Plant data analysis

By Kirill Markin

Introduction

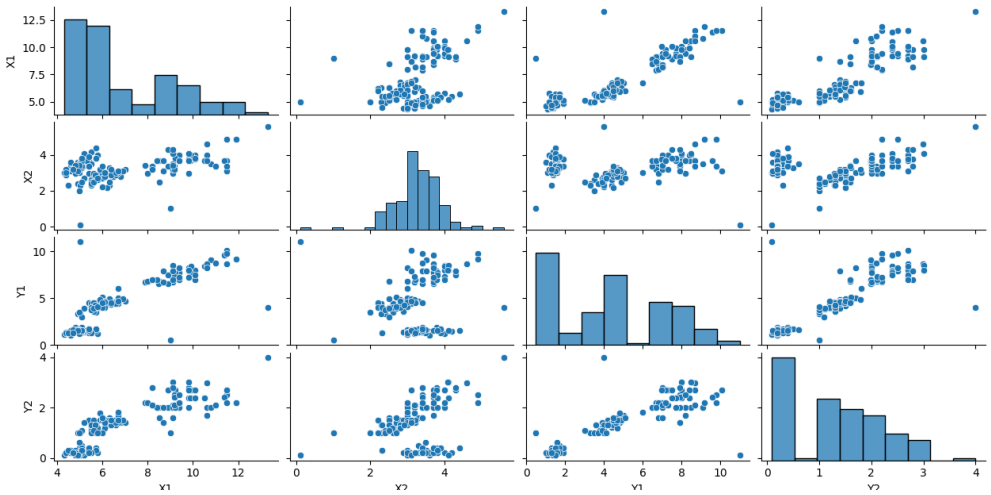
The focus of this analysis lies in understanding the essence of size measurement data related to a specific plant species. Through an initial analysis, various statistical methods such as scatter plots and correlation calculations were employed to discern fundamental trends and connections within the dataset. Notable patterns emerged from this preliminary examination. To deepen the understanding of these observations, Principal Component Analysis (PCA) was used, allowing for a more thorough interpretation of the dataset's underlying patterns. In addition, K-means algorithm was used to identify distinct groupings within the data, offering potential insights into the plant's size variations. These analytical approaches collectively contribute to a more comprehensive understanding of the dataset.

General analysis

Exploratory data analysis was conducted to gain general knowledge about dataset and its parameters. This included the generation of a matrix scatter plot, box plots to identify the distribution and outliers of individual variables, and a correlation matrix. These preliminary steps provided general overview of the dataset's structure and guided further analysis.

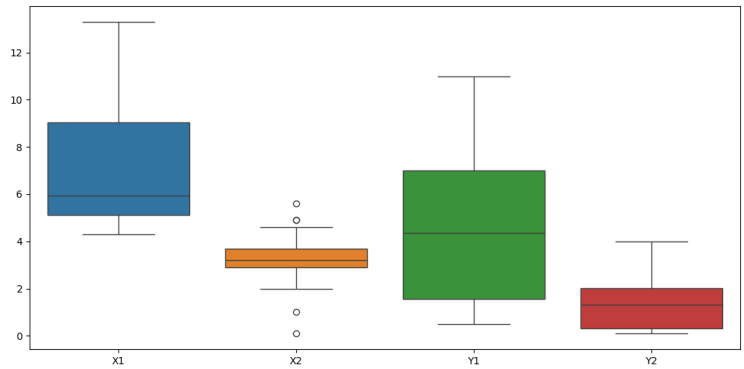
The matrix scatter plot used to visually demonstrate complex relationships within the multidimensional dataset (figure 1). By visualizing pairwise associations between variables, distinct clusters and potential trends became visible, providing valuable insights into the underlying structure of the data. Notably, while some clusters exhibited discernible patterns aligned with diagonal trends or linear relationships, others appeared more scattered, suggesting non-linear associations or complex dependencies. Moreover, the presence of outliers across multiple plots underscores the importance of robust analysis techniques to account for potential anomalies. These initial observations lay the groundwork for further exploration.

Figure 1. Matrix Scatter plot



Box plots provide a visual summary of the distribution and variability within each variable (figure 2). The varying characteristics of the box plots across different variables offered insights into their respective distributions and potential outliers. X1 exhibited a wider interquartile range and higher median, suggesting greater variability compared to other variables. On the other hand, Y2 showcased a narrower interquartile range, indicating less dispersion in its data. The presence of outliers in X2 highlighted the potential extreme values in data which have to be further investigated. Additionally, it is evident from the box plot that some variables exhibit larger values than others, with all variables positioned either above or below on the scale. This discrepancy highlights the need for normalization to ensure fair comparisons and accurate modeling across variables.

