Stony Brook University

EEO388

Fall 2024

Assignment 1

Pete Mills
ID: 115009163

Professor
Vibha Mane

September 22, 2024

Overview

In this assignment we made use of scatter plots and histograms to analyze distributions and relationships, separabiility, and irregularities in data sets.

1 Exercise Set 1A

```python
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from datetime import datetime

# Create a NumPy array with 1 column and 100 rows, filled with random numbers
array_1d = np.random.randn(100)


# Convert the 1D array into a pandas Series
series_1d = pd.Series(array_1d)
print(series_1d)


# Create a NumPy array with 3 columns and 100 rows, filled with random integers
array_3d = np.random.randn(100, 3)


# Convert the 3-column array into a DataFrame with column labels
df = pd.DataFrame(array_3d, columns=['X1', 'X2', 'X3'])
print(df)


# Use Seaborn to create a pairplot of the DataFrame
sns.pairplot(df)
#plt.show()

# Generate current timestamp in the format yyyy-mm-dd-hh-mm-ss
timestamp = datetime.now().strftime("%Y-%m-%d-%H-%M-%S")

# Create the file name with timestamp prepended
filename = f"{timestamp}_pairplot.png"
plt.savefig(filename)


```

Figure 1: Source code.

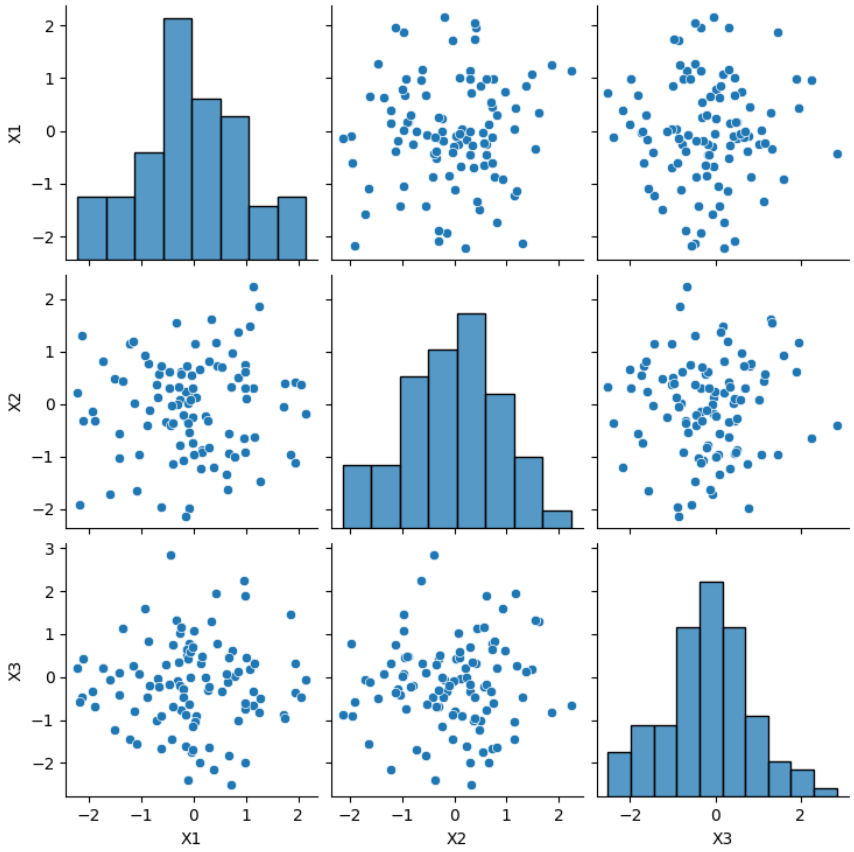

Figure 2: The Pair plot set for these random data.

2 Exercise Set 1B

2.1 Raisin

The data in this dataset are categorical. Based on the plots there do appear to be some outliers. There are 450 samples in the Kecimen class, and 450 samples in the Bensi class.

