



# Final Project

Detection of fraudulent banana labelling



# Summary :

Introduction

Data Analysis

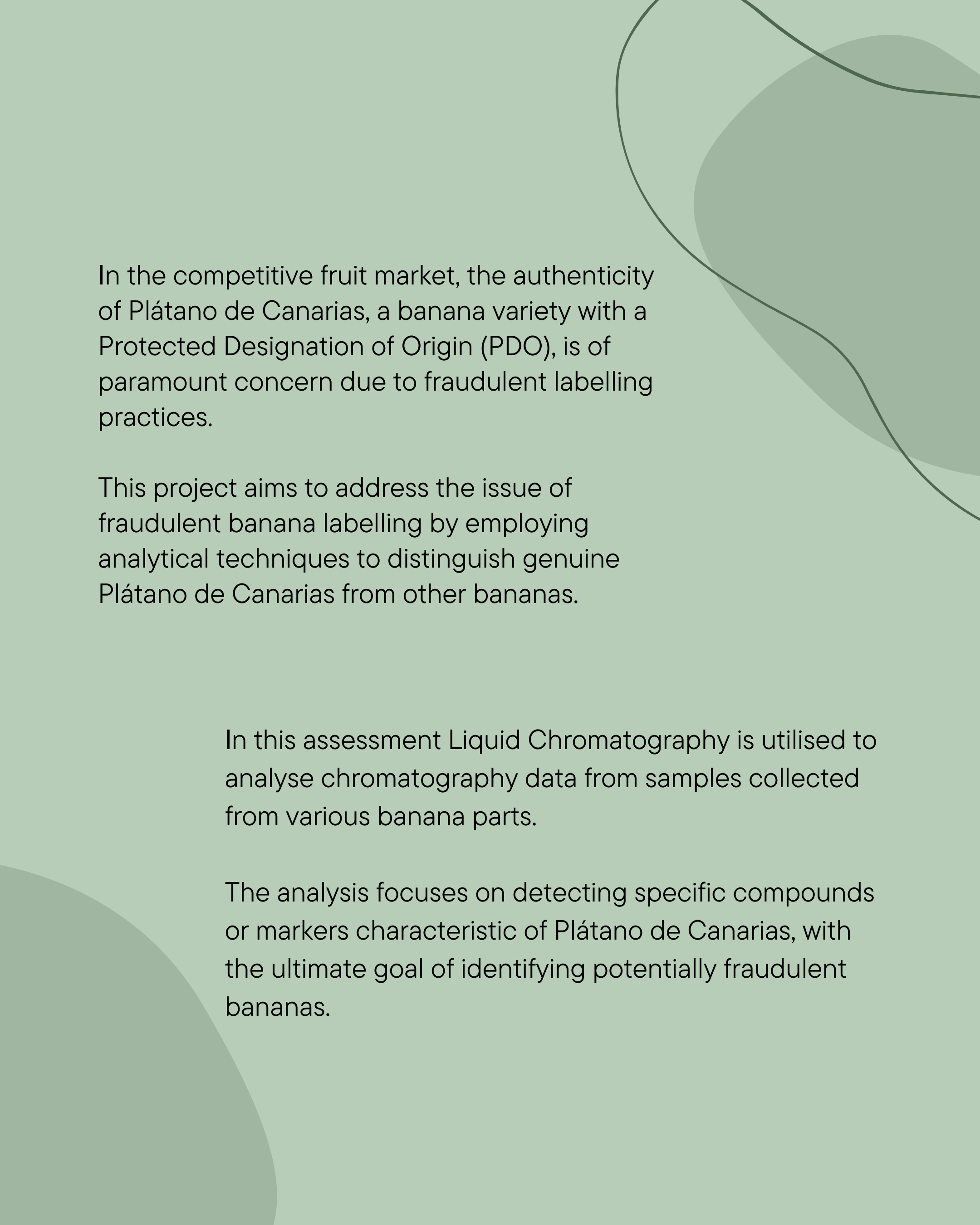
Results

Comments

Conclusion

# Introduction





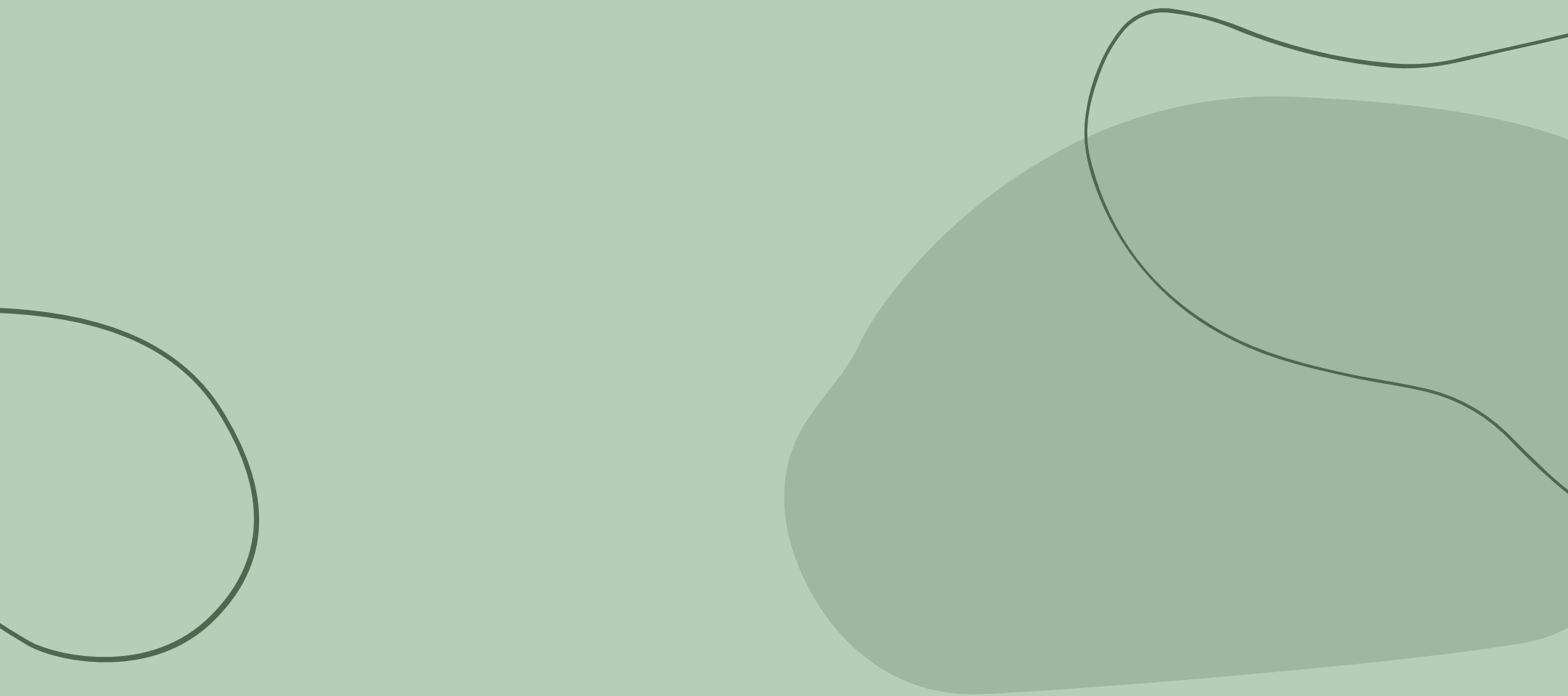
In the competitive fruit market, the authenticity of Plátano de Canarias, a banana variety with a Protected Designation of Origin (PDO), is of paramount concern due to fraudulent labelling practices.

This project aims to address the issue of fraudulent banana labelling by employing analytical techniques to distinguish genuine Plátano de Canarias from other bananas.

In this assessment Liquid Chromatography is utilised to analyse chromatography data from samples collected from various banana parts.

The analysis focuses on detecting specific compounds or markers characteristic of Plátano de Canarias, with the ultimate goal of identifying potentially fraudulent bananas.