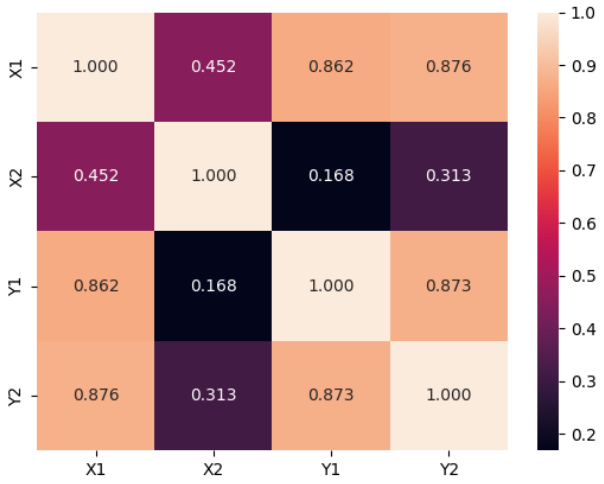
Figure 2. Box plot



Correlation matrix is crucial for understanding the strength and direction of possible linear relationships among variables (figure 3), complementing the insights gained from the matrix scatter plot where some linear trends were noticeable. Notably, strong positive correlations were observed between X1 and both Y1 and Y2, as well as

between Y1 and Y2, suggesting potentially significant associations. However, the correlation between X2 and other variables was notably weaker.

Figure 3. Correlation matrix



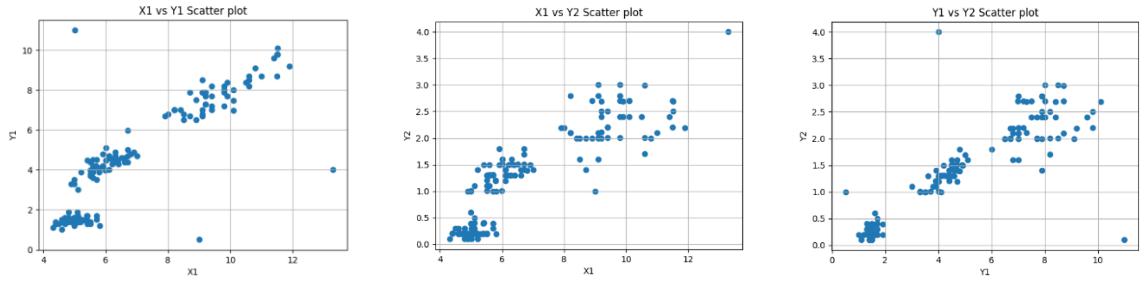
The exploratory analysis has provided valuable insights into the complex relationships within the dataset. The matrix scatter plot revealed distinct clusters and noticeable linear trends, supported by high correlation coefficients observed in the correlation matrix. Particularly, the strong positive correlations between X1 and both Y1 and Y2, as well as between Y1 and Y2, signify potentially significant associations warranting further investigation.

Pattern analysis

We dive deeper into the relationships between variables X1, Y1, and Y2, focusing particularly on patterns observed in their interactions. Building upon the initial exploratory analysis, our attention is directed towards understanding any trends or associations that may exist between these variables.

X1 and both Y1 and Y2, as well as Y1 and Y2 demonstrate strong positive correlation coefficients, so we will look into those variables. Scatter plots of X1 vs Y1, X1 vs Y2, and Y1 vs Y2 (figure 4) reveal distinct clusters, each with noticeable dispersion and arrangement patterns. While X1 vs Y1 shows clusters somewhat aligned along a straight line, the other plots display dispersed and rounded clusters.

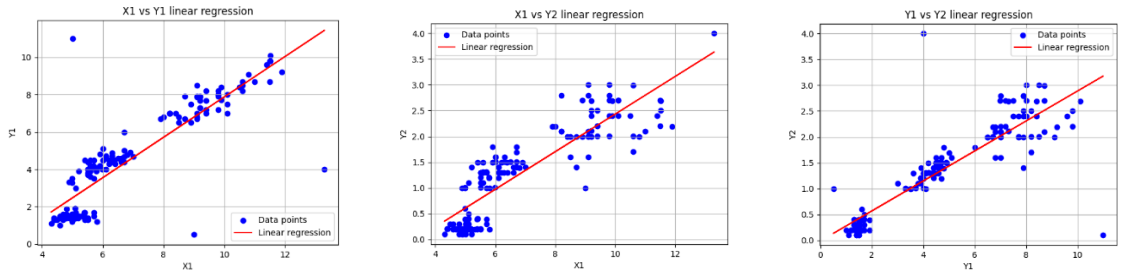
Figure 4. Scatter plots



The high correlation coefficients observed among variables X1, Y1, and Y2, along with the linearity apparent in their scatter plots, provide sufficient evidence of their interrelation. The presence of distinct clusters in these plots further underscores the existence of complex but interconnected patterns among the variables. It can be reasonably concluded that there is indeed a relationship between X1, Y1, and Y2.

