HW_PCA_MJD

November 3, 2019

1 PCA HW - Mark Deutsch

In this HW, I use Principal Coordinate Analysis on a dataset of my own. The dataset I am using is the wine dataset that has been used previously in this class. The dataset has 13 attribute of wine, and then classifiers as to what class the wine falls in to. I go through and break down the dataset before I cluster it using PCA.

[61]:
```
import pandas as pd

df = pd.read_csv('wine.data.csv')
df.columns = ['Class', 'Alcohol', 'Malic acid', 'Ash', 'Alcalinity of ash',␣
 ↪'Magnesium', 'Total phenols', 'Flavanoids', 'Nonflavanoid phenols',␣
 ↪'Proanthocyanins', 'Color intensity', 'Hue', 'OD280/OD315 of diluted wines',␣
 ↪'Proline']
```

[62]:
```
df.head()
```

[62]:

	Class	Alcohol	Malic acid	Ash	Alcalinity of ash	Magnesium	\
0	1	13.20	1.78	2.14	11.2	100	
1	1	13.16	2.36	2.67	18.6	101	
2	1	14.37	1.95	2.50	16.8	113	
3	1	13.24	2.59	2.87	21.0	118	
4	1	14.20	1.76	2.45	15.2	112	

	Total phenols	Flavanoids	Nonflavanoid phenols	Proanthocyanins	\
0	2.65	2.76	0.26	1.28	
1	2.80	3.24	0.30	2.81	
2	3.85	3.49	0.24	2.18	
3	2.80	2.69	0.39	1.82	
4	3.27	3.39	0.34	1.97	

	Color intensity	Hue	OD280/OD315 of diluted wines	Proline
0	4.38	1.05	3.40	1050
1	5.68	1.03	3.17	1185
2	7.80	0.86	3.45	1480
3	4.32	1.04	2.93	735
4	6.75	1.05	2.85	1450

[63]:
```
df.shape
```

[63]:
```
(177, 14)
```

[64]:
```
features = list(df.columns[1:14])
```

[65]:
```
df_features = df[features]
```

[66]:
```
df_features.corr(method = 'pearson')
```

[66]:

	Alcohol	Malic acid	Ash	\
Alcohol	1.000000	0.099963	0.210964	
Malic acid	0.099963	1.000000	0.164955	
Ash	0.210964	0.164955	1.000000	
Alcalinity of ash	-0.303350	0.286148	0.446698	
Magnesium	0.258742	-0.049049	0.287107	
Total phenols	0.284543	-0.333512	0.128176	
Flavanoids	0.230133	-0.409324	0.114084	
Nonflavanoid phenols	-0.151445	0.291501	0.187354	
Proanthocyanins	0.127561	-0.217975	0.008082	
Color intensity	0.547883	0.250053	0.258643	
Hue	-0.075375	-0.560854	-0.075181	
OD280/OD315 of diluted wines	0.057417	-0.366720	0.001503	
Proline	0.641068	-0.189512	0.222979	

