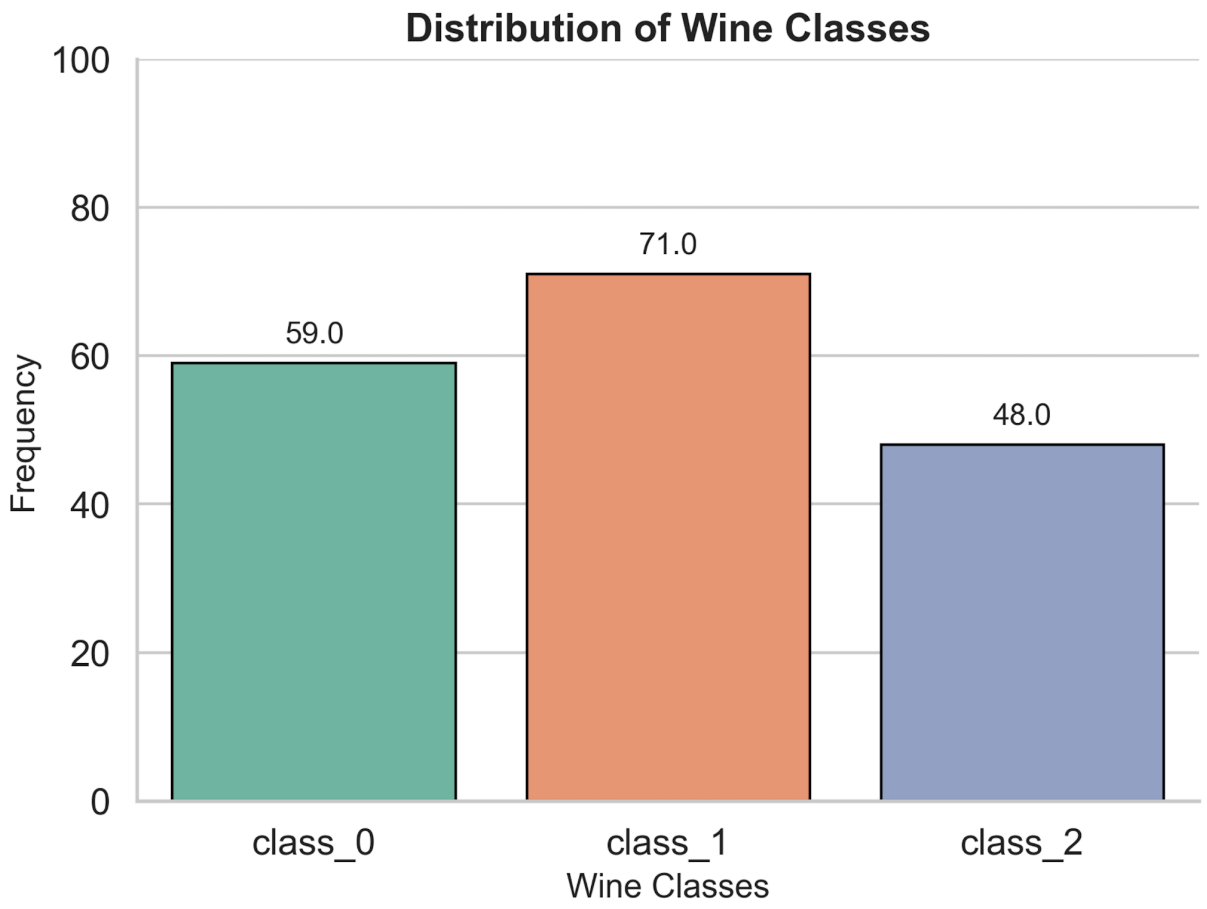