The linear regression plots show important evidence of linearity, especially in the X1 vs Y1 plot (figure 5), where data points align along a straight line with minimal deviation. Such a consistent linear trends indicate a systematic relationship between the variables, which could potentially be leveraged for predictive modeling or causal inference. Understanding the extent and nature of this linearity is paramount as it informs the interpretation of the data and provides a foundation for further analysis and modeling.

Figure 5. Linear regressions



In conclusion, our analysis of the relationships between variables X1, Y1, and Y2 has revealed notable patterns and associations. The distinct clusters observed in the scatter plots, particularly the distinct linear positioning evident in the X1 vs Y1 plot, highlight the presence of systematic relationships among these variables. The combination of high correlation coefficients, interrelationship among 3 variables, and pronounced linearity provides sufficient evidence for relationship existence, suggesting that further exploration of these patterns may yield valuable insights.

Analysis of components (PCA)

We conducted Principal Component Analysis (PCA), a method for dimensionality reduction, and explore clustering within multidimensional datasets. By transforming variables, PCA reveals underlying data structure, while analyzing variance and loadings provides insights into significant sources of variation. This analysis not only simplifies complex data but also facilitates clustering exploration, offering a comprehensive understanding of the dataset's structure and patterns.

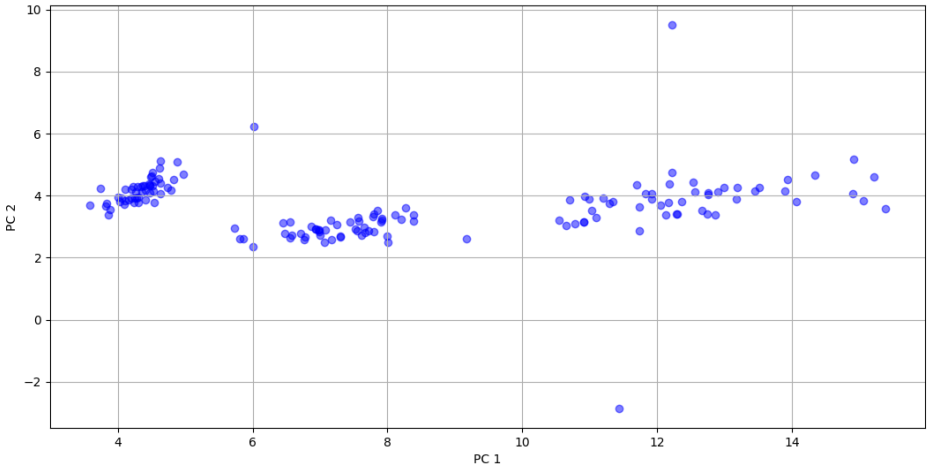
PCA has been conducted to identify the most significant sources of variance within the data. The table provided showcases the percentage contribution of each principal component (PC) to the total variance in the dataset (table 1). Notably, PC1 accounts for the majority of the variance (90.1%), followed by PC2 (7.31%), with PC3 and PC4 contributing smaller proportions. Given that PC1 and PC2 collectively explain 97.41% of the variance, retaining these two components is sufficient for capturing the majority of the dataset's variability. It is crucial for guiding further analysis and interpretation, as it informs the selection of the most informative components for subsequent modeling and clustering endeavors.

