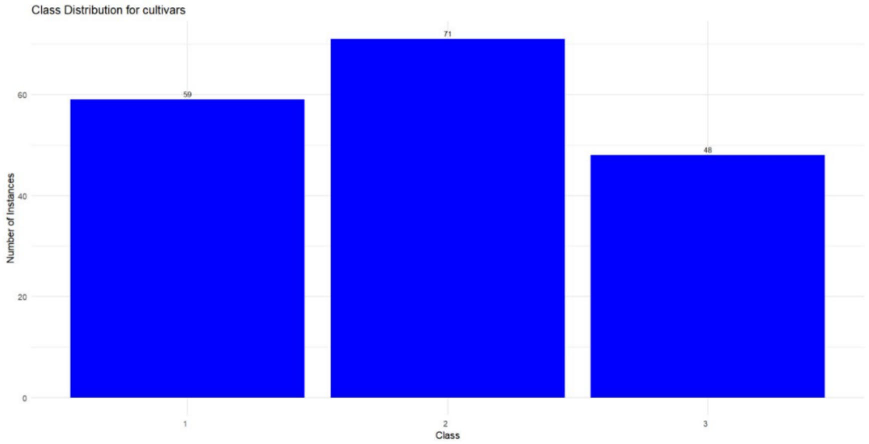
PCA and FastICA

WINE DATASET

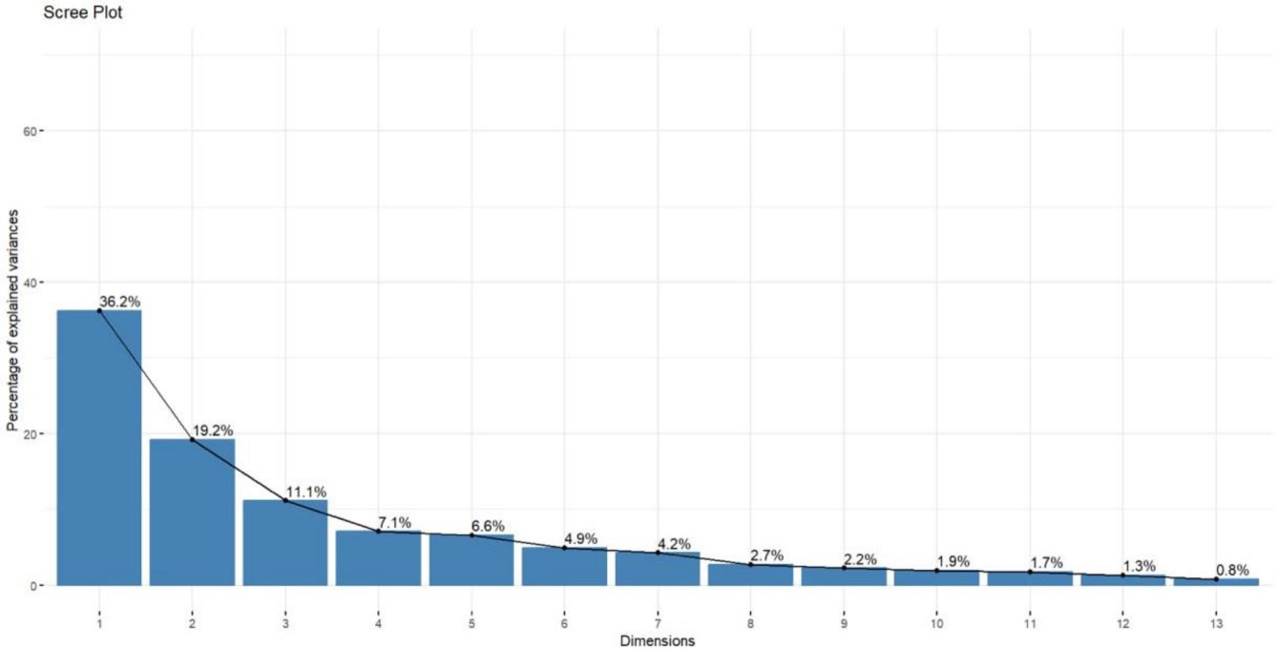
This dataset contains 178 cases and 14 features, of which 13 refers to chemical/physical characteristics of the wine and one, the ‘cultivars’ feature is the wine identifier (i.e., 1, 2 or 3).



