DATA SCIENCE AND BUSINESS ANALYTICS AT THE SPARK FOUNDATION

Girija Kumaran

TASK 2

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

In [2]:
```
df = pd.read_csv("C:/Users/Girija/Downloads/Iris.csv")
print("The data set is loaded successfully!!")
pd.set_option("display.max_rows",None)
print(df)
```

```
The data set is loaded successfully!!
      Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm  \
0      1            5.1           3.5            1.4           0.2
1      2            4.9           3.0            1.4           0.2
2      3            4.7           3.2            1.3           0.2
3      4            4.6           3.1            1.5           0.2
4      5            5.0           3.6            1.4           0.2
5      6            5.4           3.9            1.7           0.4
6      7            4.6           3.4            1.4           0.3
7      8            5.0           3.4            1.5           0.2
8      9            4.4           2.9            1.4           0.2
9     10            4.9           3.1            1.5           0.1
10    11            5.4           3.7            1.5           0.2
11    12            4.8           3.4            1.6           0.2
12    13            4.8           3.0            1.4           0.1
13    14            4.3           3.0            1.1           0.1
14    15            5.8           4.0            1.2           0.2
15    16            5.7           4.4            1.5           0.4
16    17            5.4           3.9            1.3           0.4
17    18            5.1           3.5            1.4           0.3
18    19            5.7           3.8            1.7           0.3
19    20            5.1           3.8            1.5           0.3
20    21            5.4           3.4            1.7           0.2
21    22            5.1           3.7            1.5           0.4
22    23            4.6           3.6            1.0           0.2
23    24            5.1           3.3            1.7           0.5
24    25            4.8           3.4            1.9           0.2
25    26            5.0           3.0            1.6           0.2
26    27            5.0           3.4            1.6           0.4
27    28            5.2           3.5            1.5           0.2
28    29            5.2           3.4            1.4           0.2
29    30            4.7           3.2            1.6           0.2
30    31            4.8           3.1            1.6           0.2
31    32            5.4           3.4            1.5           0.4
32    33            5.2           4.1            1.5           0.1
33    34            5.5           4.2            1.4           0.2
34    35            4.9           3.1            1.5           0.1
35    36            5.0           3.2            1.2           0.2
36    37            5.5           3.5            1.3           0.2
37    38            4.9           3.1            1.5           0.1
38    39            4.4           3.0            1.3           0.2
39    40            5.1           3.4            1.5           0.2
40    41            5.0           3.5            1.3           0.3
41    42            4.5           2.3            1.3           0.3
42    43            4.4           3.2            1.3           0.2
43    44            5.0           3.5            1.6           0.6
44    45            5.1           3.8            1.9           0.4
45    46            4.8           3.0            1.4           0.3
46    47            5.1           3.8            1.6           0.2
47    48            4.6           3.2            1.4           0.2
48    49            5.3           3.7            1.5           0.2
49    50            5.0           3.3            1.4           0.2
50    51            7.0           3.2            4.7           1.4
51    52            6.4           3.2            4.5           1.5
52    53            6.9           3.1            4.9           1.5
53    54            5.5           2.3            4.0           1.3
54    55            6.5           2.8            4.6           1.5
55    56            5.7           2.8            4.5           1.3
```