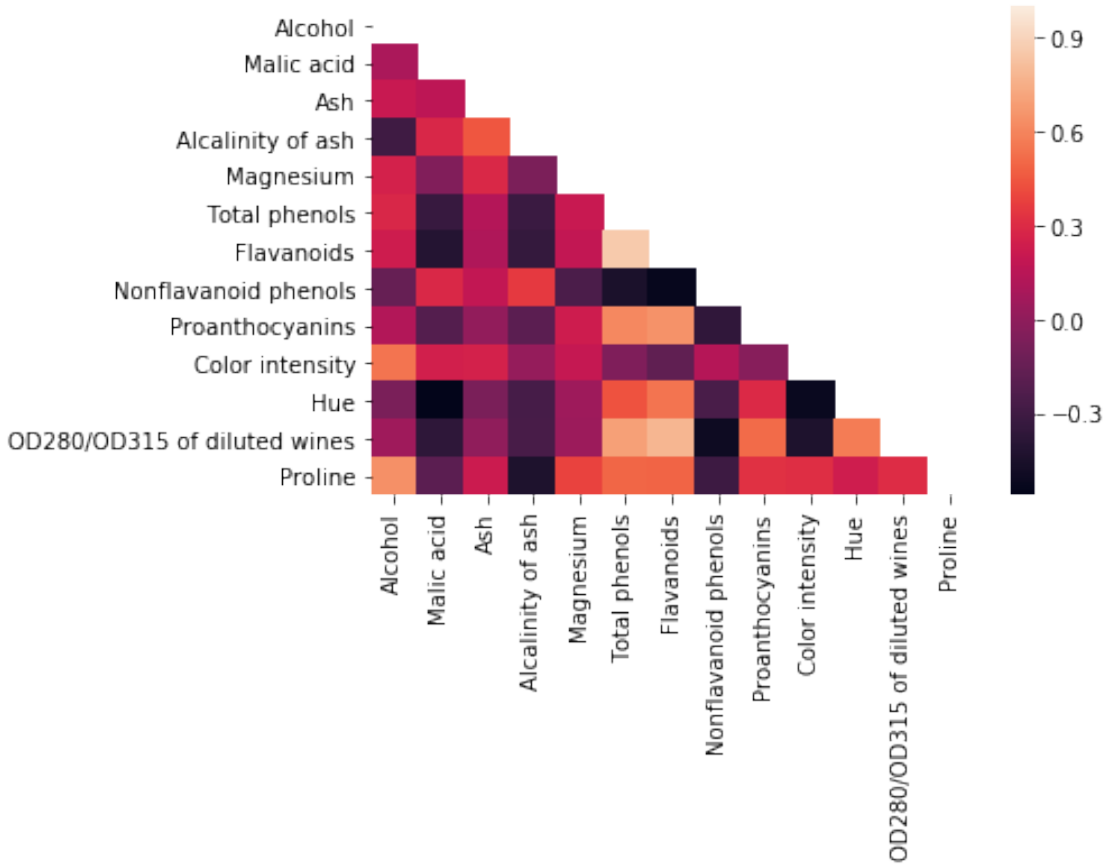
	Alcalinity of ash	Magnesium	Total phenols	\
Alcohol	-0.303350	0.258742	0.284543	
Malic acid	0.286148	-0.049049	-0.333512	
Ash	0.446698	0.287107	0.128176	
Alcalinity of ash	1.000000	-0.071707	-0.317583	
Magnesium	-0.071707	1.000000	0.208200	
Total phenols	-0.317583	0.208200	1.000000	
Flavanoids	-0.346922	0.187101	0.864046	
Nonflavanoid phenols	0.359395	-0.252091	-0.448301	
Proanthocyanins	-0.190779	0.226504	0.610533	
Color intensity	0.020478	0.199337	-0.056401	
Hue	-0.272719	0.052042	0.432987	
OD280/OD315 of diluted wines	-0.268186	0.046961	0.699566	
Proline	-0.436858	0.387542	0.495839	

	Flavanoids	Nonflavanoid phenols	\
Alcohol	0.230133	-0.151445	
Malic acid	-0.409324	0.291501	
Ash	0.114084	0.187354	
Alcalinity of ash	-0.346922	0.359395	
Magnesium	0.187101	-0.252091	
Total phenols	0.864046	-0.448301	
Flavanoids	1.000000	-0.536326	
Nonflavanoid phenols	-0.536326	1.000000	
Proanthocyanins	0.650254	-0.363268	
Color intensity	-0.174411	0.140192	

```
Hue                             0.543208             -0.261709
OD280/OD315 of diluted wines    0.786372             -0.501859
Proline                         0.491180             -0.308886

                              Proanthocyanins  Color intensity       Hue  \
Alcohol                              0.127561         0.547883 -0.075375
Malic acid                          -0.217975         0.250053 -0.560854
Ash                                  0.008082         0.258643 -0.075181
Alcalinity of ash                   -0.190779         0.020478 -0.272719
Magnesium                            0.226504         0.199337  0.052042
Total phenols                        0.610533        -0.056401  0.432987
Flavanoids                           0.650254        -0.174411  0.543208
Nonflavanoid phenols                -0.363268         0.140192 -0.261709
Proanthocyanins                      1.000000        -0.027112  0.294397
Color intensity                     -0.027112         1.000000 -0.522615
Hue                                  0.294397        -0.522615  1.000000
OD280/OD315 of diluted wines         0.513415        -0.435744  0.567395
Proline                              0.325731         0.315632  0.234879

                              OD280/OD315 of diluted wines   Proline
Alcohol                                           0.057417  0.641068
Malic acid                                       -0.366720 -0.189512
Ash                                               0.001503  0.222979
Alcalinity of ash                                -0.268186 -0.436858
Magnesium                                         0.046961  0.387542
Total phenols                                     0.699566  0.495839
Flavanoids                                        0.786372  0.491180
Nonflavanoid phenols                             -0.501859 -0.308886
Proanthocyanins                                   0.513415  0.325731
Color intensity                                  -0.435744  0.315632
Hue                                               0.567395  0.234879
OD280/OD315 of diluted wines                      1.000000  0.306031
Proline                                           0.306031  1.000000
```