# Data Analysis

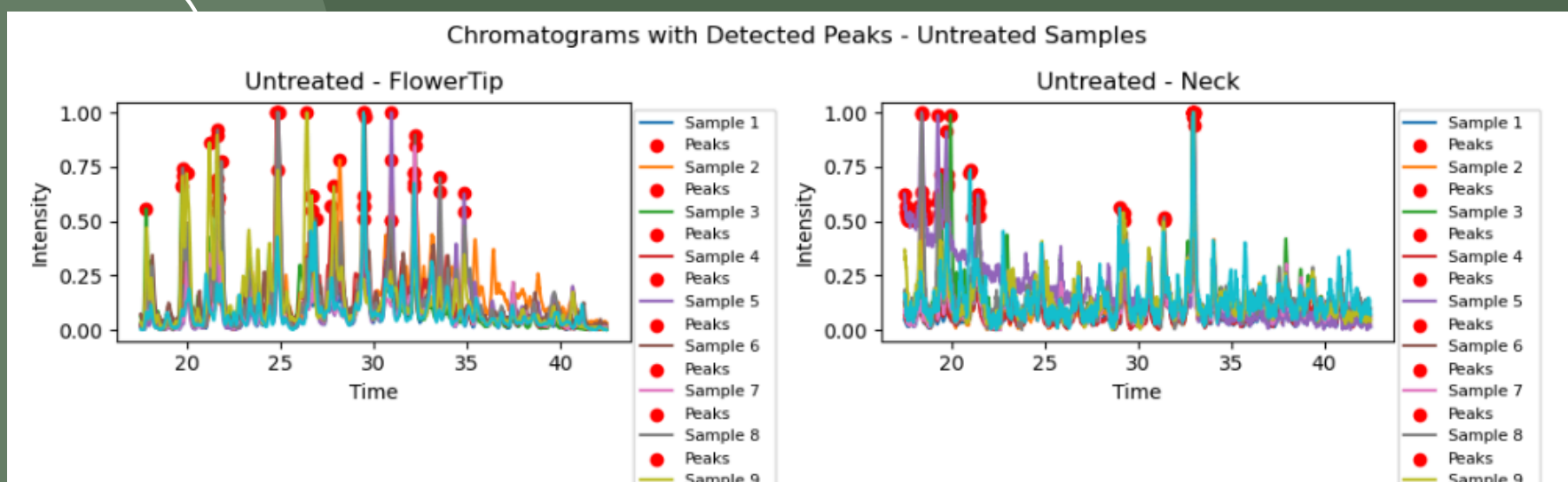


The data analysis process begins with loading the chromatography data obtained from samples of different banana parts. In this case, Liquid Chromatography is employed, the data processing includes loading and cleaning the data, normalising it using Min-Max scaling, detecting peaks in the chromatograms, and performing clustering analysis to identify clusters of samples with similar characteristics.. The chromatograms are then analysed to detect peaks that signify the presence of specific chemicals distinguishing Plátano de Canarias from other bananas.

To load the chromatography data, a function `load_chromatography_data` is defined. This function reads the data from a CSV file and returns a DataFrame containing the chromatography data.

Once the data is loaded, it can be visualized by plotting the chromatograph :

