Hands-on Project

Question 1: List the dataset(s) you chose for this project from the UCI Machine Learning respository (https://archive.ics.uci.edu/ml/datasets.php).

Iris Dataset

Question 2: Describe the dataset in your own words. How many data points, how many attributes, how many types of attributes, how many classes (if any)? Who collected it? How was it collected?

In [2]:

```
# data loading and computing functionality
import pandas as pd
import numpy as np
import scipy as sp

# datasets in sklearn package
from sklearn.datasets import load_digits

# visualization packages
import seaborn as sns
import matplotlib.pyplot as plt
import matplotlib.cm as cm

#PCA, SVD, LDA
from sklearn.decomposition import PCA
from scipy.linalg import svd
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis

#Classification Algorithms
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVC
```

In [3]:

```
iris_df = pd.read_csv('https://raw.githubusercontent.com/plotly/datasets/maste
r/iris.csv')
```

In [4]:

```
iris_df.dtypes
```

Out[4]:

```
SepalLength    float64
SepalWidth     float64
PetalLength    float64
PetalWidth     float64
Name            object
dtype: object
```

```
In [5]: iris_df.shape

Out[5]: (150, 5)
```

Dataset contains 150 instances of 5 attributes describing the characteristics of the IRIS plants.There are 3 classes (Iris-Setosa, Iris-versicolor, Iris-virginica). (a) Creator: R.A. Fisher (b) Donor: Michael Marshall (MARSHALL%PLU@io.arc.nasa.gov)

Question 3: What is your goal? Specifically, what insights do you want to learn from this data. Please be aware that clustering, classification, or itemset mining are not 'insights'. These are data mining tasks. Insights are relevant to the domain from which the data is generated.

My goal is to group the plant species based upon the attributes(SepalLength,SepalWidth,PetalLength,PetalWidth).

Question 4: List the data mining task(s) and the specific algorithms you want to perform on this data. Do not pick the tasks listed in the 'Default Task' column on the UCI page.

Clustering(K-Means)

Question 5: Before selecting the methods you listed in response to Question 3, what are all methods you originally considered to use for the selected data mining task? What was your rationale for selecting the methods you listed in response to Question 3? What was your rationale for not selecting other methods?

Firstly I performed PCA on IRIS Dataset which gave me an intuitive idea to group the data based upon the reduced attributes. Post PCA, data appeared most like blobs and henced I selected K-Means algorithm to perform clustering as this type of data is well suited for K-Means. Apart from K-Means, there are also other algorithms like EM, DBScan, Spectral clustering... but K-Means is the simplest algorithm that can ouperform all these.

Question 6: What limitations does your 'selected' method(s) has(have) that may limit your ability to accomplish the goal you have set for yourself?

Other models are typical and complex in their implementation.

Question 7: Do you have any alternative plan/strategy to overcome the above limitation(s)?

K-Means which has lineat time complexity.

Question 8: For each of the methods you want to use, what parameter choices do you want to use and why? It does not have to be one parameter choice, it could be a collection or a range of choices you may want to consider.

K=3 because I want to group the IRIS data based upon the Name of the plant(which are of 3 types).

Question 9: How will you evaluate that you are successful in your pursuing your goal at the end of the project? In other words, what is your evaluation criteria?

Post Clustering I employ Cluster Evaluation Models.

Question 10: How will you evaluate that you are successful in your pursuing your goal at the end of the project? In other words, what is your evaluation criteria?