[67]:
```
import seaborn as sns
import numpy as np
import matplotlib


corr = df_features.corr(method = 'pearson')
mask = np.zeros_like(corr)
mask[np.triu_indices_from(mask)] = True
sns.heatmap(corr, mask=mask)
```

[67]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x2357d659e10>
```



[68]:
```
X_data = df.iloc[:,1:14]
Y_data = df.iloc[:,0]
```

[85]:
```
X_data.head()
```

[85]:

	Alcohol	Malic acid	Ash	Alcalinity of ash	Magnesium	Total phenols	\
0	13.20	1.78	2.14	11.2	100	2.65	
1	13.16	2.36	2.67	18.6	101	2.80	
2	14.37	1.95	2.50	16.8	113	3.85	
3	13.24	2.59	2.87	21.0	118	2.80	
4	14.20	1.76	2.45	15.2	112	3.27	

	Flavanoids	Nonflavanoid phenols	Proanthocyanins	Color intensity	Hue	\
0	2.76	0.26	1.28	4.38	1.05	
1	3.24	0.30	2.81	5.68	1.03	
2	3.49	0.24	2.18	7.80	0.86	
3	2.69	0.39	1.82	4.32	1.04	
4	3.39	0.34	1.97	6.75	1.05	

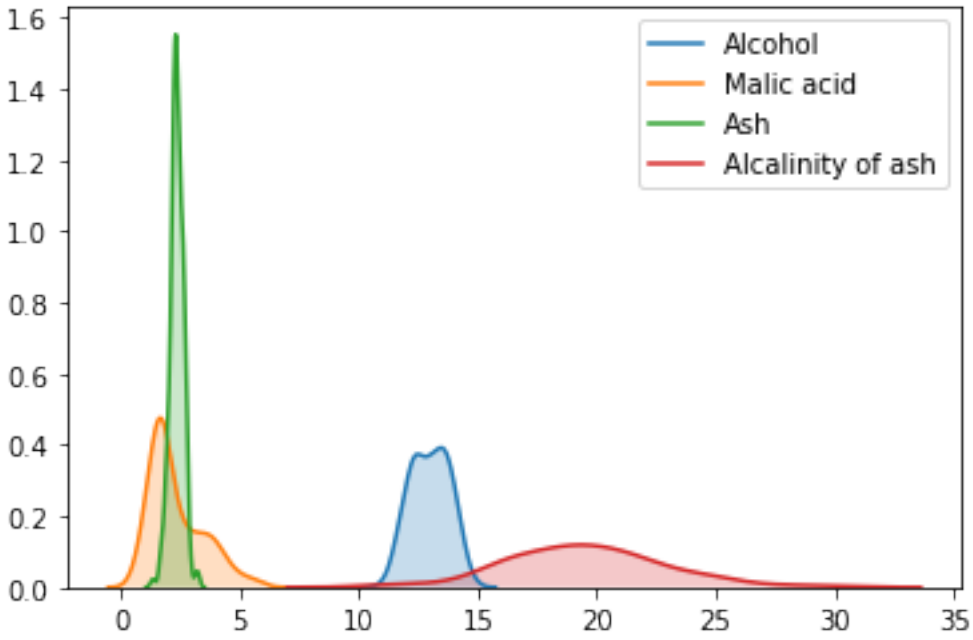
	OD280/OD315 of diluted wines	Proline
0	3.40	1050

```
1                       3.17      1185
2                       3.45      1480
3                       2.93       735
4                       2.85      1450
```

[70]:
```
from sklearn.preprocessing import StandardScaler