## Example :

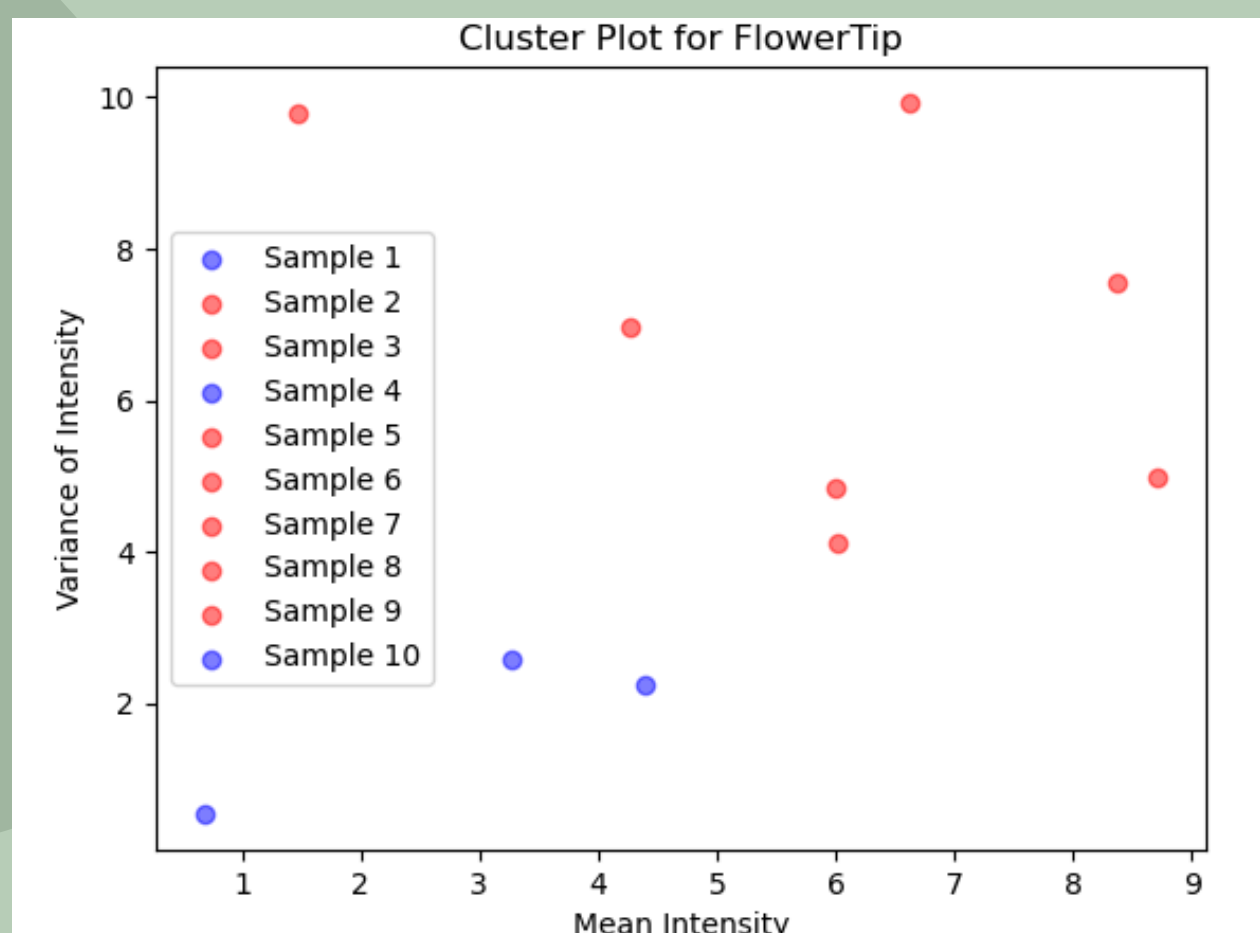


Next, the data is analysed for clustering, which helps in identifying patterns and potential markers indicative of fraudulent labelling. For this, KMeans clustering is applied to the chromatography data. A function `cluster_analysis` is defined to perform this clustering analysis.

The clustering analysis is then applied to the chromatography data, and the clusters are visualized for reference. The function `cluster_analysis` is called within the main code to perform this analysis.

Only thing left to do is to analyse those cluster counting the occurrence of each sample in each cluster to find which of those seems to be fraudulent.

## Example :





The full code and plots are  
available in the deposit area.

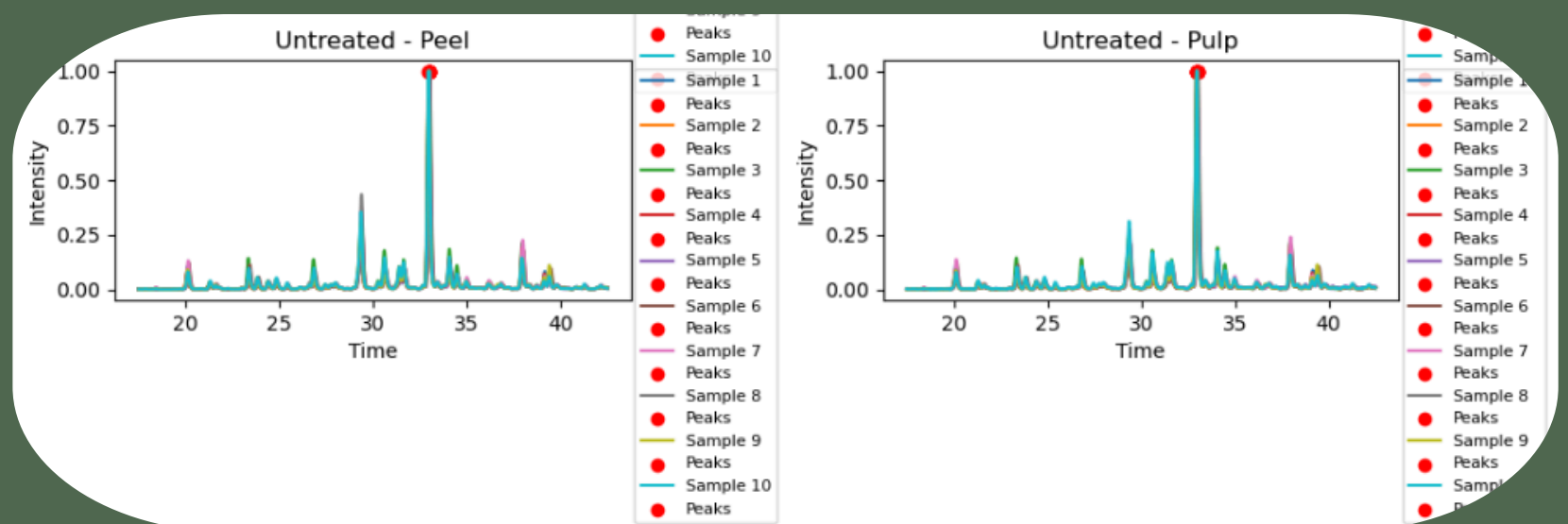




Results :

Based on the analysis of the chromatography data, several findings emerge regarding the potential markers of fraudulent labelling :

The results showcased in the plot of the chromatophy data, suggest that certain parts of the banana, as the peel or pulp, may contain more significant markers of fraudulent labelling than others.