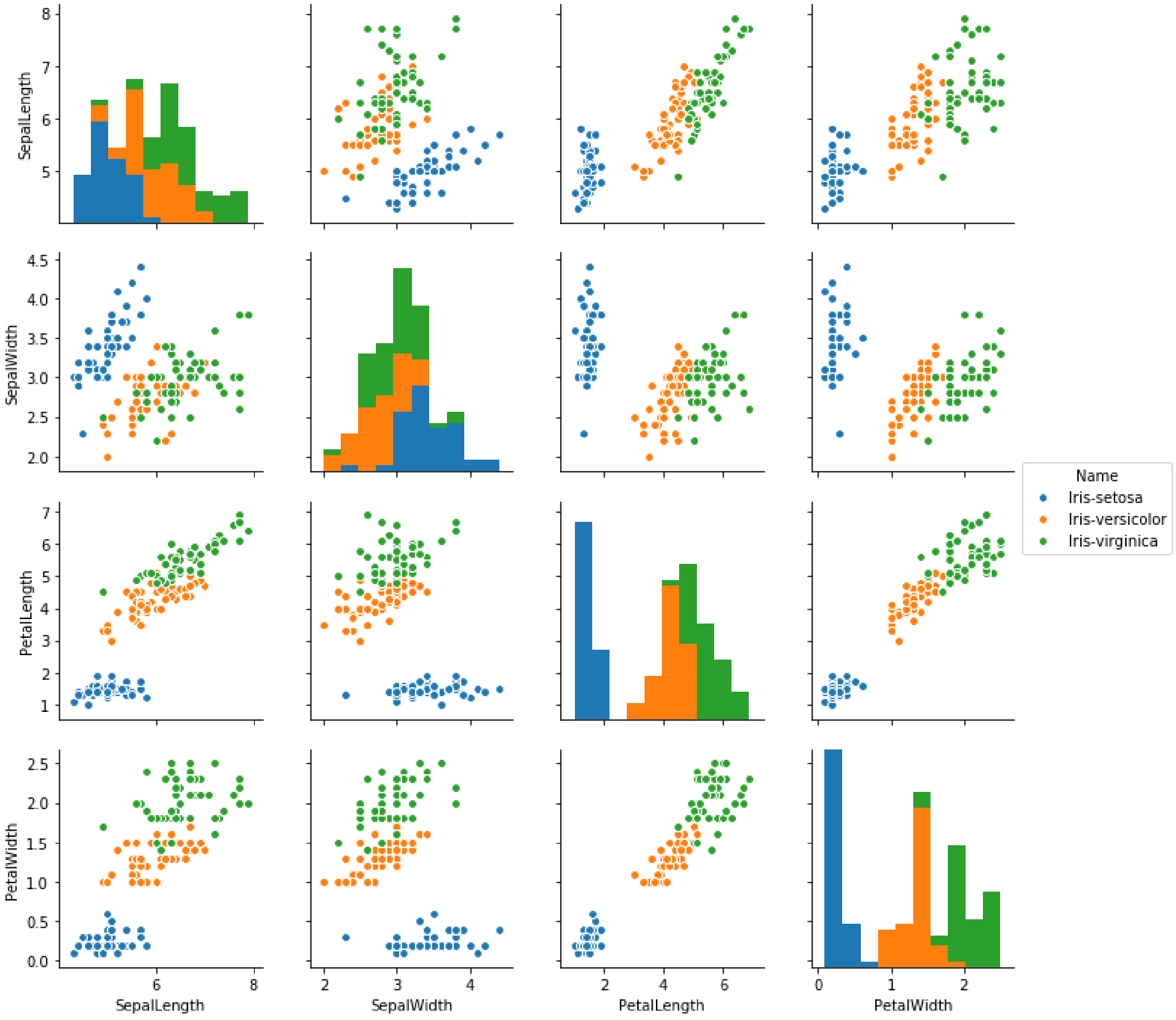
Post Clustering I employ Cluster Evaluation Models.

Question 11: Show any visualizations you may have generated to understand your data. Please include the code you used and the plots below. If you borrowed code (entirely or partially) from the hands-on projects or anywhere else, clearly provide a link to your source.

You may use this package to load UCI data in python: https://github.com/SkafteNicki/py_uci (https://github.com/SkafteNicki/py_uci)

```
In [6]: #Pairplots to figure out the correlation among attributes
        import seaborn as sns
        sns.pairplot(iris_df, hue="Name")

Out[6]: <seaborn.axisgrid.PairGrid at 0x2b449185dc50>
```

