# lab_0

March 8, 2021

## 1 Adam Jochna

[3]:
```
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.preprocessing import StandardScaler
import seaborn as sns
from sklearn.decomposition import PCA
import time
```

[4]:
```
PROJECT_PATH = '/home/adam/Desktop/ml_labs'
```

[5]:
```
datasets_data = {
    'iris': {
        'columns': [
            'sepal_length',
            'sepal_width',
            'petal_length',
            'petal_width',
            'class'
        ],
        'y_column': 'class',
        'class_mapping': {
            0: 'Iris-setosa',
            1: 'Iris-versicolor',
            2: 'Iris-virginica'
        },
        'plot_0_cols': ['sepal_width', 'sepal_length']
    },
    'glass': {
        'columns': [
            'Id',
            'RI',
            'Na',
            'Mg',
            'Al',
            'Si',
            'K',
```

```
            'Ca',
            'Ba',
            'Fe',
            'type_of_glass'
        ],
        'y_column': 'type_of_glass',
        'class_mapping': None,
        'class_mapping': {
            1: 'building_windows_float_processed',
            2: 'building_windows_non_float_processed',
            3: 'vehicle_windows_float_processed',
            4: 'vehicle_windows_non_float_processed (none in this database)',
            5: 'containers',
            6: 'tableware',
            7: 'headlamps',
        },
        'plot_0_cols': ['RI', 'Si']
    },
    'wine': {
        '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'
        ],
        'y_column': 'class',
        'class_mapping': {
            1: 'class_1',
            2: 'class_2',
            3: 'class_3',
        },
        'plot_0_cols': ['alcohol', 'malic_acid']
    },
}
```

[6]:
```
def perform_analysis(dataset_name):
    assert dataset_name in ['iris', 'glass', 'wine']
```

```
    df = pd.read_csv(
        '{}/lab_0/datasets/{}/{}.data'.format(
            PROJECT_PATH,
            dataset_name,
            dataset_name
        ),
        header=None
    )

    df.columns = datasets_data[dataset_name]['columns']
    y_col = datasets_data[dataset_name]['y_column']

    x_cols = datasets_data[dataset_name]['columns'].copy()
    x_cols.remove(datasets_data[dataset_name]['y_column'])

    if 'Id' in x_cols:
        x_cols.remove('Id')

    if dataset_name == 'iris':
        class_inv_mapping = {v: k for k, v in␣
↪datasets_data[dataset_name]['class_mapping'].items()}
        df[y_col] = df[y_col].apply(lambda x: class_inv_mapping[x])

    df = df[x_cols + [y_col]]
    df.columns = x_cols + ['class_idx']

    class_counts = df['class_idx'].value_counts().to_frame()
    class_counts = class_counts.reset_index()
    class_counts.columns = ['class_idx', 'class_count']
    class_counts['class_name'] = class_counts['class_idx'].apply(lambda x:␣
↪datasets_data[dataset_name]['class_mapping'][x])
    class_counts = class_counts[['class_name', 'class_idx', 'class_count']]
    class_counts['class_perc'] = class_counts['class_count']/
↪class_counts['class_count'].sum()*100

    print('#'*30)
    print('DATASET NAME: {}'.format(dataset_name))

    print()
    print('CLASS DISTRIBUTION ANALYSIS:')

    print(class_counts)

    print()
    print('DATASET ATRIBUTES:')

    print(x_cols)
```

