

Statistical foundation of Data Sciences

Practical- 10

Roll number: GF202344767

Name: Ishita Mehta

Workflow and Interpretations

Importing the Dataset & Viewing Summary Statistics

The first step is loading the Wine dataset and generating **basic statistics using `describe()`**. This helps understand:

- The range of each feature
- Mean, variance, spread
- Whether values differ significantly between features
- Identifying skewed or highly varying variables

This overview is necessary before performing any visualizations.

Boxplot by Output Labels

For the boxplot, the variable "**alcohol**" was chosen because:

- It shows clear variation across wine classes
- Alcohol content is a primary chemical characteristic that can separate wine types
- It is easy to interpret in a visual format
- It often appears in wine quality studies as a distinguishing factor

Plotting alcohol vs. class helps check whether **any wine class consistently has higher or lower alcohol levels**.

Scatterplot Using Two Variables

The scatterplot uses:

- **Alcalinity of Ash**

- **Malic Acid**

These variables were selected because:

- They represent different chemical measurements (one mineral-related, one acid-related)
- Both vary widely within the dataset
- They help explore relationships between acidity and mineral composition
- They provide a good 2D spread instead of forming a narrow cluster

A scatterplot with these two allows us to see whether any **natural grouping or separation** appears among the wines.

Covariance Matrix Plot

The covariance matrix helps understand **how pairs of features grow or shrink together**.

Covariance was chosen (instead of correlation) because your practical specifically requires studying how raw variables co-vary before scaling.

We adjust the heatmap color scale because covariance values vary greatly (some around 0.1, others around 300+).

This ensures differences become visible instead of the heatmap appearing a single color.

The covariance matrix helps identify which chemical properties rise or fall together — useful for dimensionality reduction later.

Data Scaling (Standardization)

Scaling is important because:

- Wine features are on **different units** (mg/L, g/L, alcohol %, ash concentration)
- Some features have values around **1 or 2**, others reach **1000+**
- PCA and distance-based methods require equal weighting

Standardization transforms all features to **mean = 0 and standard deviation = 1**, ensuring fair comparison.

PCA for Better Class Separation

Principal Component Analysis (PCA) reduces high-dimensional data into new axes (principal components) that capture maximum variance.

Why PCA was used:

- To visualize wine classes in 2D
- To check if chemical properties naturally separate wine types
- To remove noise and highlight the most informative features

Why PCA works well on the Wine dataset:

- Wine features are strongly correlated (high covariance)
- PCA condenses this shared information into fewer dimensions
- After scaling, PCA components show excellent class separation

Using **PC1 and PC2**, wines from different classes tend to cluster separately, demonstrating that chemical composition effectively differentiates wine types.

GitHub Repository link

https://github.com/pineapplesdontbelongonpizza/CSU1658_practical1_Testing_Pandas_and_Numpy.git