Table 1. PC variance contributions

Principal Component	Contribution
PC 1	90.1%
PC 2	7.31%
PC 3	1.57%
PC 4	1.03%

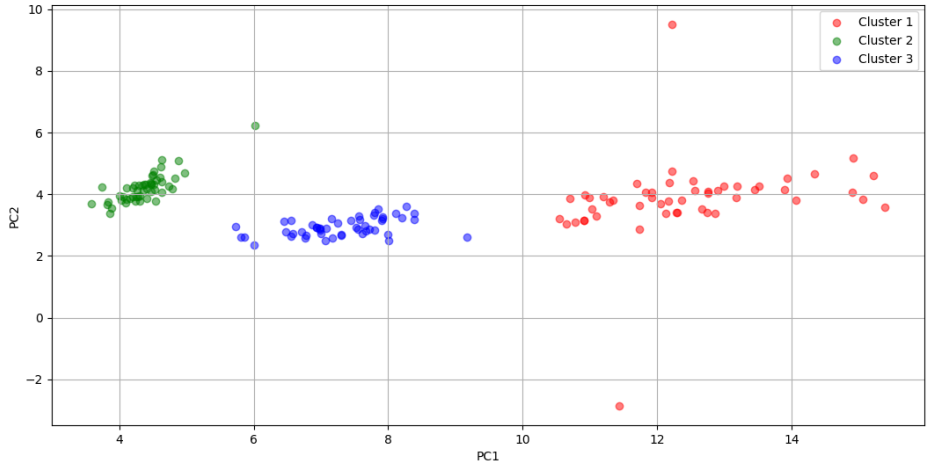
PCA scatter plot represents the projection of observations onto the two most significant principal components identified through PCA (figure 6). Notably, the plot reveals the presence of three distinct clusters, each exhibiting different shapes and levels of dispersion. The leftmost cluster appears to be the most compact, while the other two clusters display more elongated shapes with varying degrees of spread. Additionally, only four observations are identified as outliers, suggesting the overall coherence and structure within the dataset.

Figure 6. PCA Scatter plot



Implementing the K-means algorithm further validates the clustering hypothesis suggested by scatter plots. By segmenting the data into three distinct clusters, K-means shows underlying groupings within the dataset. In the scatter plot, each cluster is delineated by a distinct color: red for the rightmost cluster with the highest spread, green for the leftmost compact cluster, and blue for the centrally positioned cluster (figure 7). This color-coded representation highlights the algorithm's effectiveness in accurately partitioning the data, including appropriately assigning outliers to their nearest clusters.

Figure 7. PCA Scatter plot with clusters



In conclusion, our analysis of components through Principal Component Analysis (PCA) has revealed clear clustering patterns within the dataset. The presence of distinct clusters identified in the PCA scatter plot underscores the inherent grouping and structure within the data. Additionally, the identification of only four outliers

suggests a high level of coherence within the clusters. With PC1 and PC2 collectively explaining a significant portion of the dataset's variability, findings highlight the robustness of the clustering hypothesis. These observations affirm the utility of PCA in uncovering underlying patterns and facilitating the interpretation of complex datasets. The application of the K-means algorithm further supported the clustering hypothesis, providing additional validation of the presence of distinct groups within the data.

Conclusion

In summary, our analysis has uncovered significant insights into the relationships and structures within the dataset. Firstly, we have established that variables X1, Y1, and Y2 exhibit linear relationships, as evidenced by high correlation coefficients, the clear linearity observed in the scatter plots, and the clustering of data points around certain lines. This finding suggests a strong correlation between plant size measurements, implying that some measurements may redundantly capture similar information.

Secondly, our exploration has revealed the presence of distinct clusters within the dataset. This clustering phenomenon was evident across multiple analyses, including the observation of clear clusters on scatter plots, the arrangement of distinct clusters in a line on the PCA plot, and the successful segmentation of data into three clusters using the K-means algorithm. The identification of these clusters suggests the presence of three distinct species of plants in the dataset.

Overall, these findings underscore the importance of careful analysis and interpretation of data, providing valuable insights that can inform decision-making processes and further research.