By setting hue based on class, we can visualize the two classes of raisins on the same scatter plot. We can then inspect for clustering of data which reveals there is some separability in the data. We can use this to identify characteristics for class sorting.

```python
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Load the data
df_raisin = pd.read_csv('../data-sets/Raisin_Dataset.csv')

# Scatter plot matrix
sns.pairplot(df_raisin, hue='Class', markers=["o", "s"])
plt.show()

# Histograms
df_raisin.hist(bins=30, figsize=(10, 7))
plt.tight_layout()
plt.show()

# Check for class distribution
print(df_raisin['Class'].value_counts())

```

Figure 3: Source code.

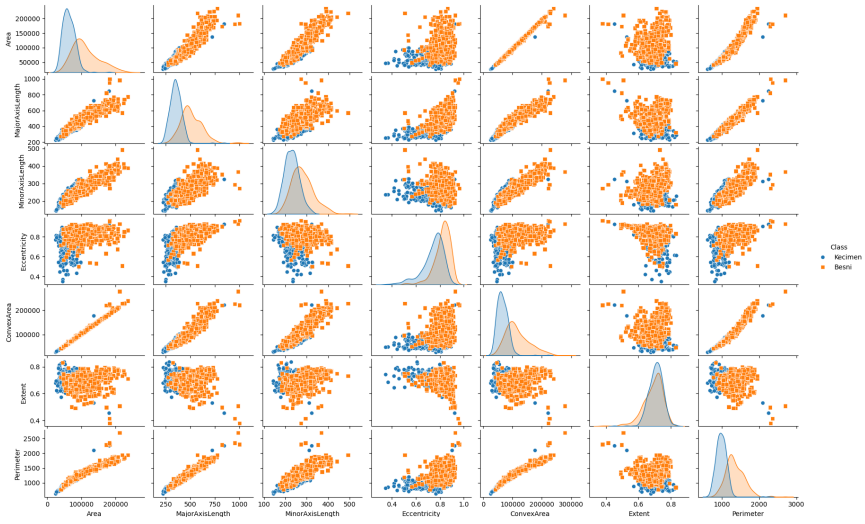

Figure 4: Scatter plots.

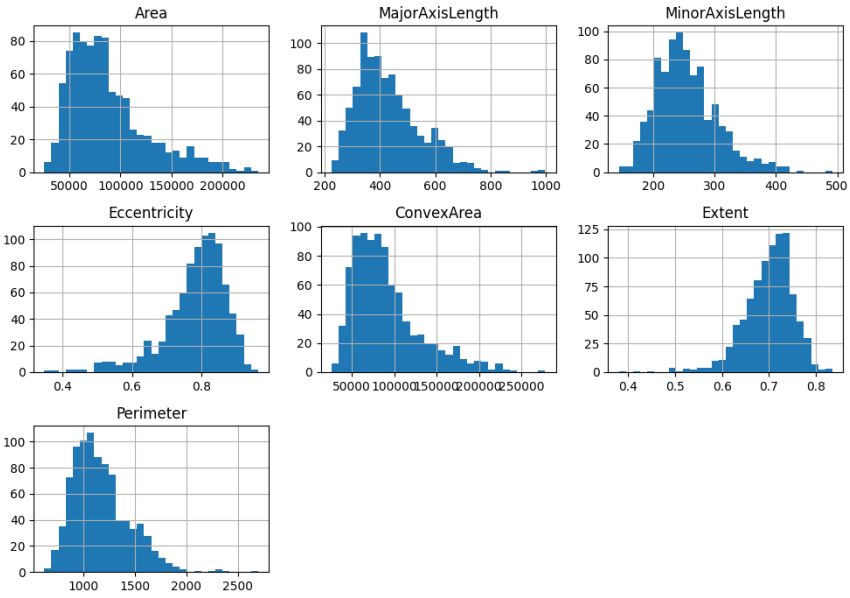

Figure 5: Histograms.

2.2 Deep Space

The data in this dataset are numerical. There does not appear to be any outliers.

```python
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Load the data
df_deepspace = pd.read_csv('../data-sets/DeepSpaceData.csv')

# Scatter plot matrix without class labels
sns.pairplot(df_deepspace)
plt.show()


# Histograms
df_deepspace.hist(bins=30, figsize=(10, 7))
plt.tight_layout()
plt.show()
```

Figure 6: Source code.