Peaks observed in the chromatograms may indicate the presence of specific chemicals characteristic of Plátano de Canarias.

Clustering analysis helps identify clusters of samples with similar chromatographic profiles, potentially highlighting fraudulent samples that deviate from the norm. As we can see on the table bellow the sample that are the most likely to be fraudulent are **1, 7, 2, 4 and 5**

	Occurrences in Cluster 1	Occurrences in Cluster 2
Banana		
1	5	4
2	5	2
3	3	2
4	2	4
5	1	5
6	3	0
7	3	5
8	5	2
9	4	1
10	2	3

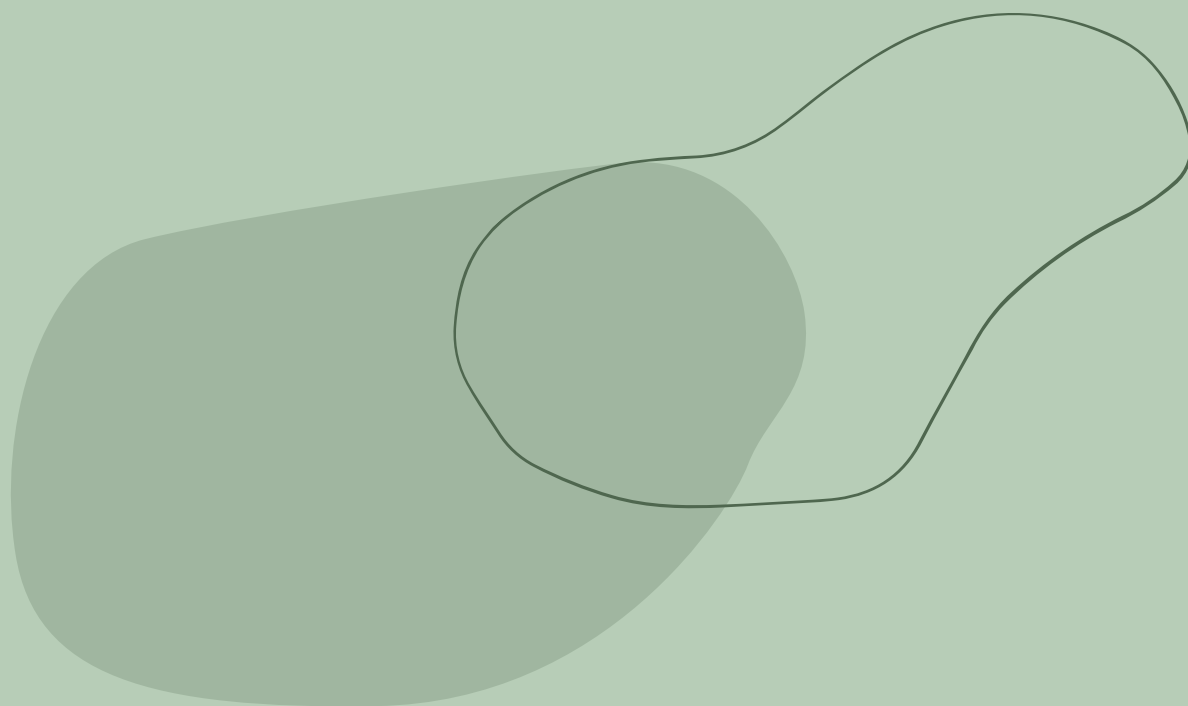
Additionally, the impact of sample concentration on fraud detection is assessed, with variations in peak clarity, resolution, and signal-to-noise ratio observed.