Code

```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from mpl_toolkits.mplot3d import Axes3D
from sklearn.linear_model import LinearRegression
from sklearn.cluster import KMeans

data = pd.read_excel("Assignment8_data.xlsx").set_axis(["X1", "X2", "Y1", "Y2"], axis=1)

# matrix pair plot
sns.pairplot(data)
plt.show()

# box plot
sns.boxplot(data)
plt.show()

# correlation matrix
corr_matrix = data.corr()
print("\nCorrelation matrix:")
print(corr_matrix)
sns.heatmap(corr_matrix,annot=True,fmt=".3f")
plt.show()

# covariance matrix
cov_matrix = data.cov()
sns.heatmap(cov_matrix,annot=True,fmt=".2f")
plt.show()
```

```python

# Scatter plots
# X1 vs Y1
plt.scatter(data["X1"].values, data["Y1"].values)
plt.title("X1 vs Y1 Scatter plot")
plt.xlabel("X1")
plt.ylabel("Y1")
plt.grid(True)
plt.show()
# X1 vs Y2
plt.scatter(data["X1"].values, data["Y2"].values)
plt.title("X1 vs Y2 Scatter plot")
plt.xlabel("X1")
plt.ylabel("Y2")
plt.grid(True)
plt.show()
# Y1 vs Y2
plt.scatter(data["Y1"].values, data["Y2"].values)
plt.title("Y1 vs Y2 Scatter plot")
plt.xlabel("Y1")
plt.ylabel("Y2")
plt.grid(True)
plt.show()

# 3D scatter plot of X1, Y1, Y2
fig = plt.figure()
ax = fig.add_subplot(111, projection='3d')
ax.scatter(data["X1"], data["Y1"], data["Y2"])
ax.set_xlabel('X1')
ax.set_ylabel('X2')
ax.set_zlabel('X3')
ax.set_title('3D Scatter plot')
plt.show()
```

```
# LINEAR REGRESSIONS
# X1 vs Y1
X = data['X1'].values.reshape(-1, 1)
Y = data['Y1'].values
model = LinearRegression()
model.fit(X, Y)
m = model.coef_[0]
b = model.intercept_
print("X1 vs Y1: y = {:.3f}x + {:.3f}".format(m, b))
# X1 vs Y1
plt.scatter(X, Y, color='blue', label='Data points')
plt.plot(X, model.predict(X), color='red', label='Linear regression')
plt.title('X1 vs Y1 linear regression')
plt.xlabel('X1')
plt.ylabel('Y1')
plt.legend()
plt.grid(True)
plt.show()
# X1 vs Y2
X = data['X1'].values.reshape(-1, 1)
Y = data['Y2'].values
model = LinearRegression()
model.fit(X, Y)
m = model.coef_[0]
b = model.intercept_
print("X1 vs Y2: y = {:.3f}x + {:.3f}".format(m, b))
plt.scatter(X, Y, color='blue', label='Data points')
plt.plot(X, model.predict(X), color='red', label='Linear regression')
plt.title('X1 vs Y2 linear regression')
plt.xlabel('X1')
plt.ylabel('Y2')
plt.legend()
plt.grid(True)
plt.show()
```

```
# Y1 vs Y2
X = data['Y1'].values.reshape(-1, 1)
Y = data['Y2'].values
model = LinearRegression()
model.fit(X, Y)
m = model.coef_[0]
b = model.intercept_
print("X1 vs Y2: y = {:.3f}x + {:.3f}".format(m, b))
plt.scatter(X, Y, color='blue', label='Data points')
plt.plot(X, model.predict(X), color='red', label='Linear regression')
plt.title('Y1 vs Y2 linear regression')
plt.xlabel('Y1')
plt.ylabel('Y2')
plt.legend()
plt.grid(True)
plt.show()

# PCA
eigenvalues, eigenvectors = np.linalg.eig(cov_matrix)
sorted_indices = np.argsort(eigenvalues)[::-1]
eigenvalues = eigenvalues[sorted_indices]
eigenvectors = eigenvectors[:, sorted_indices]
total_variance = np.sum(eigenvalues)
percent_contributions = (eigenvalues / total_variance) * 100
print("\nEigenvalues:")
print(eigenvalues)
print("\nPercent contribution to variance:")
for i, percent in enumerate(percent_contributions):
    print(f"PC{i + 1}: {percent:.2f}%")
eigenvector_pc1 = eigenvectors[:, 0]
eigenvector_pc2 = eigenvectors[:, 1]
```

```
129
130 print("\nEigenvector for PC1:")
131 print(eigenvector_pc1)
132 print("\nEigenvector for PC2:")
133 print(eigenvector_pc2)
134
135 pc1_values = np.dot(data, eigenvector_pc1)
136 pc2_values = np.dot(data, eigenvector_pc2)
137 pca_data = pd.DataFrame({"pc1":pc1_values,"pc2":pc2_values})
138 # PCA plot
139 plt.figure(figsize=(8, 6))
140 plt.scatter(pc1_values, pc2_values, c='b', alpha=0.5)
141 plt.xlabel("PC 1")
142 plt.ylabel("PC 2")
143 plt.grid(True)
144 plt.show()
145
146 # cluster analysis on PCA data
147 pca_num_data = pca_data[["pc1","pc2"]].values
148 kmeans = KMeans(n_clusters=3)
149 kmeans.fit(pca_num_data)
150 cluster_labels = kmeans.labels_
151 pca_data["Cluster"] = cluster_labels
152 pca_data["Cluster"] += 1
153 # cluster plot
154 colors = ["red", "green", "blue"]
155 plt.figure(figsize=(8, 6))
156 for cluster_id, group in pca_data.groupby("Cluster"):
157     plt.scatter(group["pc1"], group["pc2"], color=colors[cluster_id-1], alpha=0.5, label=f"Cluster {cluster_id}")
158 plt.xlabel("PC1")
159 plt.ylabel("PC2")
160 plt.grid(True)
161 plt.legend()
162 plt.show()
163
```