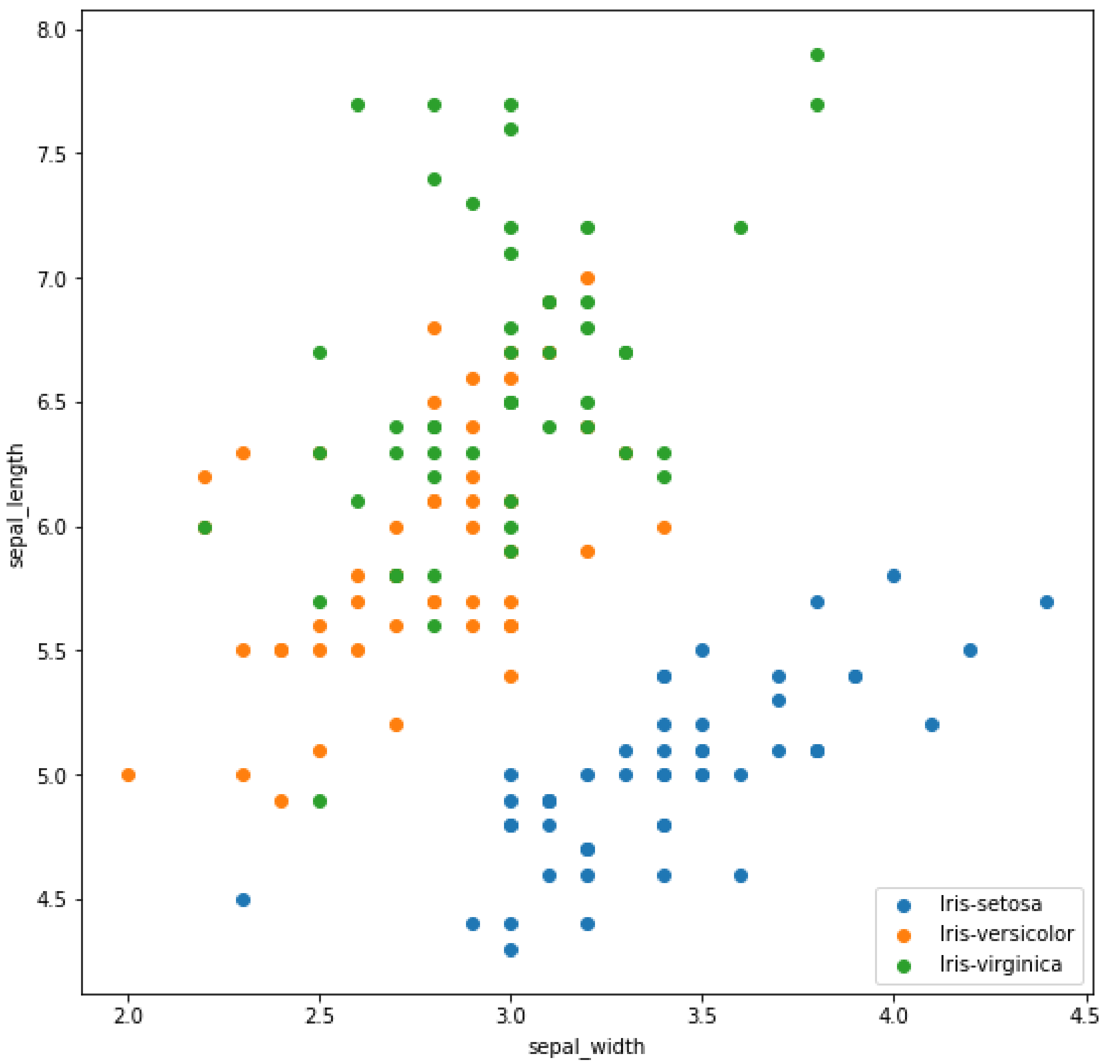
```python
print()
print('DATASET ANALYSIS:')

print(df[x_cols].describe())

# SCATTER PLOT

plt.figure(figsize=(9, 9))

for class_idx in datasets_data[dataset_name]['class_mapping'].keys():
    df_class = df.loc[df['class_idx'] == class_idx]
    class_name = datasets_data[dataset_name]['class_mapping'][class_idx]
    col_0, col1 = datasets_data[dataset_name]['plot_0_cols']

    plt.scatter(df_class[col_0], df_class[col1], label=class_name)

plt.legend(loc='lower right')
plt.xlabel(col_0)
plt.ylabel(col1)

# PAIRGRID PLOT

df['class_name'] = df['class_idx'].apply(lambda x:␣
↪datasets_data[dataset_name]['class_mapping'][x])
g = sns.PairGrid(df[x_cols + ['class_name']], hue="class_name")
g.map_diag(plt.hist)
g.map_offdiag(plt.scatter)
g.add_legend()

# PCA SCATTER PLOT

x = df.loc[:, x_cols].values
y = df.loc[:,['class_idx']].values
x = StandardScaler().fit_transform(x)

pca = PCA(n_components=2)
pca_components = pca.fit_transform(x)
df_pca = pd.DataFrame(
    data=pca_components,
    columns=['comp_0', 'comp_1']
)
df_pca = pd.concat([df_pca, df[['class_idx']]], axis=1)

plt.figure(figsize=(9, 9))

for class_idx in datasets_data[dataset_name]['class_mapping'].keys():
```

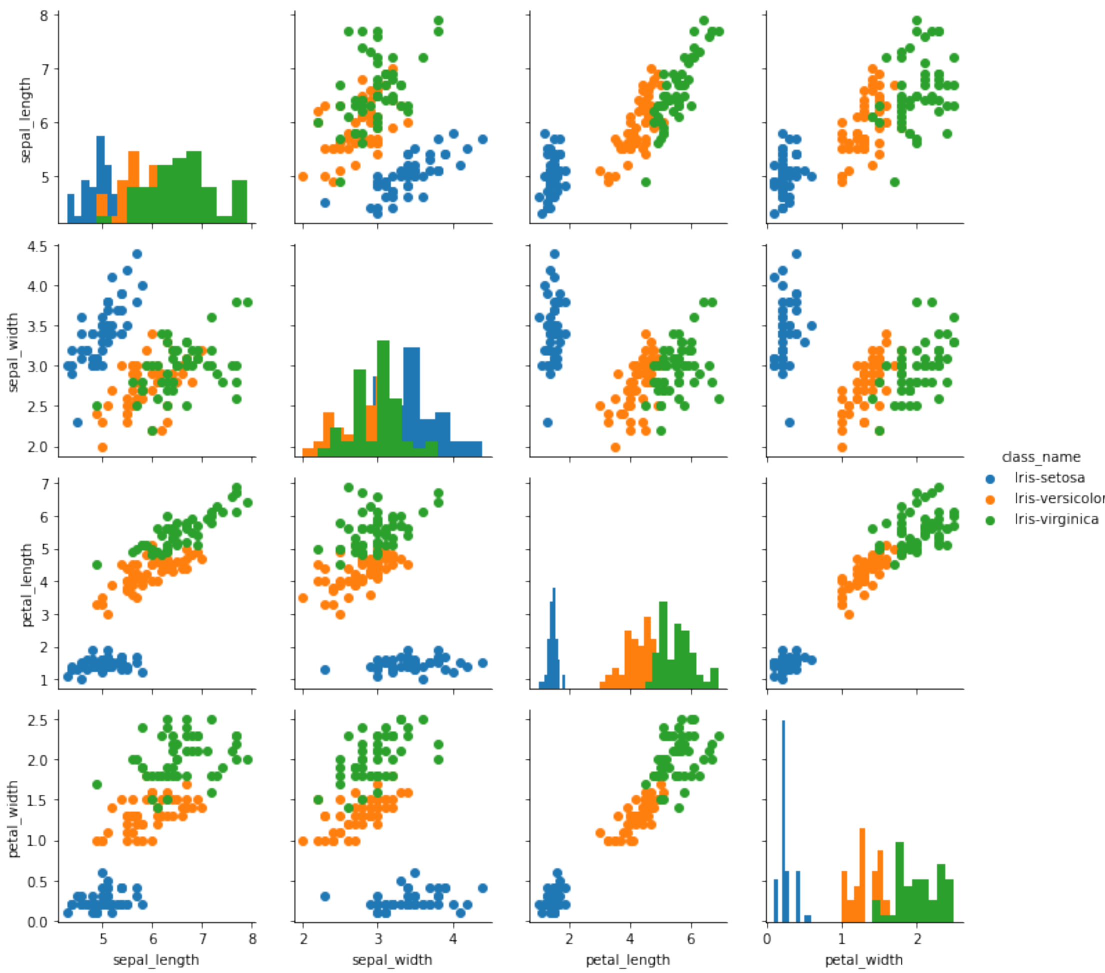
```python
        df_class = df_pca.loc[df_pca['class_idx'] == class_idx]
        class_name = datasets_data[dataset_name]['class_mapping'][class_idx]
        col_0, col1 = 'comp_0', 'comp_1'

        plt.scatter(df_class[col_0], df_class[col1], label=class_name)

    plt.legend(loc='lower right')
    plt.xlabel(col_0)
    plt.ylabel(col1)
```