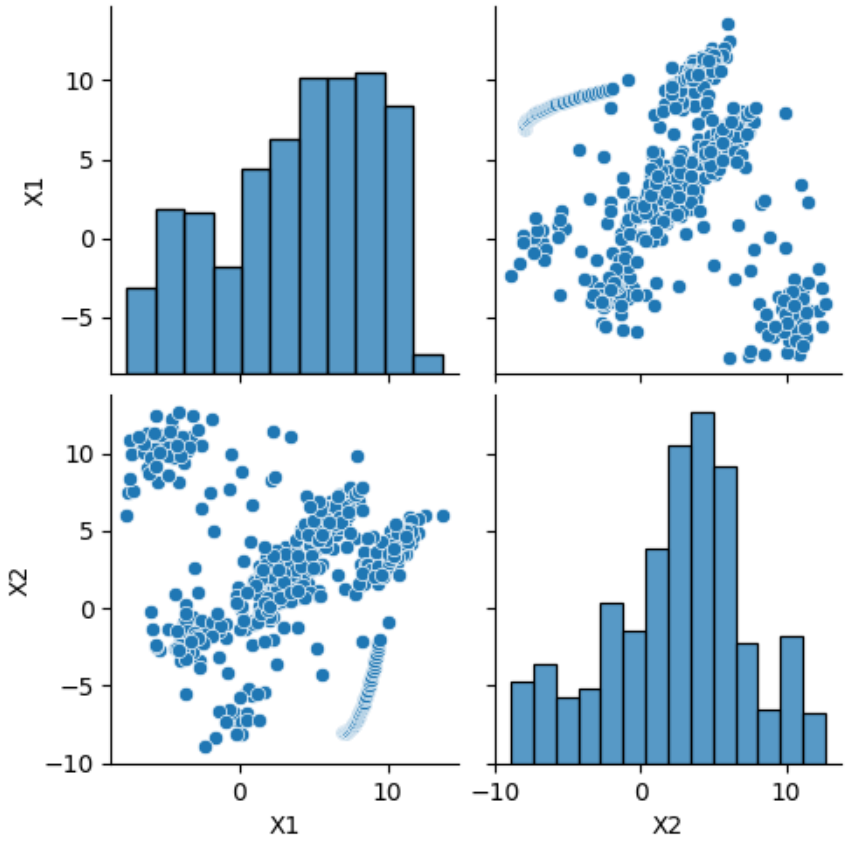

Figure 7: Scatter Plots.

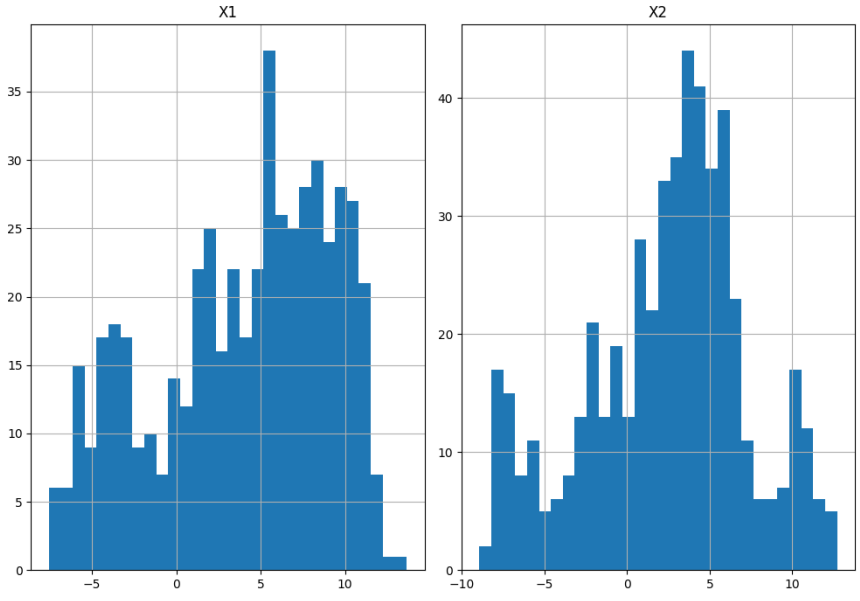

Figure 8: Histograms.