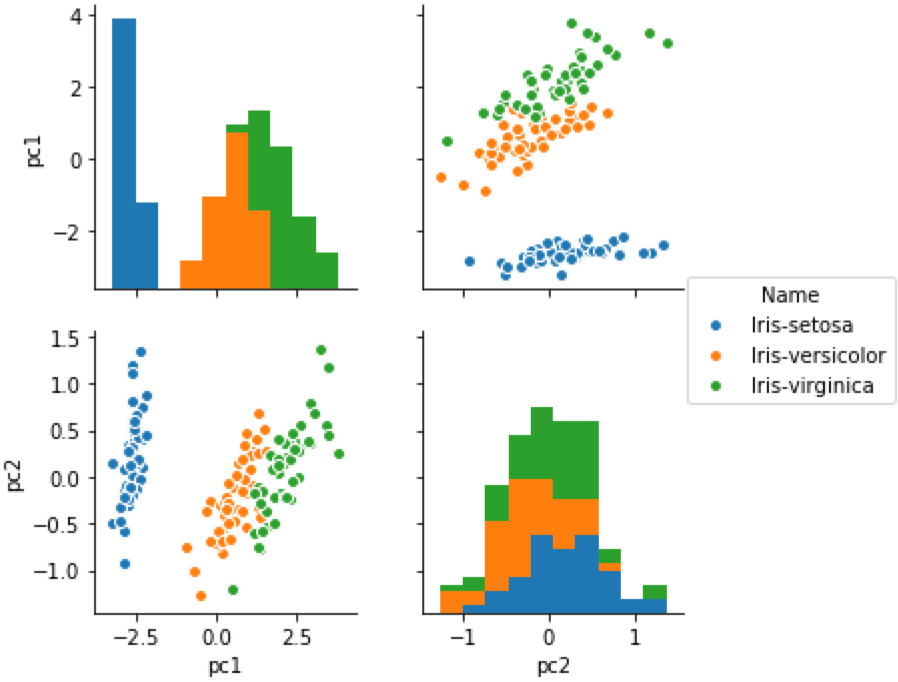
In [7]:
```
plt.show()
```



In [8]:
```
#PCA to reduce dimensions :
iris_data = iris_df.values[:,0:4]
pca = PCA(2)
projected = pca.fit_transform(iris_data)
print(projected.shape)
```

```
(150, 2)
```

In [9]:
```
p_columns = {'pc1': projected[:, 0].tolist(), 'pc2': projected[:, 1].tolist(),
'Name': iris_df['Name']}
p_data_df=pd.DataFrame(p_columns)
sns.pairplot(p_data_df,hue="Name")
plt.show()
```



Question 12: **Perform data mining, evaluate your work and report your findings.** This should include code, plots and results you may have generated. If you borrowed code (entirely or partially) from the hands-on projects or anywhere else, clearly provide a link to your source.

In [10]:
```
p_data_df.head()
```

Out[10]:

	pc1	pc2	Name
0	-2.684207	0.326607	Iris-setosa
1	-2.715391	-0.169557	Iris-setosa
2	-2.889820	-0.137346	Iris-setosa
3	-2.746437	-0.311124	Iris-setosa
4	-2.728593	0.333925	Iris-setosa

In [11]:

```
p_data_df.Name[p_data_df.Name == 'Iris-setosa'] = 0
p_data_df.Name[p_data_df.Name == 'Iris-versicolor'] = 1
p_data_df.Name[p_data_df.Name == 'Iris-virginica'] = 2
```

```
/usr/local/anaconda5/lib/python3.6/site-packages/ipykernel_launcher.py:1: Set
tingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/st
able/user_guide/indexing.html#returning-a-view-versus-a-copy
  """Entry point for launching an IPython kernel.
/usr/local/anaconda5/lib/python3.6/site-packages/ipykernel_launcher.py:2: Set
tingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/st
able/user_guide/indexing.html#returning-a-view-versus-a-copy

/usr/local/anaconda5/lib/python3.6/site-packages/ipykernel_launcher.py:3: Set
tingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/st
able/user_guide/indexing.html#returning-a-view-versus-a-copy
  This is separate from the ipykernel package so we can avoid doing imports u
ntil
```

In [12]:

```
p_data_df
```

Out[12]:

	pc1	pc2	Name
0	-2.684207	0.326607	0
1	-2.715391	-0.169557	0
2	-2.889820	-0.137346	0
3	-2.746437	-0.311124	0
4	-2.728593	0.333925	0
...	...	...	...
145	1.944017	0.187415	2
146	1.525664	-0.375021	2
147	1.764046	0.078519	2
148	1.901629	0.115877	2
149	1.389666	-0.282887	2