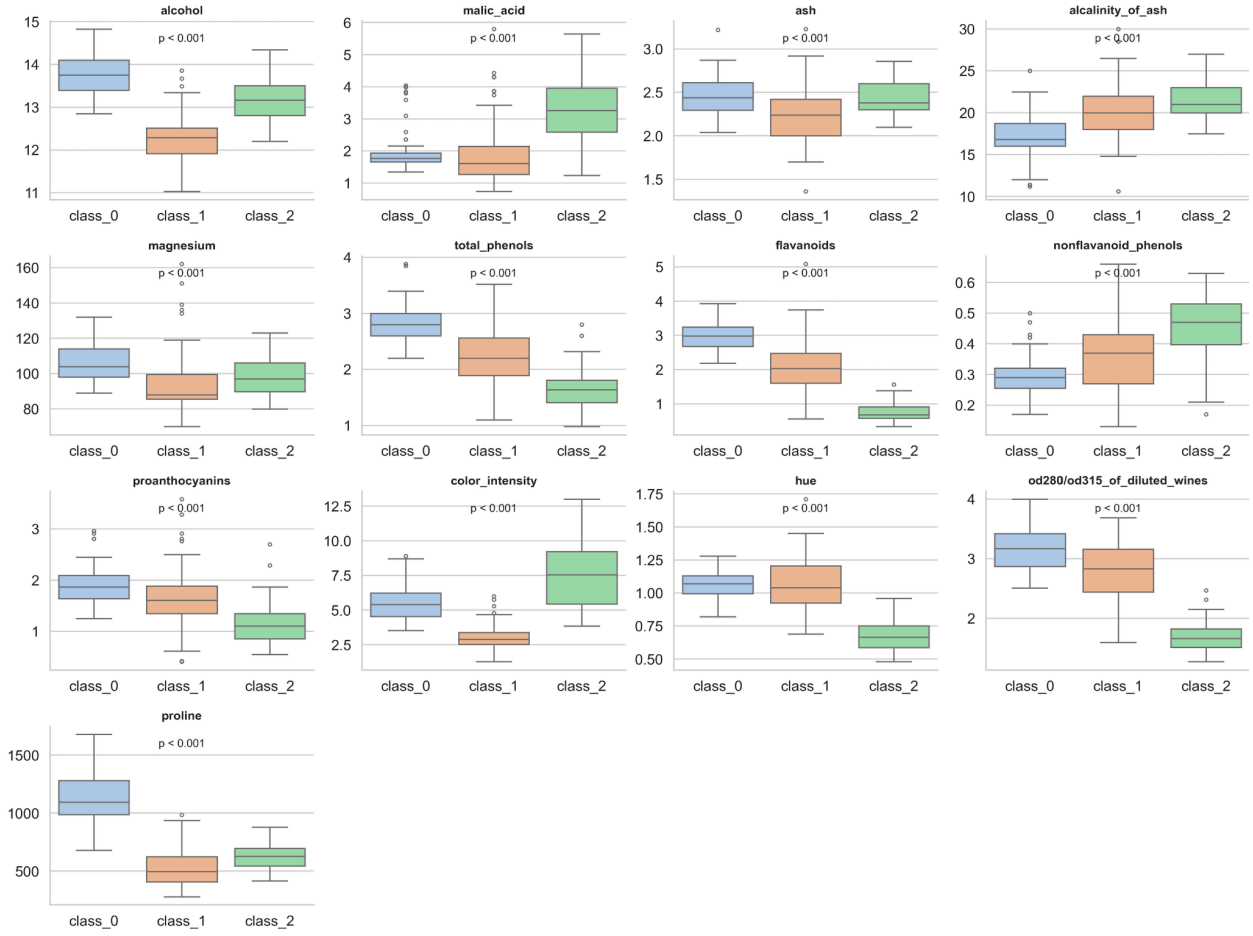
Wine_DescriptiveAnalysis.py