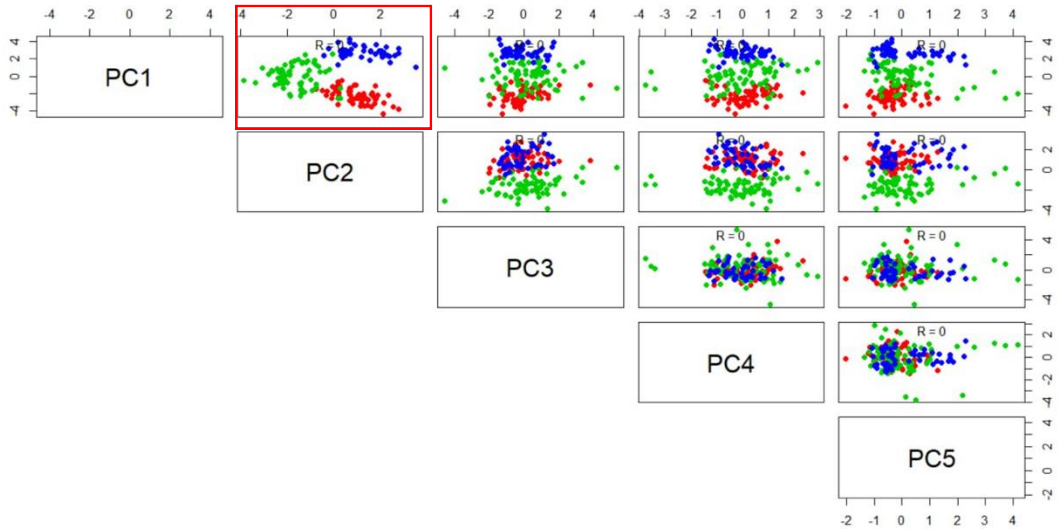
In order to analyze the data effectively, it was necessary to scale it due to variations in the scales of different features. This scaling ensures that all features are treated equally in terms of importance. To achieve this, a standardization process has been employed, resulting in the total variance being determined by the total number of variables.

1) PCA

PCA, or Principal Component Analysis, serves as a method for reducing the dimensions of the dataset. The principal components (PC) are novel variables formed through linear combinations of the original ones. These combinations are designed to make these new variables uncorrelated while preserving most of the data's information in the initial components. It is important to note that when talking about "information," we are referring to variability, where greater variance indicates more information. Even though the purpose of PCA is to decrease the dataset's dimensionality, it generates as many PCs as there are original variables in the dataset, thus conserving the total variance. However, in the end, only the first PCs, which explain most of the variance, are typically selected. This selection is often determined using the 'Elbow method' applied to the Scree Plot.



The Scree plot provides insight into the proportion of variance explained by each PC. By examining the summary of the PCA results, it becomes apparent that the cumulative proportion attributed to the first five PCs is 0.80162. This indicates that when we utilize these five out of the 13 components, approximately 80% of the variance is captured.



Scatterplot Matrix for PCA:

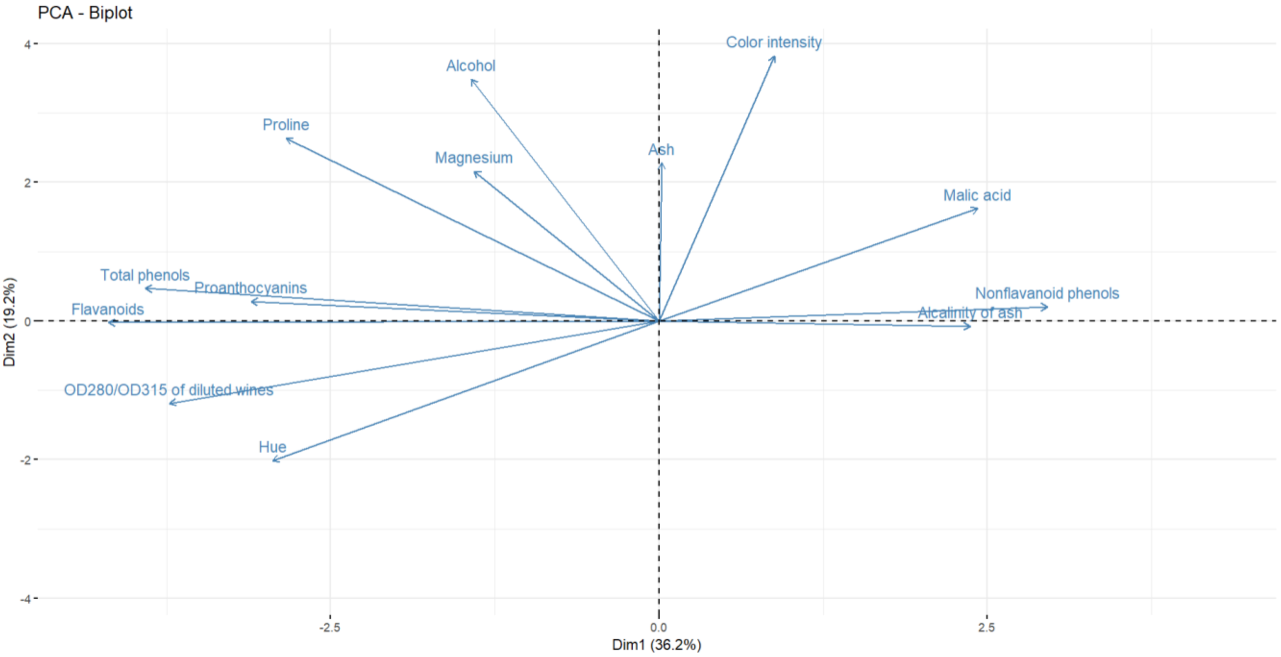
Examining the graph reveals significant overlap among points colored according to the wine type, particularly in the PC3 vs. PC4, PC3 vs. PC5, and PC4 vs. PC5 combinations. When comparing PC2 against the others, we notice a more distinct separation of the green points from the red and blue ones, although these two still tend to overlap considerably. This suggests that the PC2 variable might be more effective in distinguishing one of the wine cultivars (associated with the green points) from the other two.