Comments :

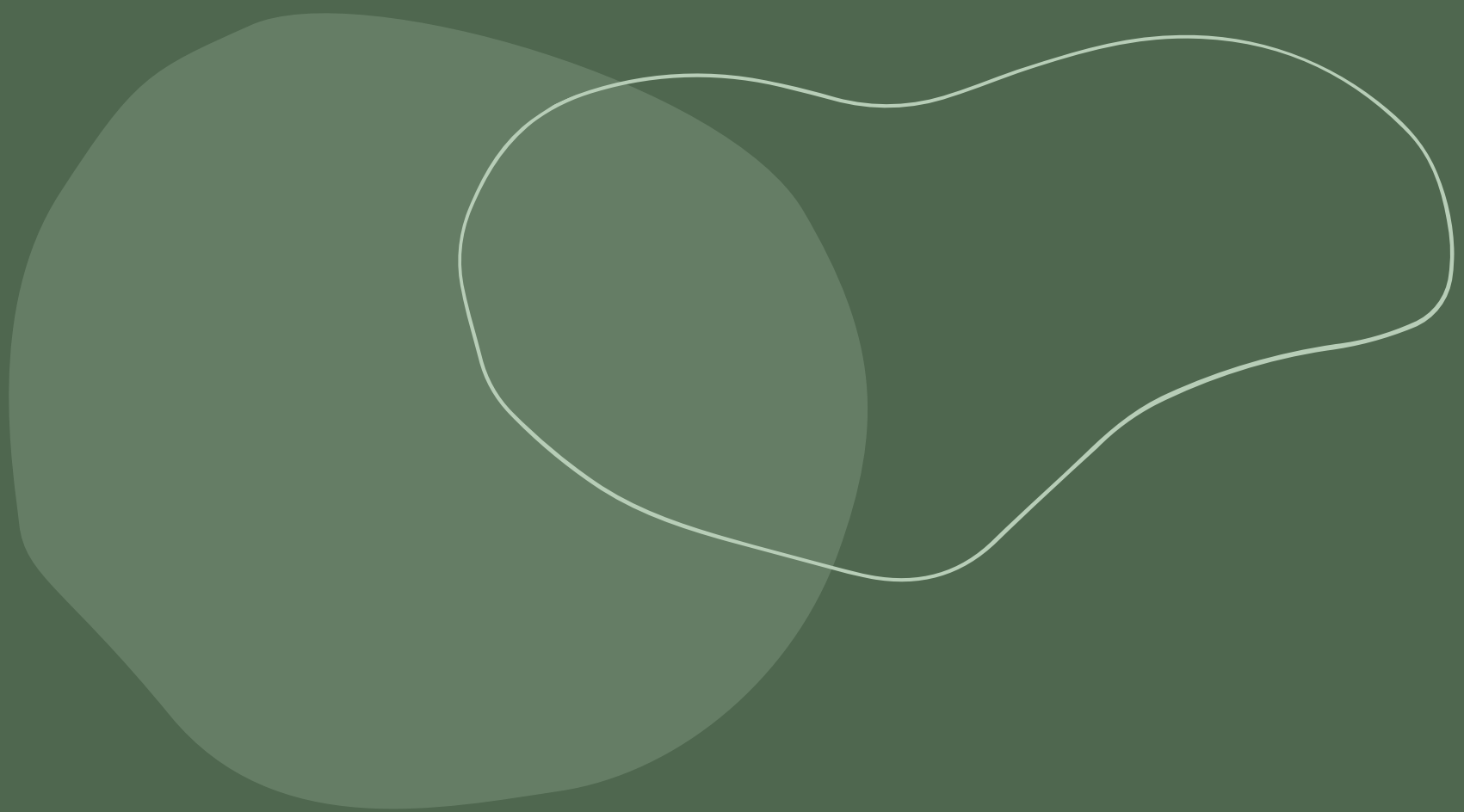
Continuing with the data analysis, after performing the clustering analysis to identify potential markers indicative of fraudulent labelling, the next step is to evaluate the results and draw conclusions based on the findings. Here are main points of possible errors :

The clustering analysis is performed using the KMeans algorithm, which partitions the chromatography data into clusters based on similarity. However, it's essential to interpret the results carefully and consider any limitations of the technique used.

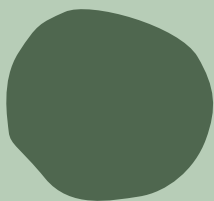


Cluster Interpretation: The clusters obtained from the analysis may not always correspond directly to fraudulent and non-fraudulent samples. Instead, they represent groups of samples with similar characteristics based on the chromatography data.

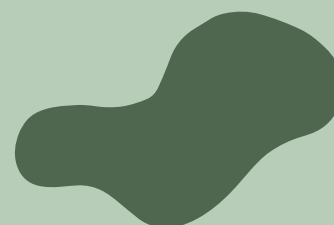
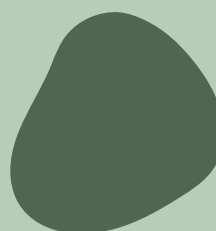
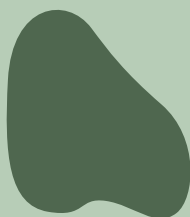
Marker Identification: While clustering can help identify potential markers indicative of fraudulent labelling, further analysis is typically required to validate these markers. This may involve comparing the chromatography data with known standards or conducting additional experiments.