56	57	6.3	3.3	4.7	1.6
57	58	4.9	2.4	3.3	1.0
58	59	6.6	2.9	4.6	1.3
59	60	5.2	2.7	3.9	1.4
60	61	5.0	2.0	3.5	1.0
61	62	5.9	3.0	4.2	1.5
62	63	6.0	2.2	4.0	1.0
63	64	6.1	2.9	4.7	1.4
64	65	5.6	2.9	3.6	1.3
65	66	6.7	3.1	4.4	1.4
66	67	5.6	3.0	4.5	1.5
67	68	5.8	2.7	4.1	1.0
68	69	6.2	2.2	4.5	1.5
69	70	5.6	2.5	3.9	1.1
70	71	5.9	3.2	4.8	1.8
71	72	6.1	2.8	4.0	1.3
72	73	6.3	2.5	4.9	1.5
73	74	6.1	2.8	4.7	1.2
74	75	6.4	2.9	4.3	1.3
75	76	6.6	3.0	4.4	1.4
76	77	6.8	2.8	4.8	1.4
77	78	6.7	3.0	5.0	1.7
78	79	6.0	2.9	4.5	1.5
79	80	5.7	2.6	3.5	1.0
80	81	5.5	2.4	3.8	1.1
81	82	5.5	2.4	3.7	1.0
82	83	5.8	2.7	3.9	1.2
83	84	6.0	2.7	5.1	1.6
84	85	5.4	3.0	4.5	1.5
85	86	6.0	3.4	4.5	1.6
86	87	6.7	3.1	4.7	1.5
87	88	6.3	2.3	4.4	1.3
88	89	5.6	3.0	4.1	1.3
89	90	5.5	2.5	4.0	1.3
90	91	5.5	2.6	4.4	1.2
91	92	6.1	3.0	4.6	1.4
92	93	5.8	2.6	4.0	1.2
93	94	5.0	2.3	3.3	1.0
94	95	5.6	2.7	4.2	1.3
95	96	5.7	3.0	4.2	1.2
96	97	5.7	2.9	4.2	1.3
97	98	6.2	2.9	4.3	1.3
98	99	5.1	2.5	3.0	1.1
99	100	5.7	2.8	4.1	1.3
100	101	6.3	3.3	6.0	2.5
101	102	5.8	2.7	5.1	1.9
102	103	7.1	3.0	5.9	2.1
103	104	6.3	2.9	5.6	1.8
104	105	6.5	3.0	5.8	2.2
105	106	7.6	3.0	6.6	2.1
106	107	4.9	2.5	4.5	1.7
107	108	7.3	2.9	6.3	1.8
108	109	6.7	2.5	5.8	1.8
109	110	7.2	3.6	6.1	2.5
110	111	6.5	3.2	5.1	2.0
111	112	6.4	2.7	5.3	1.9
112	113	6.8	3.0	5.5	2.1
113	114	5.7	2.5	5.0	2.0
114	115	5.8	2.8	5.1	2.4
115	116	6.4	3.2	5.3	2.3
116	117	6.5	3.0	5.5	1.8
117	118	7.7	3.8	6.7	2.2
118	119	7.7	2.6	6.9	2.3
119	120	6.0	2.2	5.0	1.5
120	121	6.9	3.2	5.7	2.3
121	122	5.6	2.8	4.9	2.0
122	123	7.7	2.8	6.7	2.0
123	124	6.3	2.7	4.9	1.8
124	125	6.7	3.3	5.7	2.1
125	126	7.2	3.2	6.0	1.8
126	127	6.2	2.8	4.8	1.8
127	128	6.1	3.0	4.9	1.8
128	129	6.4	2.8	5.6	2.1
129	130	7.2	3.0	5.8	1.6
130	131	7.4	2.8	6.1	1.9
131	132	7.9	3.8	6.4	2.0
132	133	6.4	2.8	5.6	2.2
133	134	6.3	2.8	5.1	1.5
134	135	6.1	2.6	5.6	1.4
135	136	7.7	3.0	6.1	2.3
136	137	6.3	3.4	5.6	2.4
137	138	6.4	3.1	5.5	1.8
138	139	6.0	3.0	4.8	1.8
139	140	6.9	3.1	5.4	2.1
140	141	6.7	3.1	5.6	2.4
141	142	6.9	3.1	5.1	2.3
142	143	5.8	2.7	5.1	1.9
143	144	6.8	3.2	5.9	2.3
144	145	6.7	3.3	5.7	2.5

```
145  146           6.7          3.0           5.2          2.3
146  147           6.3          2.5           5.0          1.9
147  148           6.5          3.0           5.2          2.0
148  149           6.2          3.4           5.4          2.3
149  150           5.9          3.0           5.1          1.8

            Species
0       Iris-setosa
1       Iris-setosa
2       Iris-setosa
3       Iris-setosa
4       Iris-setosa
5       Iris-setosa
6       Iris-setosa
7       Iris-setosa
8       Iris-setosa
9       Iris-setosa
10      Iris-setosa
11      Iris-setosa
12      Iris-setosa
13      Iris-setosa
14      Iris-setosa
15      Iris-setosa
16      Iris-setosa