[7]:

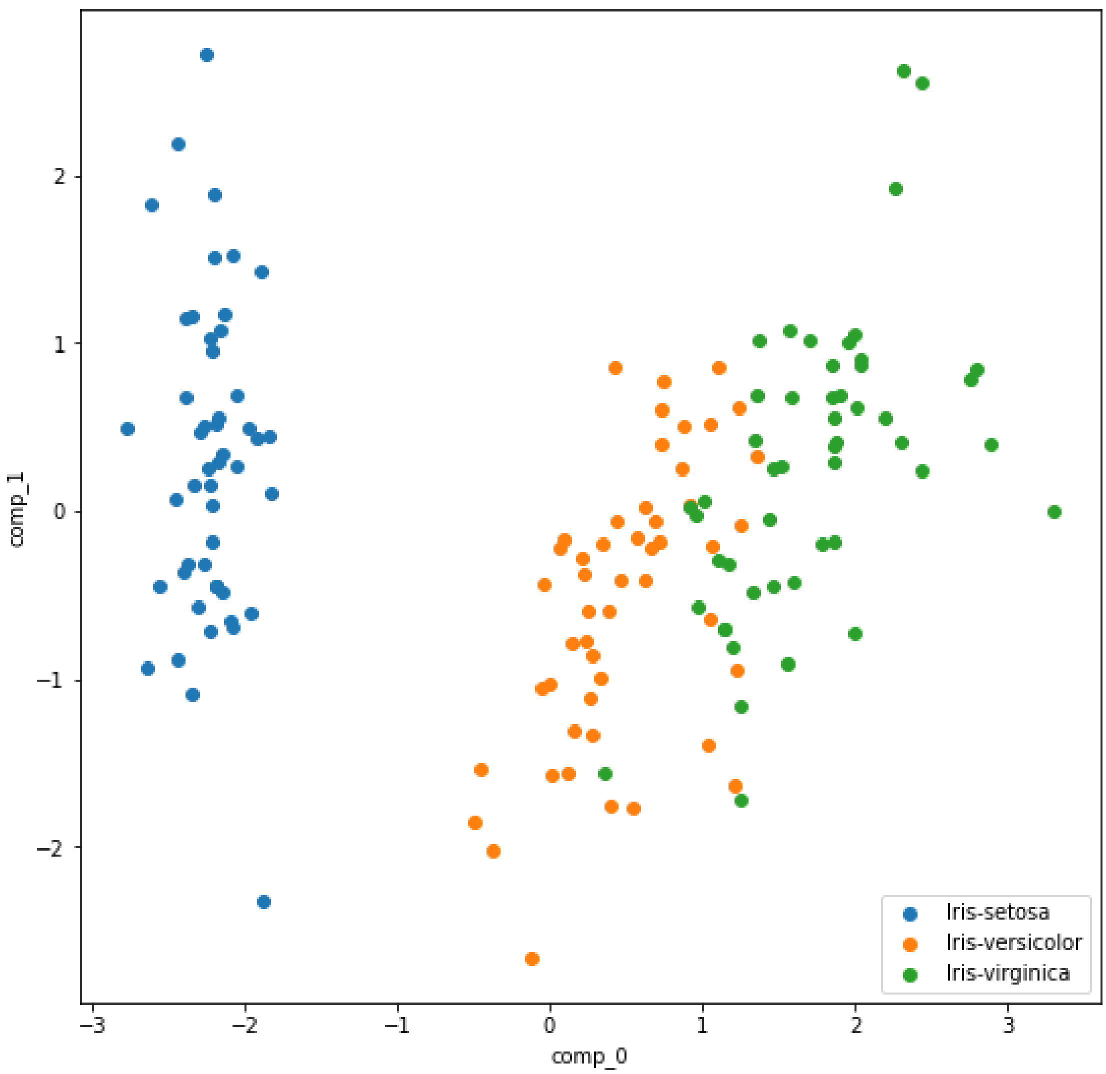
```python
perform_analysis(dataset_name='iris')
```

```
##############################
DATASET NAME: iris

CLASS DISTRIBUTION ANALYSIS:
        class_name  class_idx  class_count  class_perc
0   Iris-virginica          2           50   33.333333
1  Iris-versicolor          1           50   33.333333
2      Iris-setosa          0           50   33.333333

DATASET ATRIBUTES:
['sepal_length', 'sepal_width', 'petal_length', 'petal_width']

DATASET ANALYSIS:
       sepal_length  sepal_width  petal_length  petal_width
count    150.000000   150.000000    150.000000   150.000000
mean       5.843333     3.054000      3.758667     1.198667
std        0.828066     0.433594      1.764420     0.763161
min        4.300000     2.000000      1.000000     0.100000
25%        5.100000     2.800000      1.600000     0.300000
50%        5.800000     3.000000      4.350000     1.300000
75%        6.400000     3.300000      5.100000     1.800000
max        7.900000     4.400000      6.900000     2.500000
```