150 rows × 3 columns

In [40]:

```
#Hopkin's Stat for cluster tendency
iris_data_X = (p_data_df.values[:,0:2]).astype(int)
from sklearn.neighbors import NearestNeighbors
from random import sample
from numpy.random import uniform
from math import isnan

def hopkins(X):
    n = X.shape[0] #rows
    d = X.shape[1] #cols
    p = int(0.1 * n) #considering 10% of points
    nbrs = NearestNeighbors(n_neighbors=1).fit(X)

    rand_X = sample(range(0, n), p)

    uj = []
    wj = []
    for j in range(0, p):
        u_dist, _ = nbrs.kneighbors(uniform(np.amin(X,axis=0),np.amax(X,axis=0
),d).reshape(1, -1), 2, return_distance=True)
        uj.append(u_dist[0][1]) #distances to nearest neighbors in random data
        w_dist, _ = nbrs.kneighbors(X[rand_X[j]].reshape(1, -1), 2, return_dis
tance=True)
        wj.append(w_dist[0][1]) #distances to nearest neighbors in real data

    H = sum(uj) / (sum(uj) + sum(wj))
    if isnan(H):
        print(uj, wj)
        H = 0

    return H
hopkins(iris_data_X)
```

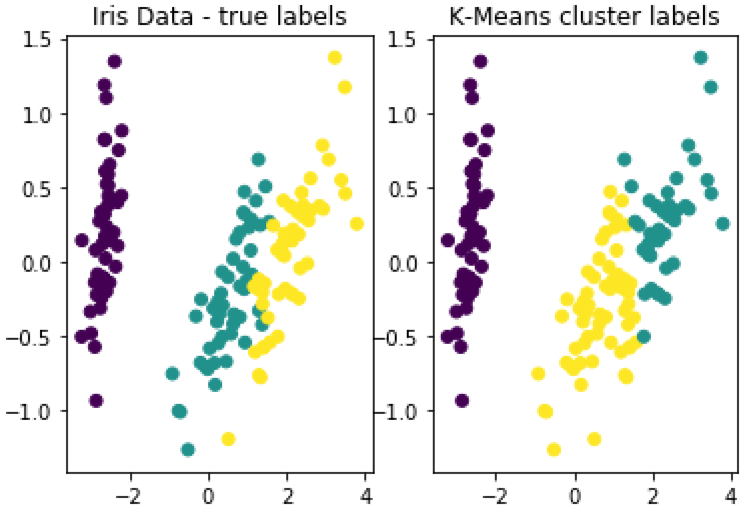
Out[40]:

```
1.0
```

This indicates that IRIS dataset is significantly a clusterable data

In [13]:

```
#K-Means Clustering
from sklearn.cluster import KMeans
n_clusters = 3
iris_X = p_data_df.values[:,0:2]
iris_Y = p_data_df.values[:,2]
kmeans = KMeans(n_clusters=n_clusters);
y_pred_iris = kmeans.fit_predict(iris_X)
fig, ax = plt.subplots()
plt.subplot(1,2,1)
plt.scatter(iris_X[:, 0], iris_X[:, 1], c=iris_Y) # true clusters
plt.title('Iris Data - true labels')
plt.subplot(1,2,2)
plt.scatter(iris_X[:, 0], iris_X[:, 1], c=y_pred_iris)  # KMeans clusters
plt.title('K-Means cluster labels')
plt.show()
```



In [37]:

```
#Rand_Index for cluster evaluation
iris_data_Y=iris_Y.astype(int)
from scipy.special import comb
def rand_index(S, T):

    Spairs = comb(np.bincount(S), 2).sum()
    Tpairs = comb(np.bincount(T), 2).sum()

    A = np.c_[(S, T)]

    f_11 = sum(comb(np.bincount(A[A[:, 0] == i, 1]), 2).sum()
             for i in set(S))

    f_10 = Spairs - f_11
    f_01 = Tpairs - f_11
    f_00 = comb(len(A), 2) - f_11 - f_10 - f_01
    return (f_00 + f_11) / (f_00 + f_01 + f_10 + f_11)
print(rand_index(iris_data_Y,y_pred_iris))
```

```
0.8737360178970918
```

Rand_Index which is approximate to 1 indicates the resultant clusters are almost similar to the original data grouped as per the Iris names.