scaled_data = StandardScaler()
scaled_X = scaled_data.fit_transform(X_data)
```

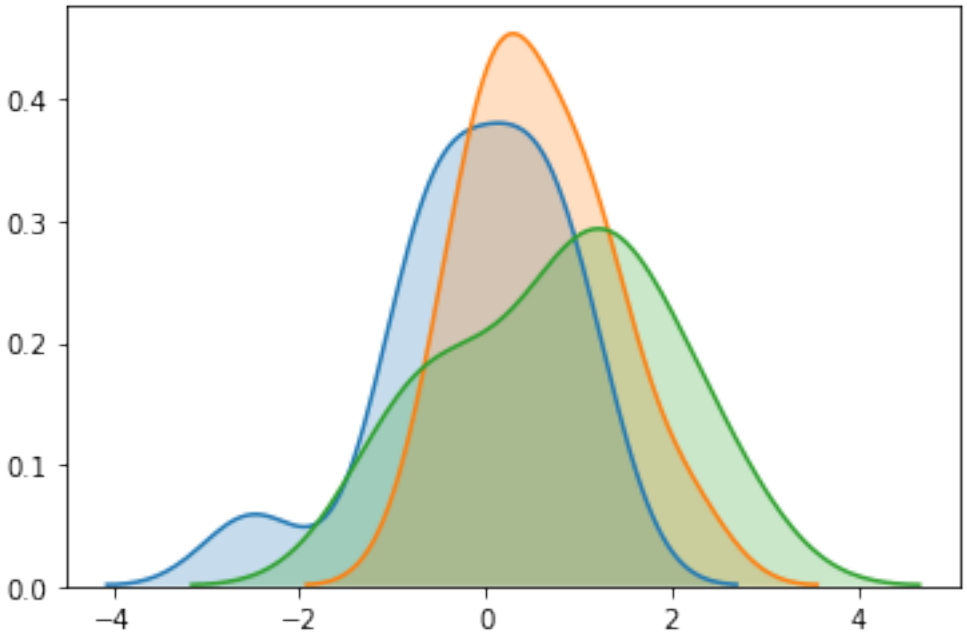
[71]:
```
sns.kdeplot(X_data.iloc[:,0], shade = True)
sns.kdeplot(X_data.iloc[:,1], shade = True)
sns.kdeplot(X_data.iloc[:,2], shade = True)
sns.kdeplot(X_data.iloc[:,3], shade = True)
```

[71]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x2357d909320>
```



[72]:
```
sns.kdeplot(scaled_X[0], shade = True)
sns.kdeplot(scaled_X[1], shade = True)
sns.kdeplot(scaled_X[2], shade = True)
```