8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
sepal_length
2.0
2.5
3.0
3.5
4.0
4.5
sepal_width
Iris-setosa
Iris-versicolor
Iris-virginica



sepal_length
sepal_width
petal_length
petal_width
class_name
Iris-setosa
Iris-versicolor
Iris-virginica




[8]:
```
perform_analysis(dataset_name='glass')
```

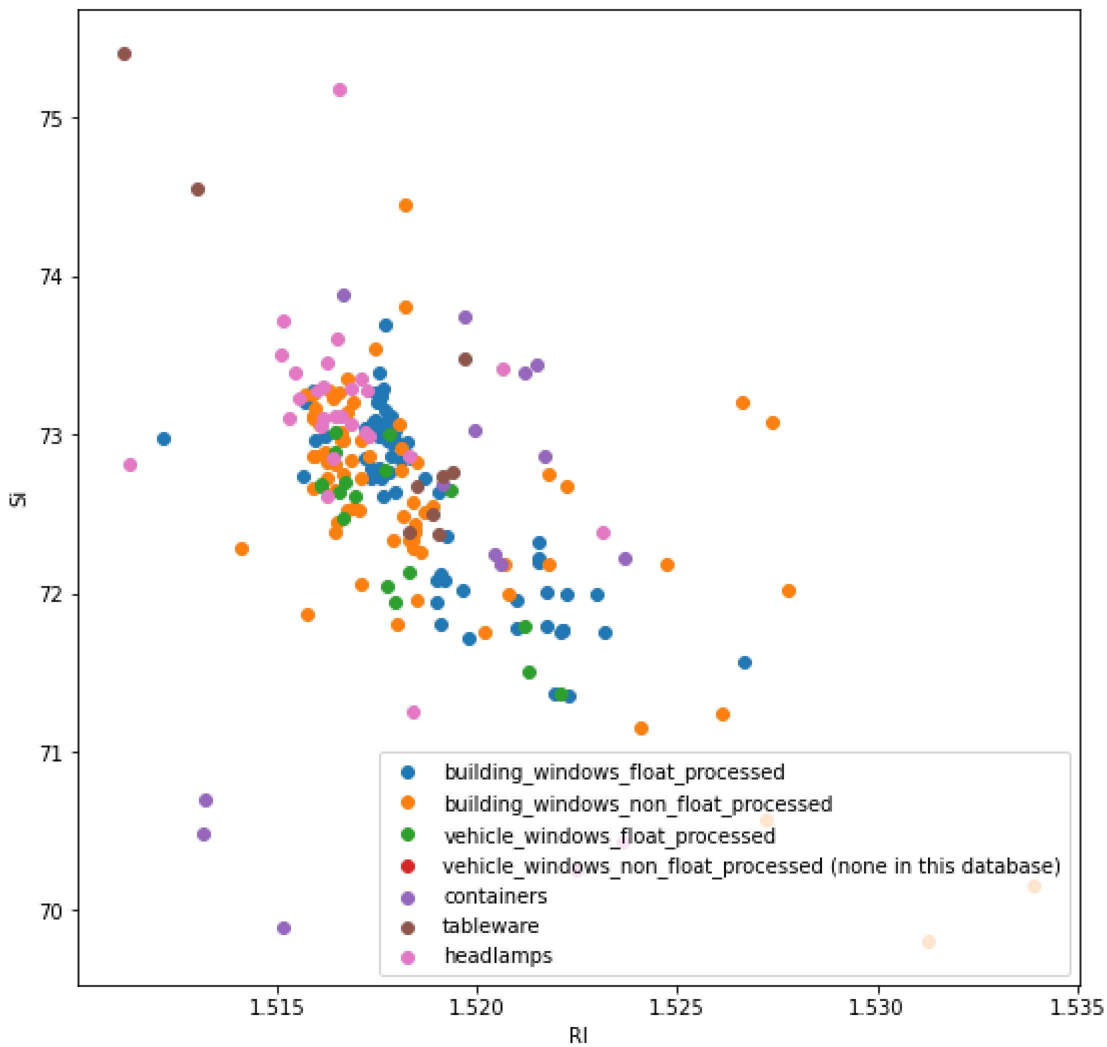
```
##############################
DATASET NAME: glass

CLASS DISTRIBUTION ANALYSIS:
                        class_name  class_idx  class_count  class_perc
0  building_windows_non_float_processed          2           76   35.514019
1      building_windows_float_processed          1           70   32.710280
2                          headlamps          7           29   13.551402
3       vehicle_windows_float_processed          3           17    7.943925
4                         containers          5           13    6.074766
5                          tableware          6            9    4.205607
```