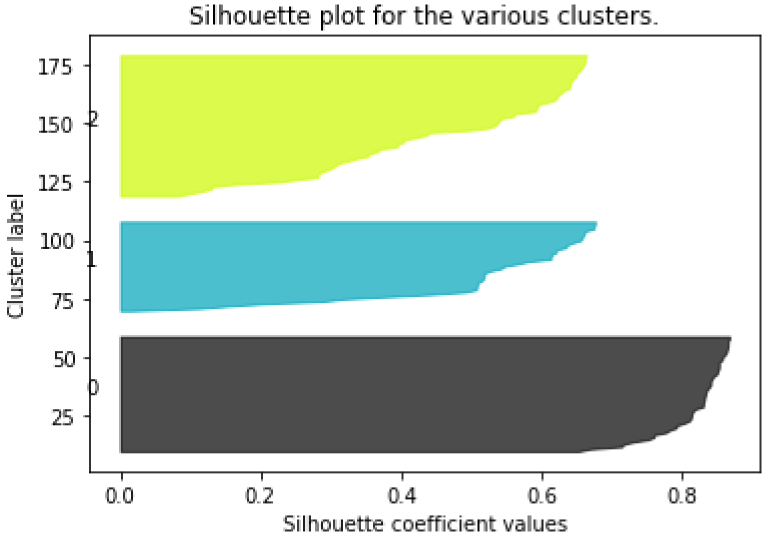
```
In [17]: #Cluster Evaluation using Silhouette Coefficient
         sample_silhouette_values = silhouette_samples(iris_X, y_pred_iris)
         np.mean(sample_silhouette_values)

Out[17]: 0.5975649100584399
```

Silhouette coefficient evaluates the cluster tendency(Checks for maximum intra cluster similarity and minimum inter cluster similarity). Here in the Iris Dataset, K-Means is successful in clustering the data but shows significantly less SC because 2 of the clusters are overlapping.

```
In [18]: #SC for Individual Clusters
         for i in range(0,n_clusters):
             print(np.mean(sample_silhouette_values[y_pred_iris==i]))

         0.8183916713573833
         0.5198366331505486
         0.46625449586926937
```

We could see that one of the clusters is having SC(0.81) which is well seperable from the other 2 clusters(Overlapping) having less SC.

In [19]:
```
from sklearn.metrics import silhouette_samples
def silhouette(X,labels):
    n_clusters = np.size(np.unique(labels));
    sample_silhouette_values = silhouette_samples(X, labels)
    y_lower = 10
    for i in range(n_clusters):
        ith_cluster_silhouette_values = sample_silhouette_values[labels == i]
        ith_cluster_silhouette_values.sort()
        size_cluster_i = ith_cluster_silhouette_values.shape[0]
        y_upper = y_lower + size_cluster_i

        color = cm.nipy_spectral(float(i) / n_clusters)
        plt.fill_betweenx(np.arange(y_lower, y_upper),
                          0, ith_cluster_silhouette_values,
                          facecolor=color, edgecolor=color, alpha=0.7)

        # Label the silhouette plots with their cluster numbers at the middle
        plt.text(-0.05, y_lower + 0.5 * size_cluster_i, str(i))
        #Compute the new y_lower for next cluster
        y_lower = y_upper + 10  # 10 for the 0 samples
    plt.title("Silhouette plot for the various clusters.")
    plt.xlabel("Silhouette coefficient values")
    plt.ylabel("Cluster label")
    plt.show()
silhouette(iris_X,y_pred_iris)
```



Question 13: Putting your findings in the context of your goal and evaluation plan, do you consider yourself successful? Provide reasons for your success or lack thereof.

Yes, I consider myself successful. I'm able to apply EDA, Classification and Clustering concepts for the IRIS dataset. But since Classification is the default task for this dataset, I checked for the clustering tendency using the evaluation measures and performed clustering.

Question 14: If you have an extra month to work on this project, what else would you do? Provide reasons.

If I was given ample time, I could have chosen a time series dataset and come up with interesting findings.

Question 15: Do you consider this project to be in the 'innovative category' or a 'good application' category? Provide your reason.

I consider this to be innovative category. For the dataset which I have picked, I performed to the maximum extent in putting together all the findings. If I wasn't restricted to perform the default task, I could have applied all my learnings from this course(except FPM).