17      Iris-setosa
18      Iris-setosa
19      Iris-setosa
20      Iris-setosa
21      Iris-setosa
22      Iris-setosa
23      Iris-setosa
24      Iris-setosa
25      Iris-setosa
26      Iris-setosa
27      Iris-setosa
28      Iris-setosa
29      Iris-setosa
30      Iris-setosa
31      Iris-setosa
32      Iris-setosa
33      Iris-setosa
34      Iris-setosa
35      Iris-setosa
36      Iris-setosa
37      Iris-setosa
38      Iris-setosa
39      Iris-setosa
40      Iris-setosa
41      Iris-setosa
42      Iris-setosa
43      Iris-setosa
44      Iris-setosa
45      Iris-setosa
46      Iris-setosa
47      Iris-setosa
48      Iris-setosa
49      Iris-setosa
50  Iris-versicolor
51  Iris-versicolor
52  Iris-versicolor
53  Iris-versicolor
54  Iris-versicolor
55  Iris-versicolor
56  Iris-versicolor
57  Iris-versicolor
58  Iris-versicolor
59  Iris-versicolor
60  Iris-versicolor
61  Iris-versicolor
62  Iris-versicolor
63  Iris-versicolor
64  Iris-versicolor
65  Iris-versicolor
66  Iris-versicolor
67  Iris-versicolor
68  Iris-versicolor
69  Iris-versicolor
70  Iris-versicolor
71  Iris-versicolor
72  Iris-versicolor
73  Iris-versicolor
74  Iris-versicolor
75  Iris-versicolor
76  Iris-versicolor
77  Iris-versicolor
78  Iris-versicolor
79  Iris-versicolor
80  Iris-versicolor
81  Iris-versicolor
```

```
82   Iris-versicolor
83   Iris-versicolor
84   Iris-versicolor
85   Iris-versicolor
86   Iris-versicolor
87   Iris-versicolor
88   Iris-versicolor
89   Iris-versicolor
90   Iris-versicolor
91   Iris-versicolor
92   Iris-versicolor
93   Iris-versicolor
94   Iris-versicolor
95   Iris-versicolor
96   Iris-versicolor
97   Iris-versicolor
98   Iris-versicolor
99   Iris-versicolor
100   Iris-virginica
101   Iris-virginica
102   Iris-virginica
103   Iris-virginica
104   Iris-virginica
105   Iris-virginica
106   Iris-virginica
107   Iris-virginica
108   Iris-virginica
109   Iris-virginica
110   Iris-virginica
111   Iris-virginica
112   Iris-virginica
113   Iris-virginica
114   Iris-virginica
115   Iris-virginica
116   Iris-virginica
117   Iris-virginica
118   Iris-virginica
119   Iris-virginica
120   Iris-virginica
121   Iris-virginica
122   Iris-virginica
123   Iris-virginica
124   Iris-virginica
125   Iris-virginica
126   Iris-virginica
127   Iris-virginica
128   Iris-virginica
129   Iris-virginica
130   Iris-virginica
131   Iris-virginica
132   Iris-virginica
133   Iris-virginica
134   Iris-virginica
135   Iris-virginica
136   Iris-virginica
137   Iris-virginica
138   Iris-virginica
139   Iris-virginica
140   Iris-virginica
141   Iris-virginica
142   Iris-virginica
143   Iris-virginica
144   Iris-virginica
145   Iris-virginica
146   Iris-virginica
147   Iris-virginica
148   Iris-virginica
149   Iris-virginica
```