a) Distribution of wine classes



Analysis: The Frequency of each wine class i.e class0, class1, class2 helps in assessing the class representation in each dataset

b) Feature distribution across classes/ Significant differences in features



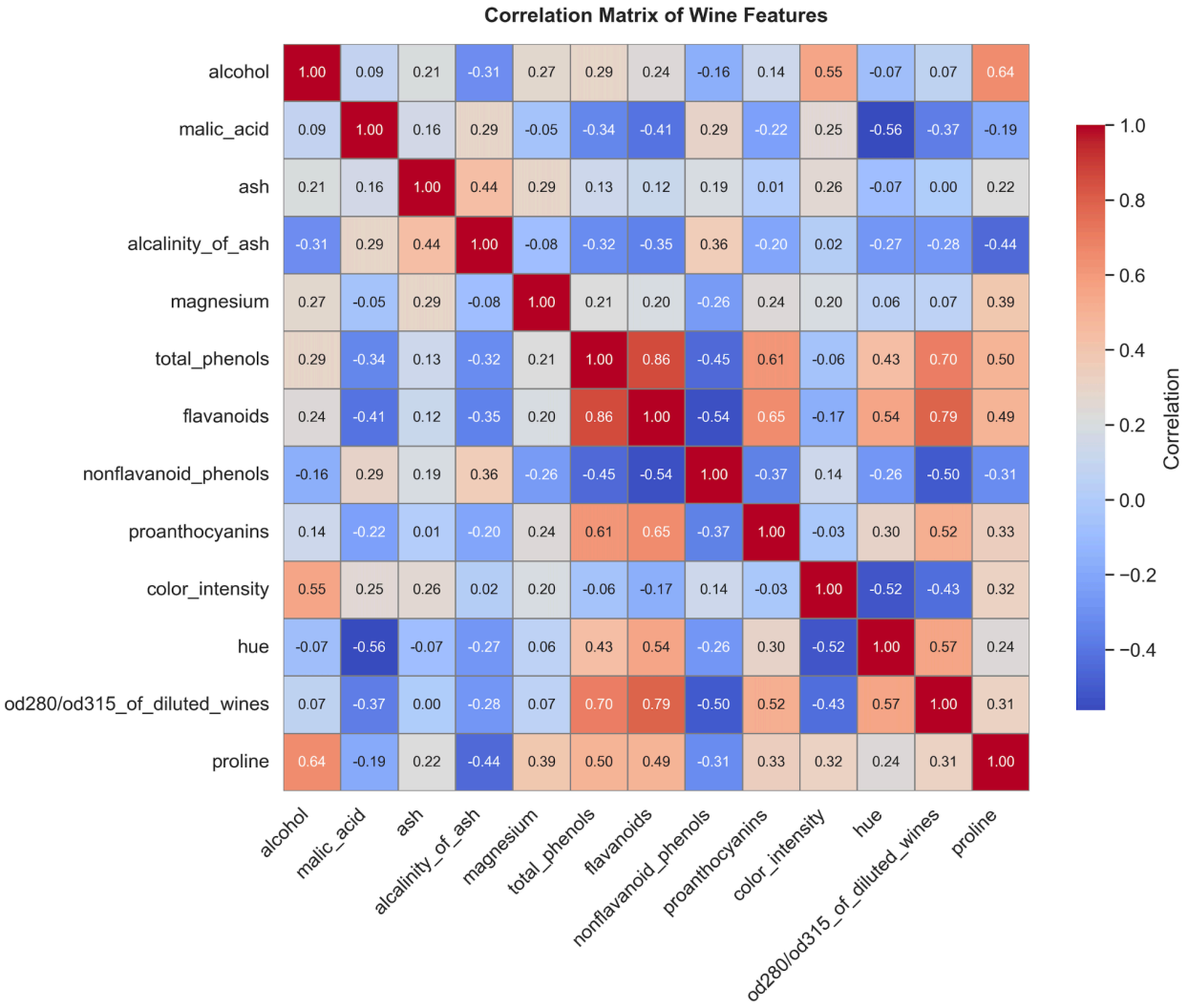
Analysis: Kruskal-Wallis Test was applied in the dataset according to the wine class. It is a type of non-parametric test which is used to determine if there is significant difference between the three classes. The boxplots display the distribution of each feature across wine classes. The test determines the distributions of the feature between the classes.