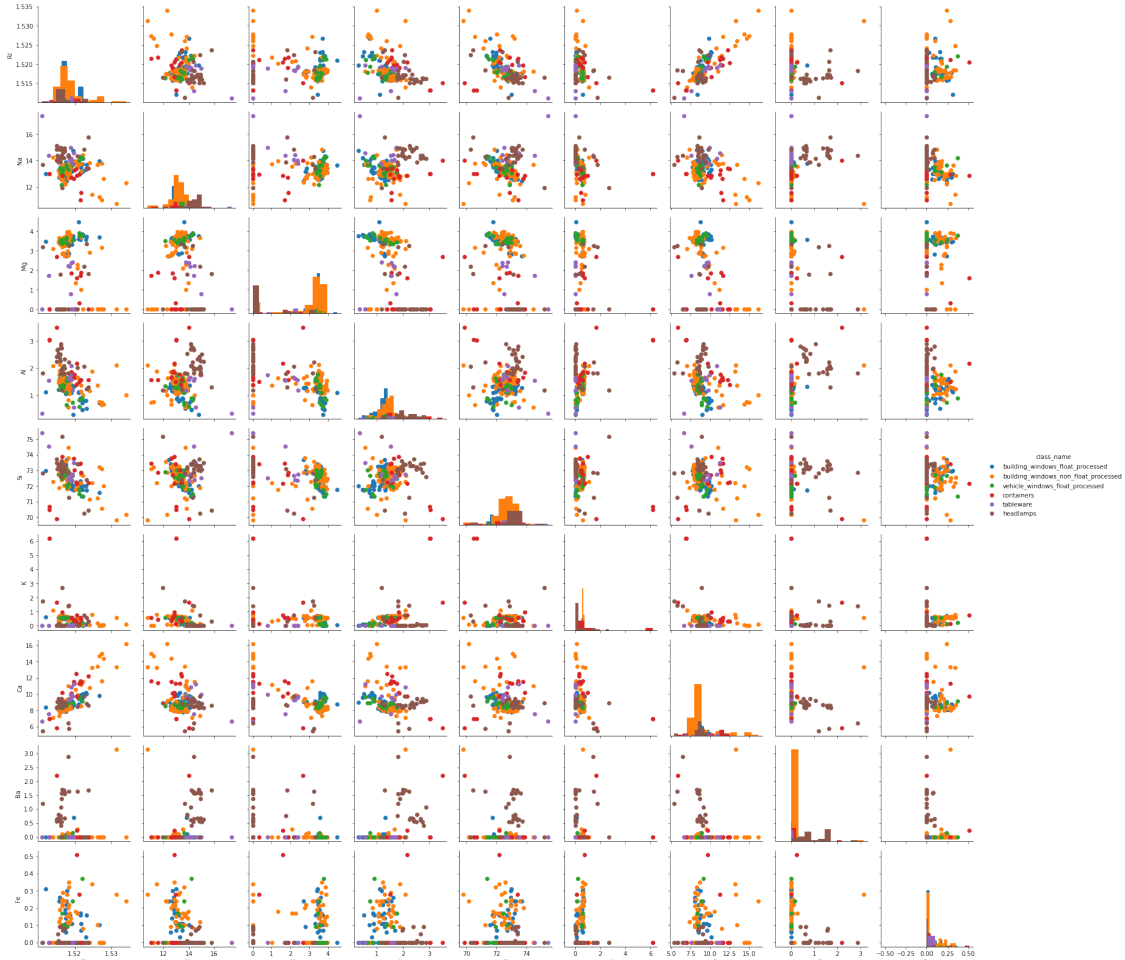
DATASET ATRIBUTES:
['RI', 'Na', 'Mg', 'Al', 'Si', 'K', 'Ca', 'Ba', 'Fe']

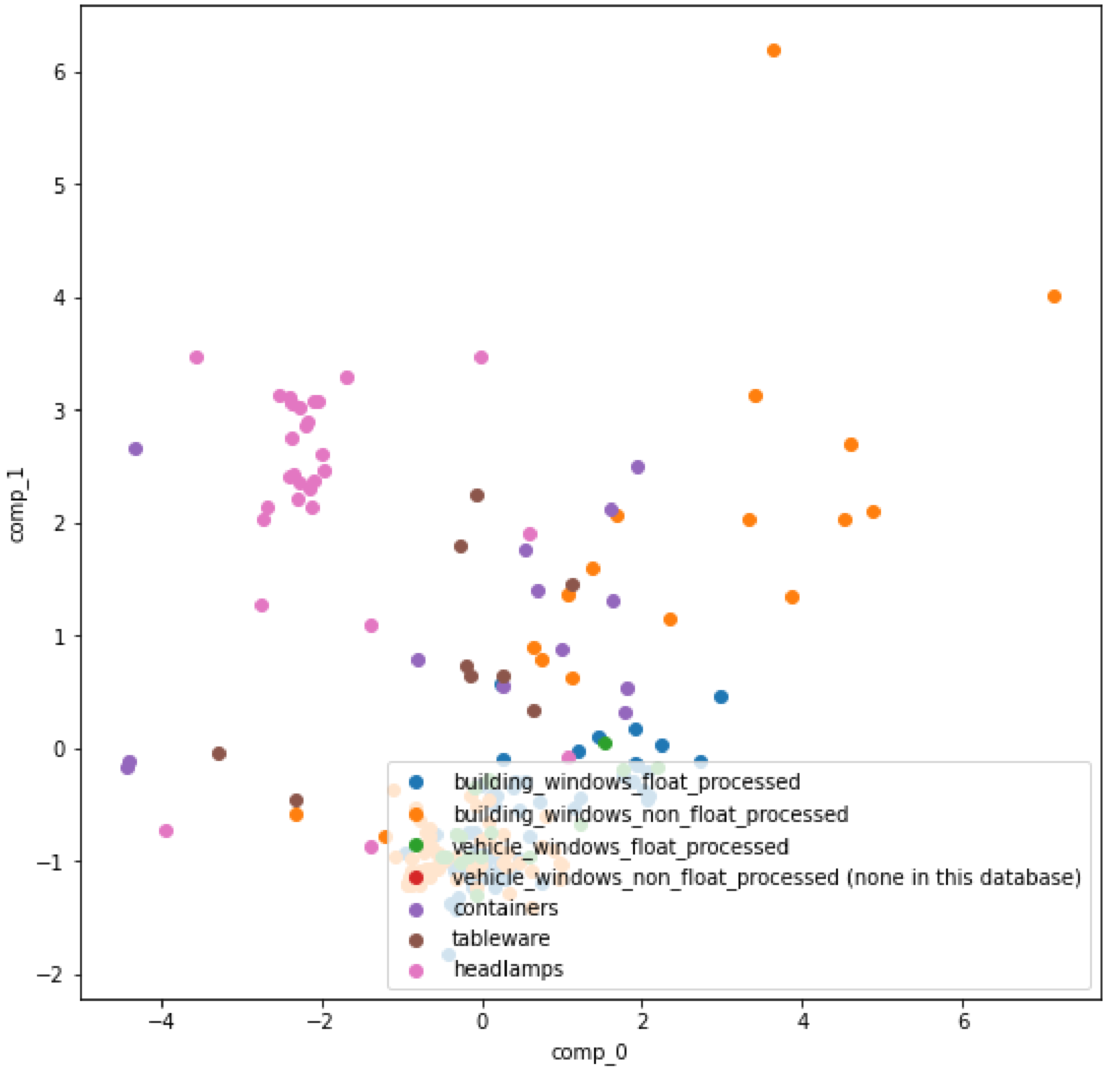
DATASET ANALYSIS:

| | RI | Na | Mg | Al | Si | K | \ |
|---|---|---|---|---|---|---|---|
| count | 214.000000 | 214.000000 | 214.000000 | 214.000000 | 214.000000 | 214.000000 | |
| mean | 1.518365 | 13.407850 | 2.684533 | 1.444907 | 72.650935 | 0.497056 | |
| std | 0.003037 | 0.816604 | 1.442408 | 0.499270 | 0.774546 | 0.652192 | |
| min | 1.511150 | 10.730000 | 0.000000 | 0.290000 | 69.810000 | 0.000000 | |
| 25% | 1.516523 | 12.907500 | 2.115000 | 1.190000 | 72.280000 | 0.122500 | |
| 50% | 1.517680 | 13.300000 | 3.480000 | 1.360000 | 72.790000 | 0.555000 | |
| 75% | 1.519157 | 13.825000 | 3.600000 | 1.630000 | 73.087500 | 0.610000 | |
| max | 1.533930 | 17.380000 | 4.490000 | 3.500000 | 75.410000 | 6.210000 | |

| | Ca | Ba | Fe |
|---|---|---|---|
| count | 214.000000 | 214.000000 | 214.000000 |
| mean | 8.956963 | 0.175047 | 0.057009 |
| std | 1.423153 | 0.497219 | 0.097439 |
| min | 5.430000 | 0.000000 | 0.000000 |
| 25% | 8.240000 | 0.000000 | 0.000000 |
| 50% | 8.600000 | 0.000000 | 0.000000 |
| 75% | 9.172500 | 0.000000 | 0.100000 |
| max | 16.190000 | 3.150000 | 0.510000 |


75
74
73
72
71
70
Si
1.515
1.520
1.525
1.530
1.535
RI
building_windows_float_processed
building_windows_non_float_processed
vehicle_windows_float_processed
vehicle_windows_non_float_processed (none in this database)
containers
tableware
headlamps