In [3]:
```
df.head()
```

Out[3]:

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa
2	3	4.7	3.2	1.3	0.2	Iris-setosa
3	4	4.6	3.1	1.5	0.2	Iris-setosa
4	5	5.0	3.6	1.4	0.2	Iris-setosa

In [4]:
```
df.tail()
```

Out[4]:

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
145	146	6.7	3.0	5.2	2.3	Iris-virginica
146	147	6.3	2.5	5.0	1.9	Iris-virginica
147	148	6.5	3.0	5.2	2.0	Iris-virginica
148	149	6.2	3.4	5.4	2.3	Iris-virginica
149	150	5.9	3.0	5.1	1.8	Iris-virginica

In [5]:
```
df["Species"].unique()
```

Out[5]:
```
array(['Iris-setosa', 'Iris-versicolor', 'Iris-virginica'], dtype=object)
```

In [6]:
```
df.dtypes
```

Out[6]:
```
Id                int64
SepalLengthCm    float64
SepalWidthCm     float64
PetalLengthCm    float64
PetalWidthCm     float64
Species           object
dtype: object
```

In [7]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150 entries, 0 to 149
Data columns (total 6 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   Id             150 non-null    int64  
 1   SepalLengthCm  150 non-null    float64
 2   SepalWidthCm   150 non-null    float64
 3   PetalLengthCm  150 non-null    float64
 4   PetalWidthCm   150 non-null    float64
 5   Species        150 non-null    object 
dtypes: float64(4), int64(1), object(1)
memory usage: 7.2+ KB
```

In [8]:
```
df.shape
```

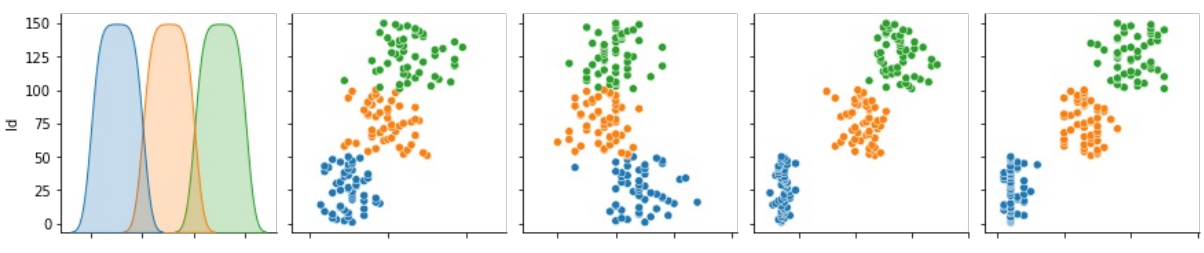
Out[8]:
```
(150, 6)
```

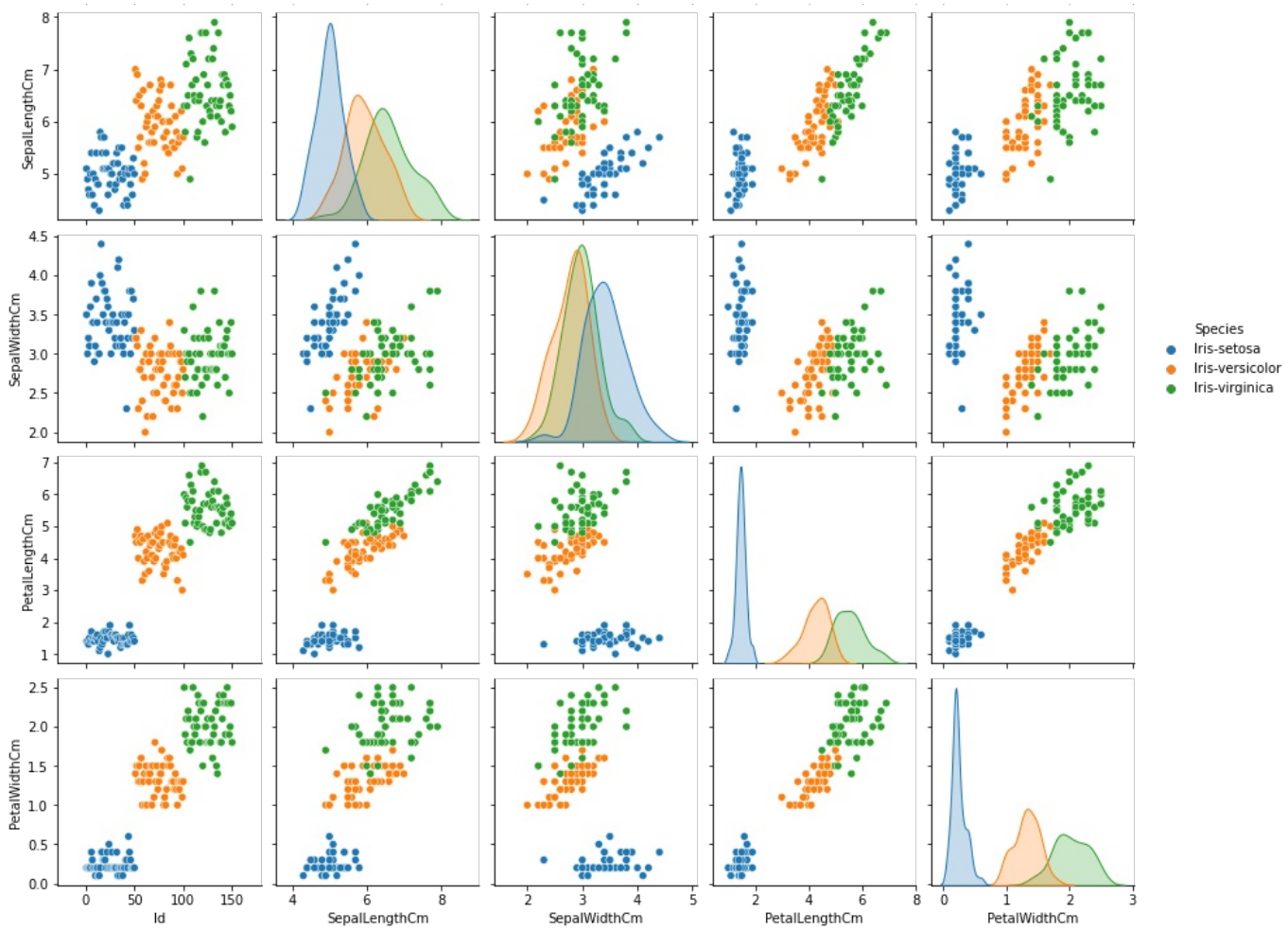
In [9]:
```
df.isnull().sum()
```

Out[9]:
```
Id               0
SepalLengthCm    0