Hypothesis:

- **Null Hypothesis :** No significant difference between the wine classes
- **Alternative Hypothesis :** Significant difference between the wine classes

Results: All p-values corresponding to each feature were **less than 0.001** and hence the null hypothesis is rejected. This means there are significant differences in the feature distributions between the wine classes.

c) Correlations between chemical components



Analysis: Pairwise correlations between features. Dark red means strong positive and dark blue means strong negative.

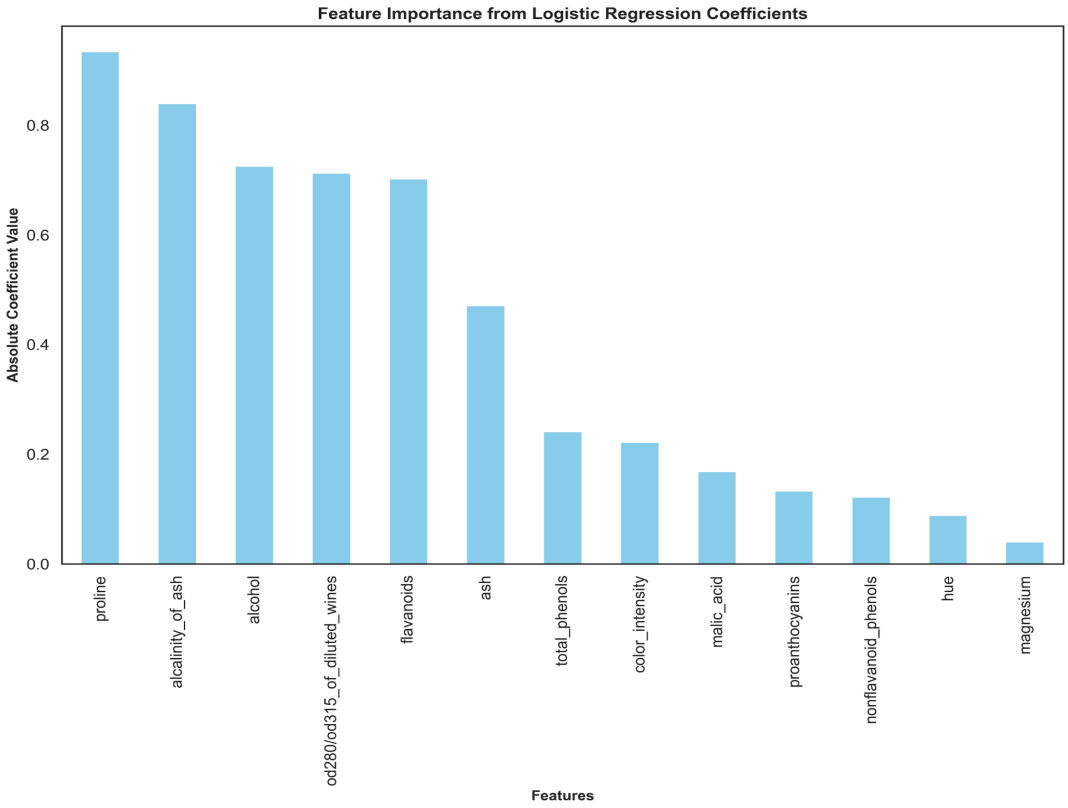
Results: Strong correlation between

- **total_phenols and flavanoids (0.86):** This suggests that wines with higher levels of total phenols also tend to have higher levels of flavanoids, which are a subclass of phenols
- **flavanoids and od280/od315_of_diluted_wines (0.95):** There is a very strong positive correlation
- **total_phenols and od280/od315_of_diluted_wines (0.70):** A strong positive correlation, similar to the relationship with flavanoids
- **color_intensity and proline (0.58):** A moderate positive relationship between the color intensity of the wine and its proline content

- **alcohol with flavanoids (0.55) and od280/od315_of_diluted_wines (0.56):** Indicates a moderate positive correlation between alcohol content and these phenolic measures.

Wine_logistic_regression_classification.py

Logistic regression was used as a baseline classification model to predict the class of wine (class 0, class1, class 2) on the basis of its chemical constituents.