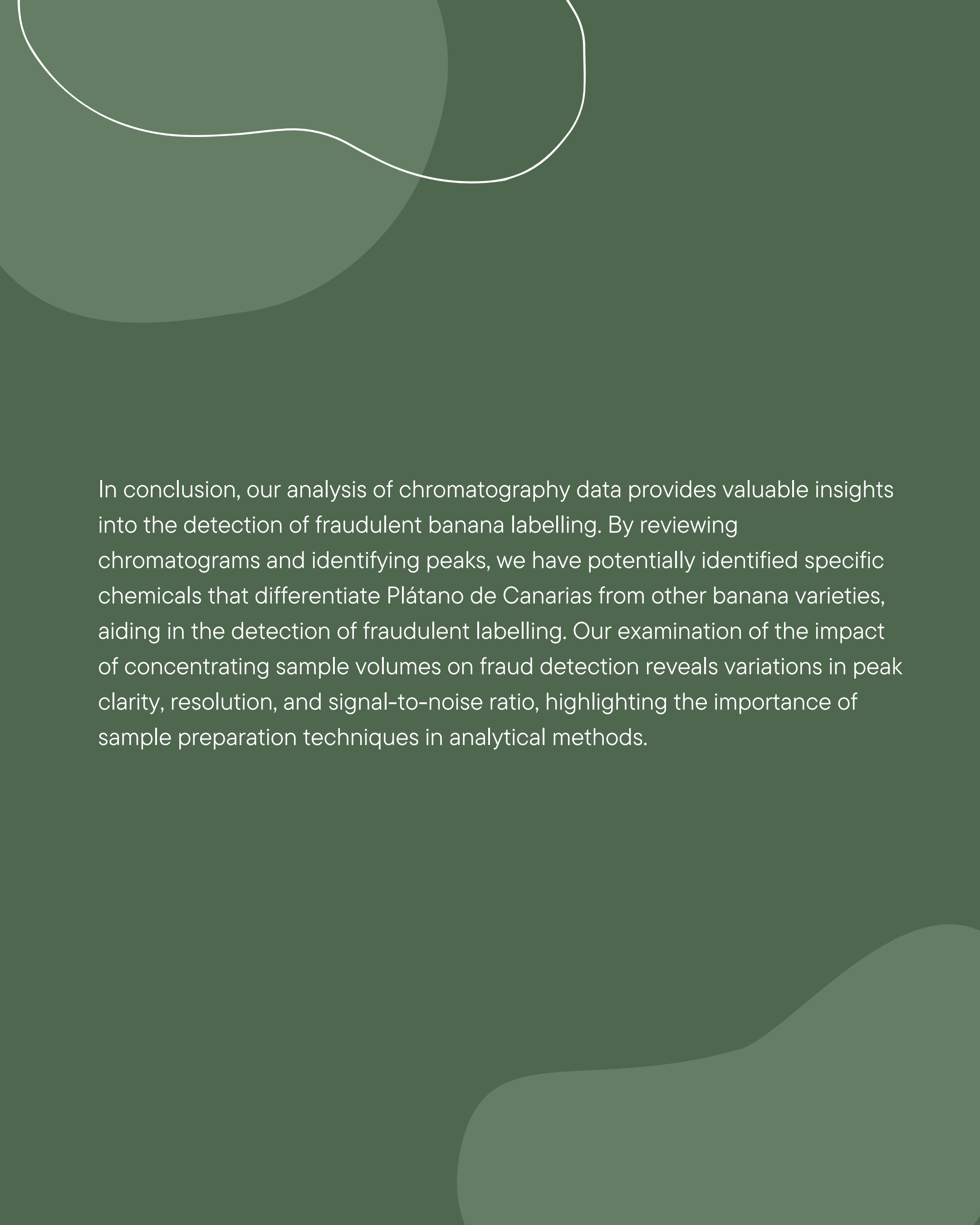


Sample Variability: The variability of the chromatography data across different samples and banana parts should be taken into account when interpreting the results. Factors such as sample preparation techniques, instrument variability, and environmental conditions can contribute to variability in the data.