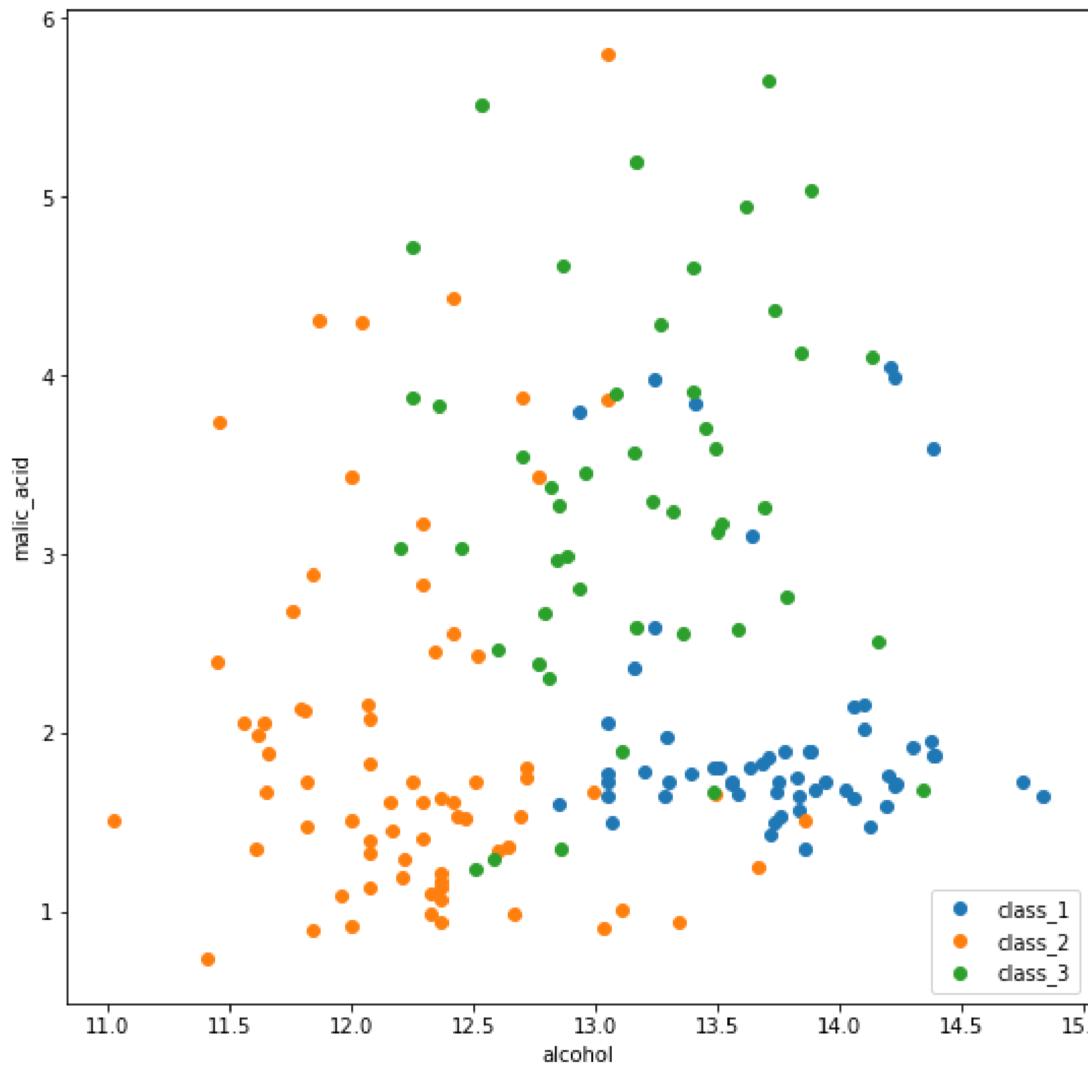

RI
Na
Mg
Al
Si
K
Ca
Ba
Fe
class_name
building_windows_float_processed
building_windows_non_float_processed
vehicle_windows_float_processed
containers
tableware
headlamps