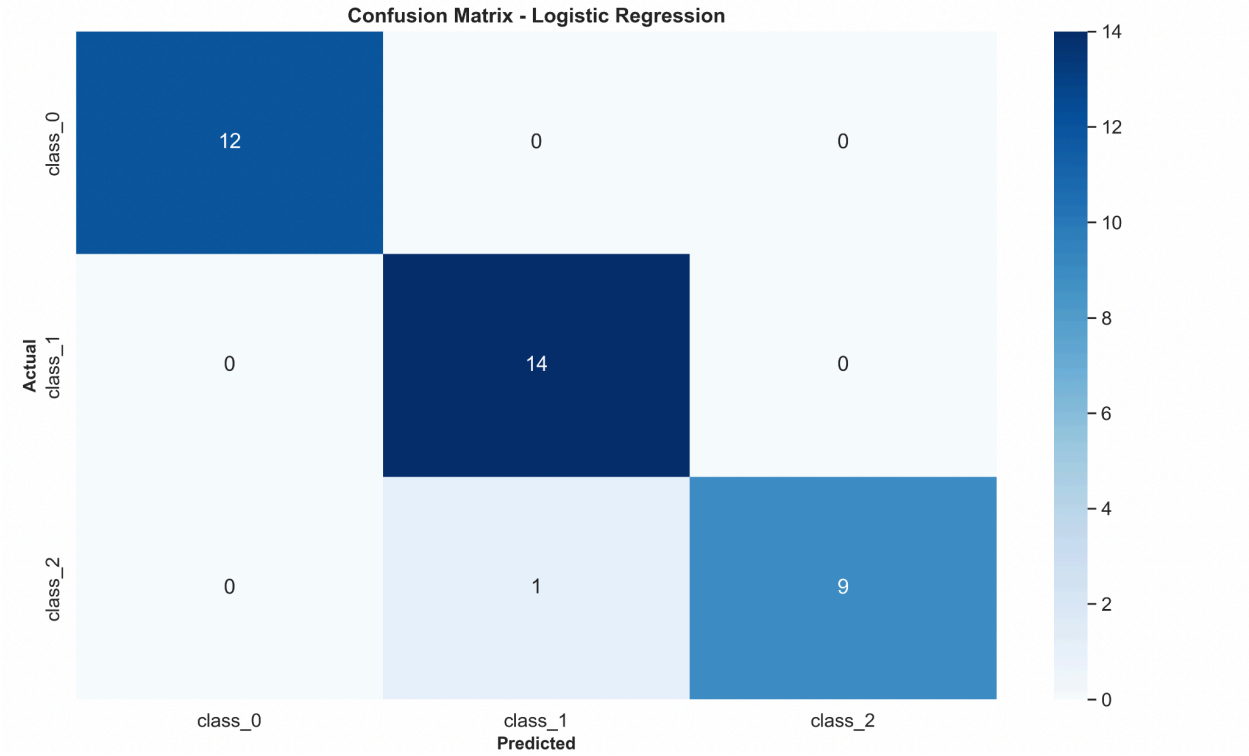
Results: The feature importance was assessed by examining the absolute values of the Logistic Regression model's coefficients. Larger coefficients indicate high influence on the model's classification decisions.

- Proline and alcalinity_of_ash has the highest coefficients, showing their strong influence on the wine classification

- Alcohol and flavanoids also contributes significantly to the model

- Features like hue and magnesium has minimal impact on classification

The accuracy of the model was 0.972 percent.




```
Accuracy: 0.9722222222222222
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        12
           1       0.93      1.00      0.97        14
           2       1.00      0.90      0.95        10

    accuracy                           0.97        36
   macro avg       0.98      0.97      0.97        36
weighted avg       0.97      0.97      0.97        36
```

Results:

- Class 0: Perfect prediction — no false positives or false negatives.

- Class 1: Slight drop in precision (0.93), meaning a few instances were misclassified as class 1 which can be seen in the confusion matrix

- Class 2: Perfect precision, but some true class 2 samples were missed (recall = 0.90)