
       [ 1.89829664e-01, -1.24957601e-01,  5.92161531e-01,
         7.90590793e-01],
       [ 1.28034050e+00,  1.06445364e-01,  9.33355755e-01,
         1.18500970e+00],
       [ 1.03800476e+00,  1.06445364e-01,  1.04708716e+00,
         1.57942861e+00],
       [ 1.28034050e+00,  1.06445364e-01,  7.62758643e-01,
         1.44795564e+00],
       [-5.25060772e-02, -8.19166497e-01,  7.62758643e-01,
         9.22063763e-01],
       [ 1.15917263e+00,  3.37848329e-01,  1.21768427e+00,
         1.44795564e+00],
       [ 1.03800476e+00,  5.69251294e-01,  1.10395287e+00,
         1.71090158e+00],
       [ 1.03800476e+00, -1.24957601e-01,  8.19624347e-01,
         1.44795564e+00],
       [ 5.53333275e-01, -1.28197243e+00,  7.05892939e-01,
         9.22063763e-01],
       [ 7.95669016e-01, -1.24957601e-01,  8.19624347e-01,
         1.05353673e+00],
       [ 4.32165405e-01,  8.00654259e-01,  9.33355755e-01,
         1.44795564e+00],
       [ 6.86617933e-02, -1.24957601e-01,  7.62758643e-01,
         7.90590793e-01]])
```

Take Advantage of the Elbow Method

WCSS

In [5]:
```python
wcss = []
# 'cl_num' is a that keeps track the highest number of clusters we want to use the WCSS method for. We have it
cl_num = 10
for i in range (1,cl_num):
    kmeans= KMeans(i)
    kmeans.fit(x_scaled)
    wcss_iter = kmeans.inertia_
    wcss.append(wcss_iter)
wcss
```

```
C:\Users\COMAP\Downloads\Anaconda\lib\site-packages\sklearn\cluster\_kmeans.py:881: UserWarning: KMeans is know
n to have a memory leak on Windows with MKL, when there are less chunks than available threads. You can avoid i
t by setting the environment variable OMP_NUM_THREADS=1.
  warnings.warn(
```

Out[5]:
```
[600.0000000000003,
 223.73200573676343,
 140.96581663074699,
 114.42970777082239,
 91.12763977985172,
 80.24972235577128,
 71.3007056411862,
 63.01282906393713,
 54.60041897786318]
```

The Elbow Method

In [6]:
```python
# Finding the optimum number of clusters for k-means classification
number_clusters = range(1,cl_num)
# Plotting the results onto a line graph,
# `allowing us to observe 'The elbow'
plt.plot(number_clusters, wcss)
plt.title('The Elbow Method')
plt.xlabel('Number of clusters')
plt.ylabel('Within-cluster Sum of Squares')
```

Out[6]:
```
Text(0, 0.5, 'Within-cluster Sum of Squares')
```

Understanding the Elbow Curve

Based on the Elbow Curve, 3 seem the most likely.

3 Clusters

In [7]:
```python
kmeans_3 = KMeans(3)
kmeans_3.fit(x_scaled)
```

Out[7]:
```
KMeans(n_clusters=3)
```

In [8]:
```python
clusters_3 = data.copy()
clusters_3['cluster_pred']=kmeans_3.fit_predict(x_scaled)
```

In [9]:
```python
plt.scatter(clusters_3['sepal_length'], clusters_3['sepal_width'],\
            c= clusters_3 ['cluster_pred'], cmap = 'rainbow')
```

Out[9]:
```
<matplotlib.collections.PathCollection at 0x25133492fa0>
```