2.3 Problem 3

Load the iris sample dataset into Python using a Pandas dataframe. Perform a PCA using the Scikit Decomposition component, and provide the percentage of variance explained by the 1st Principal Component. Use Matplotlib to plot the 1st/2nd Principal Components to recreate the scatterplot shown in class, with colored classes for each flower type.

Answer:

jupyter Sum 3 HW2 Last Checkpoint: 5 hours ago (unsaved changes)

File Edit View Insert Cell Kernel Help Python [Root]

Code CellToolbar

In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.decomposition import PCA
```

In [2]:
```
iris = pd.DataFrame(pd.read_table(r'C:\Python34\Lib\site-packages\Orange\datasets\iris.tab',sep = '\t'))
```

In [3]:
```
iris.drop('iris',axis=1,inplace=True)
```

In [4]:
```
iris1 = iris.ix[2:]
```

In [5]:
```
iris2=iris1.as_matrix()
```

In [6]:
```
pca = PCA(n_components=2)
```

In [7]:
```
pca.fit(iris2)
```

Out[7]:
```
PCA(copy=True, n_components=2, whiten=False)
```

In [8]:
```
iris3=pca.transform(iris2)
```

In [9]:
```
print(pca.explained_variance_ratio_)
```

```
[ 0.92461621  0.05301557]
```

In [10]:
```
print(pca.explained_variance_)
```

```
[ 4.19667516  0.24062861]
```

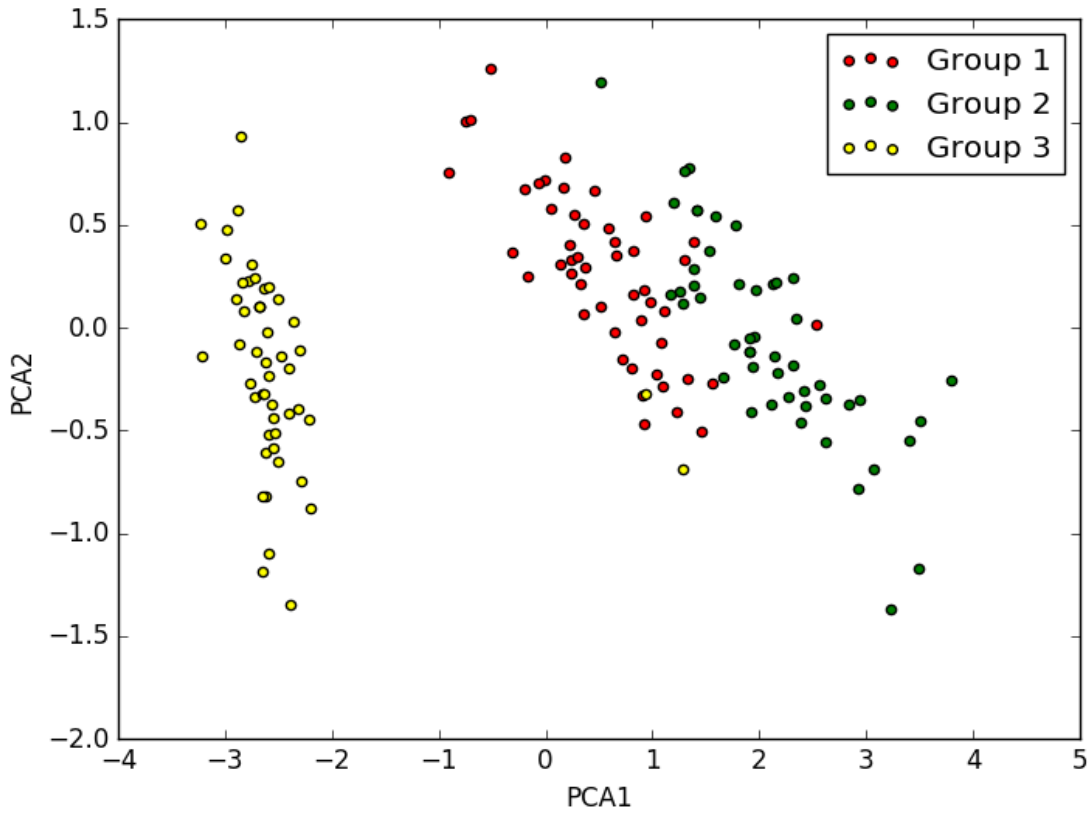
```
In [11]: iris4=pd.DataFrame(iris3)

In [12]: iris4 = iris4.ix[2:]

In [14]: irisnew = pd.DataFrame(pd.read_table(r'C:\Python34\Lib\site-packages\Orange\datasets\iris.tab',sep = '\t'))

In [15]: irisnew.drop(['sepal length','sepal width','petal length','petal width'],axis=1,inplace=True)

In [16]: irisnew1 = irisnew.ix[2:]

In [17]: irisnew2=pd.DataFrame(irisnew1)

In [18]: result = pd.concat([iris4, irisnew2], axis=1)

In [19]: result1=result.rename(columns={0:'PCA1',1:'PCA2','iris':'Iris'})

In [20]: result1
```

Out[20]:

	PCA1	PCA2	Iris
2	-2.889820	0.137346	Iris-setosa
3	-2.746437	0.311124	Iris-setosa
4	-2.728593	-0.333925	Iris-setosa
5	-2.279897	-0.747783	Iris-setosa
6	-2.820891	0.082105	Iris-setosa
7	-2.626482	-0.170405	Iris-setosa
8	-2.887959	0.570798	Iris-setosa
9	[illegible]	[illegible]	[illegible]
150	NaN	NaN	Iris-virginica
151	NaN	NaN	Iris-virginica

150 rows × 3 columns

```
In [21]: gb = result1.groupby(result1['Iris'])

In [22]: versicolor = gb.get_group('Iris-versicolor')

In [23]: virginica = gb.get_group('Iris-virginica')

In [24]: setosa = gb.get_group('Iris-setosa')

In [25]: ver = versicolor.plot.scatter(x='PCA1', y='PCA2', color='Red', label='Group 1');

In [26]: vir = virginica.plot.scatter(x='PCA1', y='PCA2', color='Green', label='Group 2', ax= ver);

In [27]: seto = setosa.plot.scatter(x='PCA1', y='PCA2', color='Yellow', label='Group 3', ax=vir);

In [28]: plt.show()

In [ ]:
```



Here, Variance as per the first PCA is 92.4%