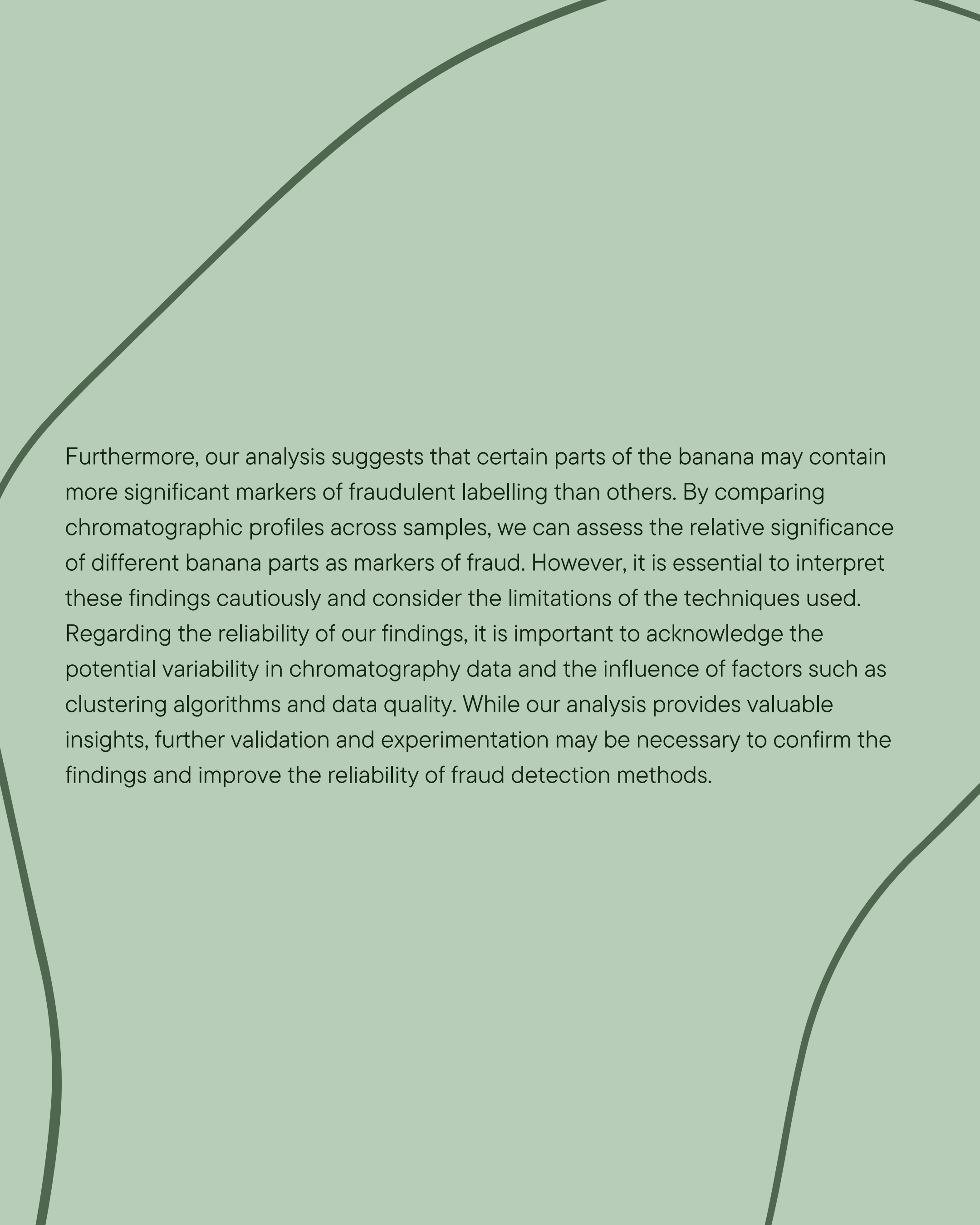
Conclusion :





In conclusion, our analysis of chromatography data provides valuable insights into the detection of fraudulent banana labelling. By reviewing chromatograms and identifying peaks, we have potentially identified specific chemicals that differentiate Plátano de Canarias from other banana varieties, aiding in the detection of fraudulent labelling. Our examination of the impact of concentrating sample volumes on fraud detection reveals variations in peak clarity, resolution, and signal-to-noise ratio, highlighting the importance of sample preparation techniques in analytical methods.





Furthermore, our analysis suggests that certain parts of the banana may contain more significant markers of fraudulent labelling than others. By comparing chromatographic profiles across samples, we can assess the relative significance of different banana parts as markers of fraud. However, it is essential to interpret these findings cautiously and consider the limitations of the techniques used. Regarding the reliability of our findings, it is important to acknowledge the potential variability in chromatography data and the influence of factors such as clustering algorithms and data quality. While our analysis provides valuable insights, further validation and experimentation may be necessary to confirm the findings and improve the reliability of fraud detection methods.