SepalWidthCm     0
PetalLengthCm    0
PetalWidthCm     0
Species          0
dtype: int64
```

In [10]:
```
sns.pairplot(df,hue="Species")
```

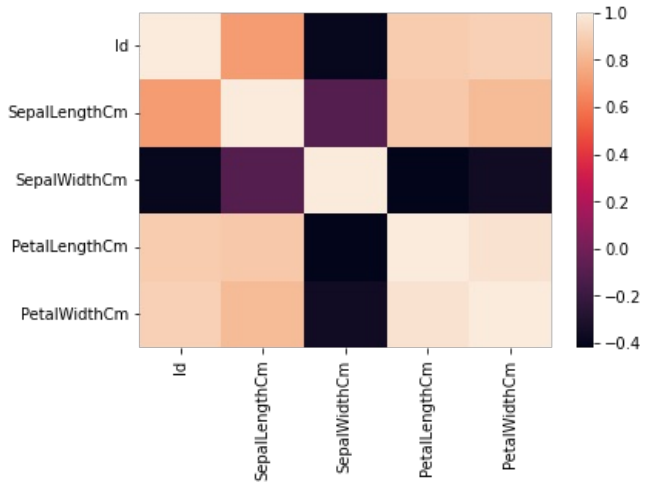
Out[10]:
```
<seaborn.axisgrid.PairGrid at 0x22fa6e0c7c0>
```





In [11]:
```
sns.heatmap(df.corr())
```

Out[11]:
<AxesSubplot:>



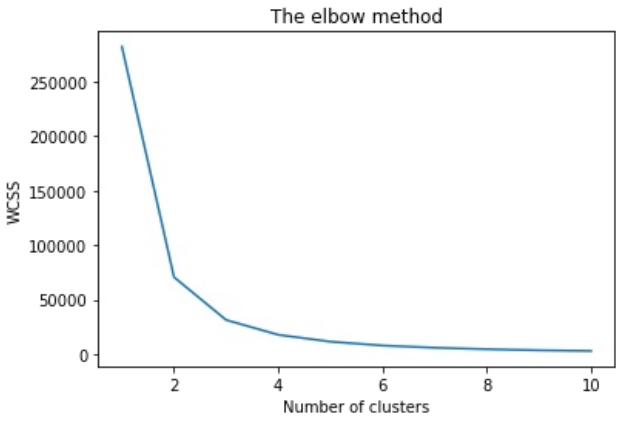
In [12]:
```
X = df.iloc[:, [0, 1, 2, 3]].values
```

In [13]:
```
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()
sns.set(rc={'figure.figsize':(5,5)})
```

```
C:\Users\Girija\anaconda3\lib\site-packages\sklearn\cluster\_kmeans.py:881: UserWarning: KMeans is known to have
a memory leak on Windows with MKL, when there are less chunks than available threads. You can avoid it by setting
the environment variable OMP_NUM_THREADS=1.
  warnings.warn(
```



In [14]:

```
kmeans = KMeans(n_clusters = 3, init = 'k-means++',
                max_iter = 300, n_init = 10, random_state = 0)
y_kmeans = kmeans.fit_predict(X)
y_kmeans
```

Out[14]:

```
array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
       2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
       2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1])
```

In [16]:

```
plt.scatter(X[y_kmeans == 0, 0], X[y_kmeans == 0, 1],
            s = 100, c = 'blue', label = 'Iris-setosa')
plt.scatter(X[y_kmeans == 1, 0], X[y_kmeans == 1, 1],
            s = 100, c = 'red', 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()

sns.set(rc={'figure.figsize':(10,8)})
```

THANK YOU !!