In summary, upon reviewing all the plots, the PC1 vs. PC2 combination stands out as providing the most effective separations among the three wine cultivars.

Interpretation of the first 2PC:

The PC interpretation is based on the loadings, and the following strategy is used:

1. Ignore loadings close to 0: Loadings near 0 signify that a variable does not significantly contribute to the respective PC. By ignoring these, attention can be directed towards variables with more substantial influence on the component.
2. Look at the positive/negative loadings and assign labels: An evaluation is made regarding the sign of the loadings and an attempt to attribute meaning (what constitutes a large or small unit?). Interpretation considers the absolute values of the loadings since they represent the strength of the relationship between the variable and the principal component. Larger absolute loadings indicate stronger associations between variables and components.
3. Refer to the Biplot for clearer visualization: The Biplot provides a more illustrative representation of the results. The direction and length of the vectors (arrows) on the Biplot convey how each variable contributes to the two principal components. A longer vector indicates a more substantial impact on the principal component, and the direction signifies whether the relationship is positive or negative.



In summary, examining this diagram based on loadings allows for the interpretation of the PCs as follows:

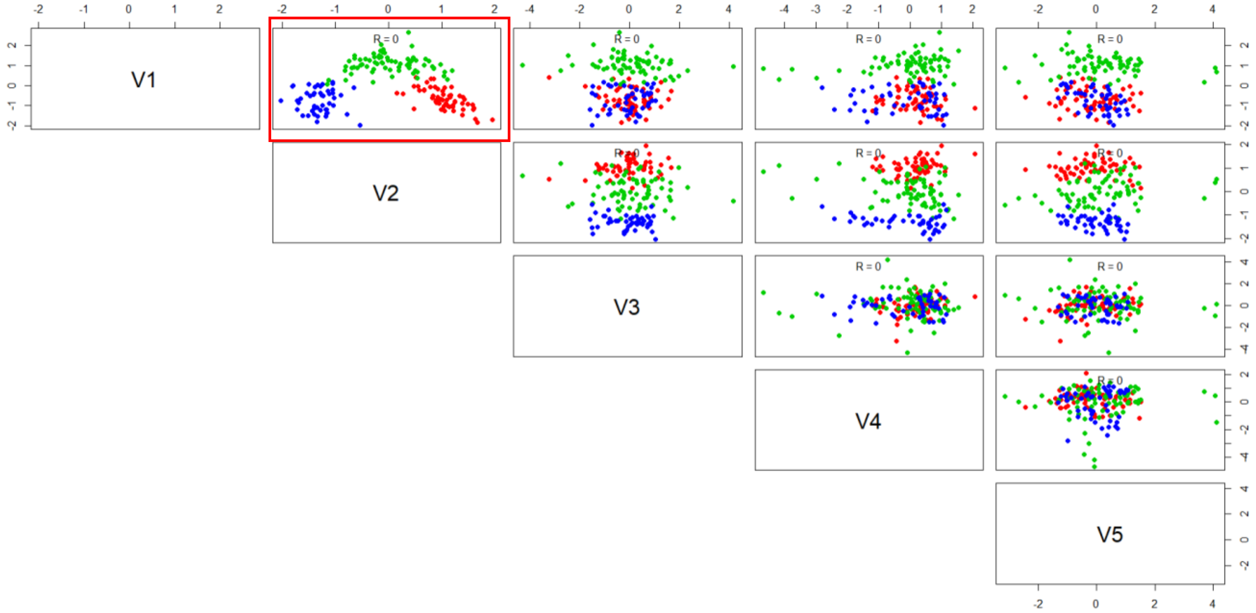
- **PC1**: Exhibits a notable negative correlation with Flavonoids, Total phenols, and the OD280/OD315 of diluted wines. This implies that wines featuring higher values for these characteristics tend to receive lower PC1 scores. Consequently, a lower PC1 score may indicate wines with a more pronounced presence of these chemical attributes.

- **PC2**: Demonstrates a positive connection with Color intensity, Alcohol, and Proline. Wines with higher values in these variables are linked to higher PC2 scores. This suggests that wines with elevated values in these characteristics will attain higher PC2 scores. Thus, a higher PC2 score might suggest wines with a heightened level of these properties.

2) FAST ICA

ICA, or Independent Component Analysis, is a method utilized to break down intricate data into distinct, unrelated segments. The term "independence" signifies that these various components do not exhibit correlations with one another. In ICA, it is not feasible to establish the sequence of these independent components. For the sake of result reproducibility, a seed was set.

Scatterplot Matrix for fastICA:



Examining this graph reveals the following:

- In each plot, R=0, confirming the statistical independence of the components.
- There is significant overlap among the points, particularly when comparing V4 and V5 against all the other variables. In conclusion, after reviewing all the plots, it becomes evident that the combination of V1 vs. V2 provides the most effective separations among the three wine cultivars.