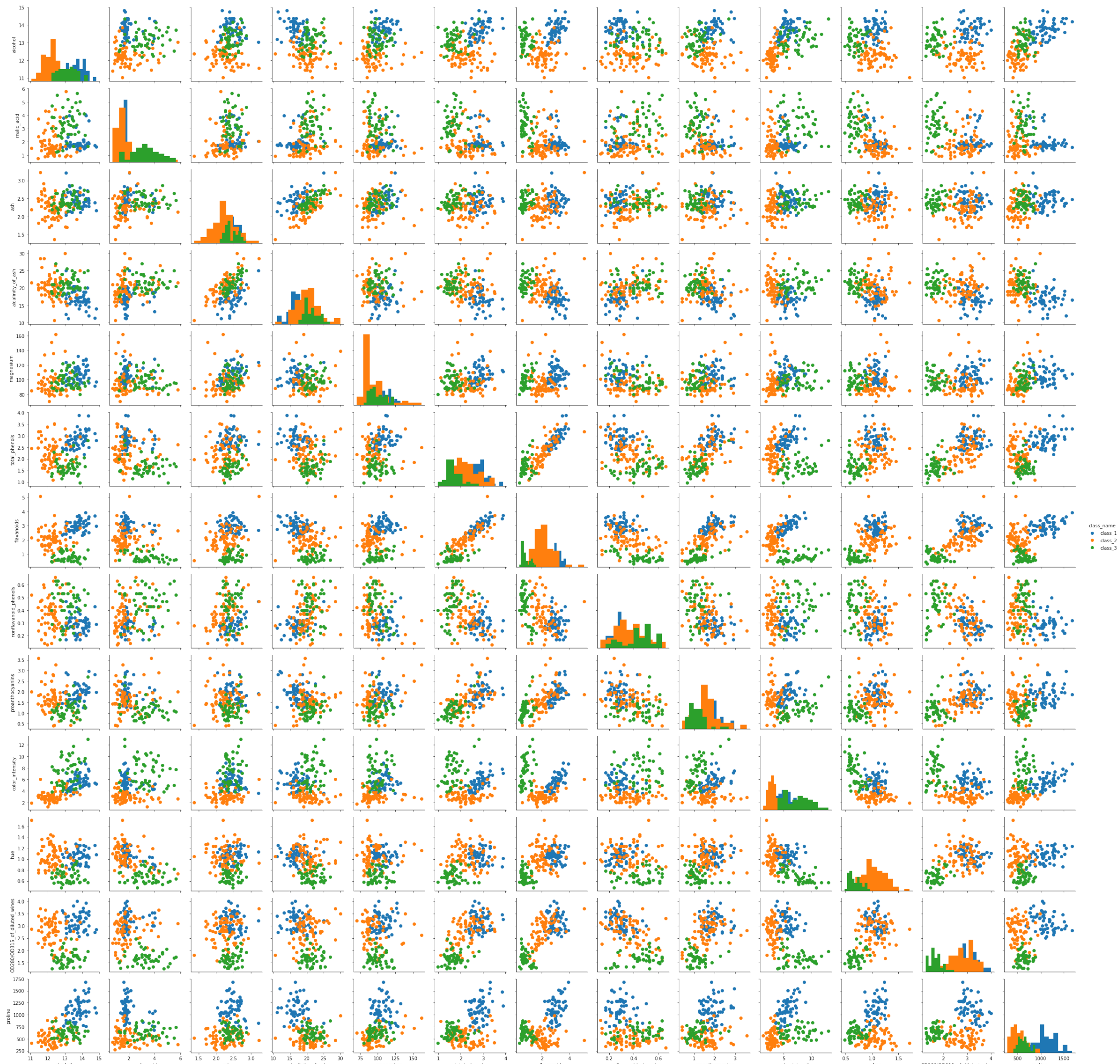


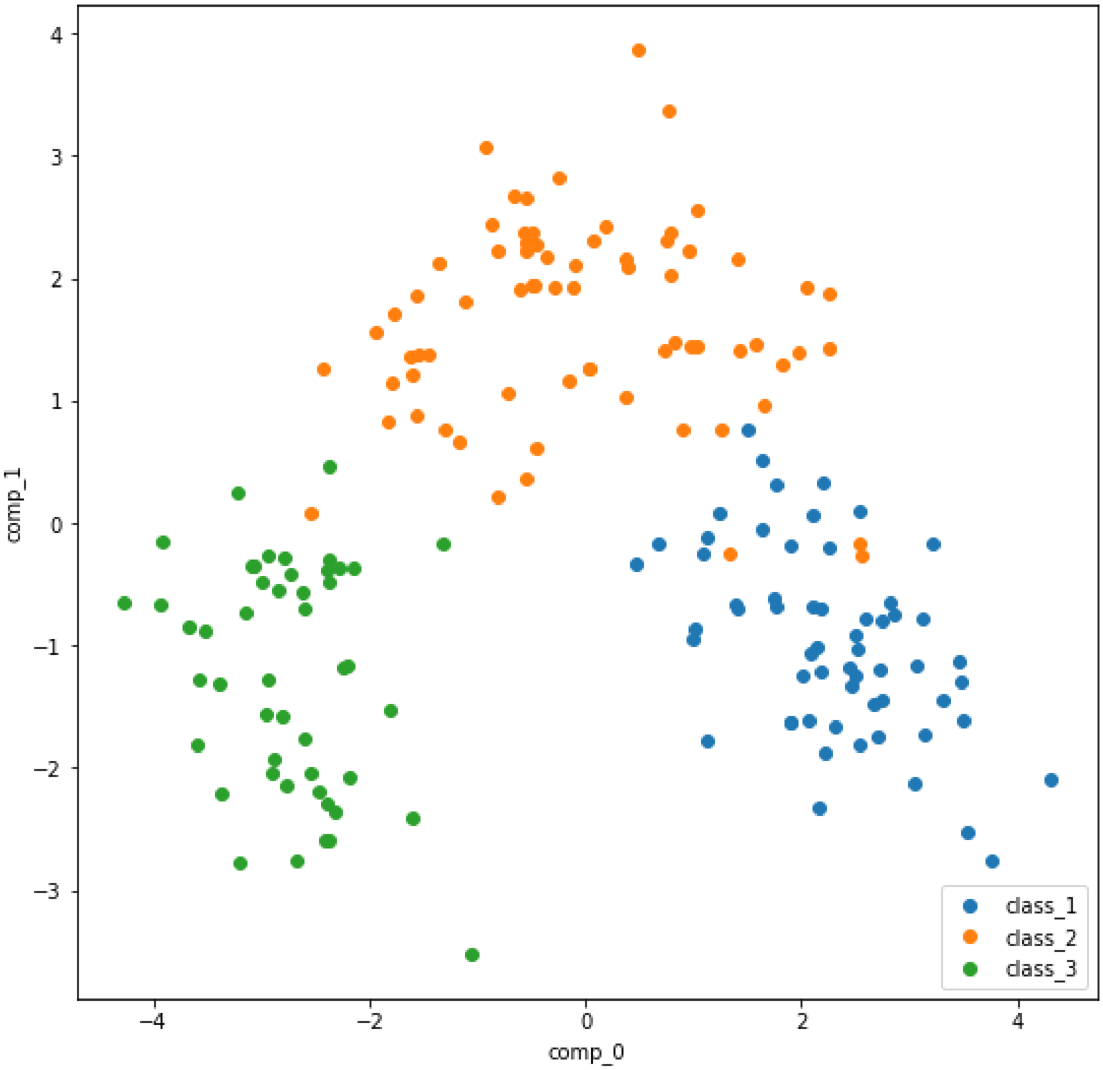

[9]:
```
perform_analysis(dataset_name='wine')
```

```
###############################
DATASET NAME: wine

CLASS DISTRIBUTION ANALYSIS:
  class_name  class_idx  class_count  class_perc
0    class_2          2           71   39.887640
1    class_1          1           59   33.146067
2    class_3          3           48   26.966292

DATASET ATRIBUTES:
['alcohol', 'malic_acid', 'ash', 'alcalinity_of_ash', 'magnesium',
'total_phenols', 'flavanoids', 'nonflavanoid_phenols', 'proanthocyanins',
```

```
'color_intensity', 'hue', 'OD280/OD315_of_diluted_wines', 'proline']

DATASET ANALYSIS:
          alcohol  malic_acid         ash  alcalinity_of_ash   magnesium  \
count  178.000000  178.000000  178.000000         178.000000  178.000000
mean    13.000618    2.336348    2.366517          19.494944   99.741573
std      0.811827    1.117146    0.274344           3.339564   14.282484
min     11.030000    0.740000    1.360000          10.600000   70.000000
25%     12.362500    1.602500    2.210000          17.200000   88.000000
50%     13.050000    1.865000    2.360000          19.500000   98.000000
75%     13.677500    3.082500    2.557500          21.500000  107.000000
max     14.830000    5.800000    3.230000          30.000000  162.000000

       total_phenols  flavanoids  nonflavanoid_phenols  proanthocyanins  \
count     178.000000  178.000000            178.000000       178.000000
mean        2.295112    2.029270              0.361854         1.590899
std         0.625851    0.998859              0.124453         0.572359
min         0.980000    0.340000              0.130000         0.410000
25%         1.742500    1.205000              0.270000         1.250000
50%         2.355000    2.135000              0.340000         1.555000
75%         2.800000    2.875000              0.437500         1.950000
max         3.880000    5.080000              0.660000         3.580000

       color_intensity         hue  OD280/OD315_of_diluted_wines      proline
count       178.000000  178.000000                    178.000000   178.000000
mean          5.058090    0.957449                      2.611685   746.893258
std           2.318286    0.228572                      0.709990   314.907474
min           1.280000    0.480000                      1.270000   278.000000
25%           3.220000    0.782500                      1.937500   500.500000
50%           4.690000    0.965000                      2.780000   673.500000
75%           6.200000    1.120000                      3.170000   985.000000
max          13.000000    1.710000                      4.000000  1680.000000
```


6
5
4
3
2
1
malic_acid
11.0
11.5
12.0
12.5
13.0
13.5
14.0
14.5
15.0
alcohol
class_1
class_2
class_3


alcohol
malic_acid
ash
alcalinity_of_ash
magnesium
total_phenols
flavanoids
nonflavanoid_phenols
proanthocyanins
color_intensity
hue
OD280/OD315_of_diluted_wines
proline
class_name
class_1
class_2
class_3



comp_1
comp_0
4
3
2
1
0
−1
−2
−3
−4
−2
0
2
4
class_1
class_2
class_3