[72]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x2357d92b9b0>
```



[86]:
```
from sklearn.decomposition import PCA
pca1 = PCA(n_components = 4)
pca1.fit(scaled_X)
trained_pca1 = pca1.transform(scaled_X)
```

[87]:
```
trained_pca1.shape
```

[87]:
```
(177, 4)
```

[88]:
```
pc_df = pd.DataFrame(data = trained_pca1, columns = ['PC1', 'PC2', 'PC3',␣
↪'PC4'])
```

[89]:
```
pc_df['Cluster'] = Y_data
```

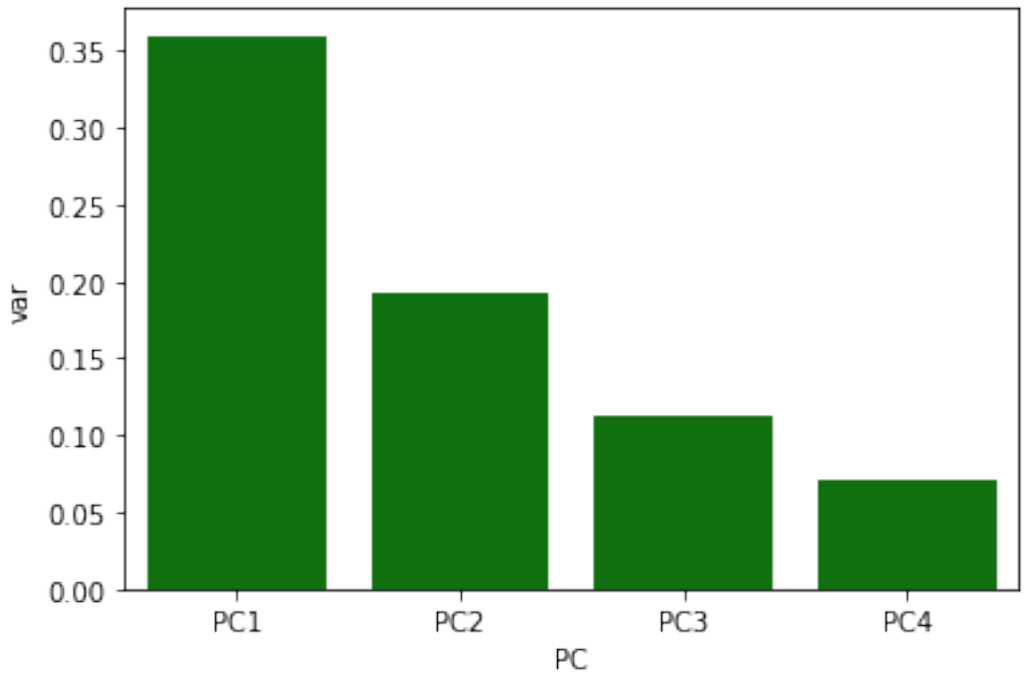
[90]:
```
pca1.explained_variance_ratio_
```

[90]:
```
array([0.35983071, 0.1924128 , 0.1117946 , 0.07111109])
```

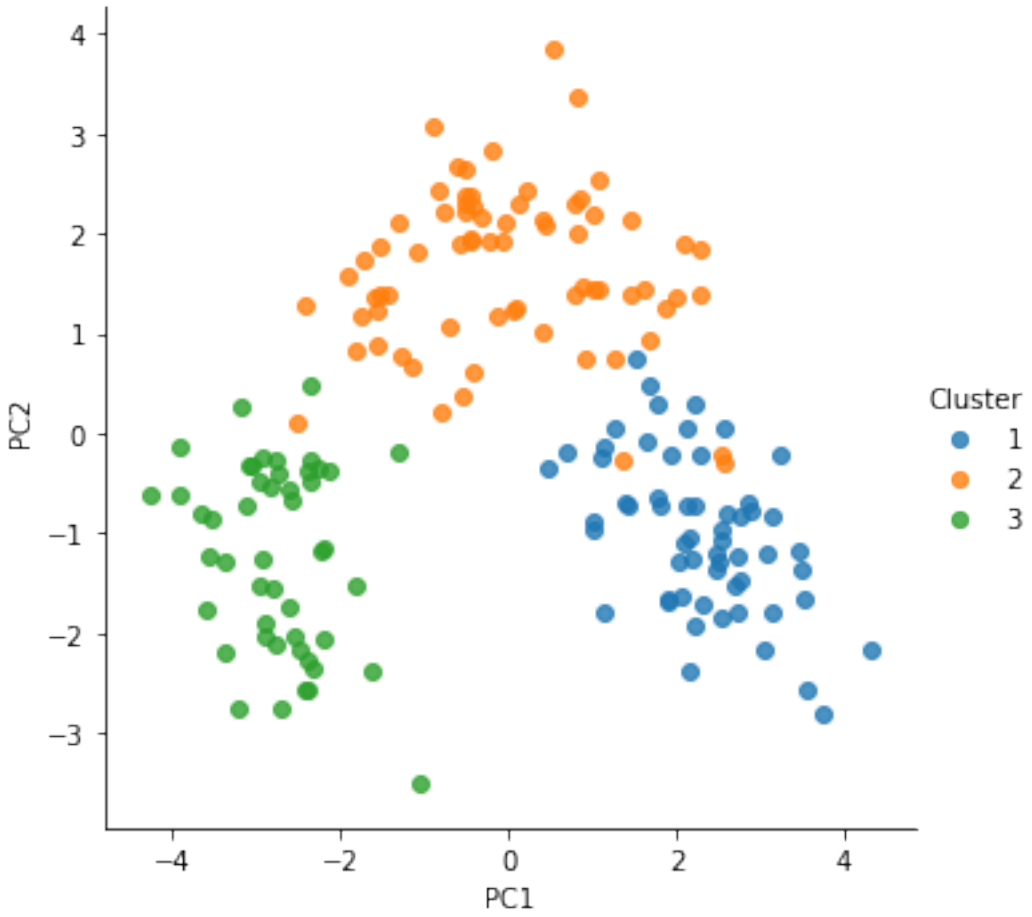
[91]:
```
df1 = pd.DataFrame({'var':pca1.explained_variance_ratio_, 'PC':['PC1', 'PC2',␣
↪'PC3', 'PC4']})
```

[92]:
```
sns.barplot(x='PC', y ='var', data=df1, color='green')
```

[92]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x2357db31b00>
```



[93]:
```
p = sns.lmplot(x='PC1', y ='PC2', data=pc_df,hue='Cluster',fit_reg=False,␣
↪legend=True)
```



2 Discussion

As seen in the above table, the PCA was able to cluster the wine data into the respective classes. It can be seen from the explained variance ratio, that PCA1 accounts for 35.9% of the variance, while PCA2 accounts for 19.2% of the variance.