Wine Clustering Analysis

Wine is an alcoholic beverage with which most of us are familiar, as it has accompanied us in special moments, from the dinner on an ordinary day after a long workday, to the toast you´ll have when you get to celebrate your wedding. We know that wine is a very versatile and elegant drink, but I don't believe the secret formula is in the public domain. I had the opportunity to conduct a thorough analysis of the 13 components of 177 wines in three vineyards, and I would like to share my findings.

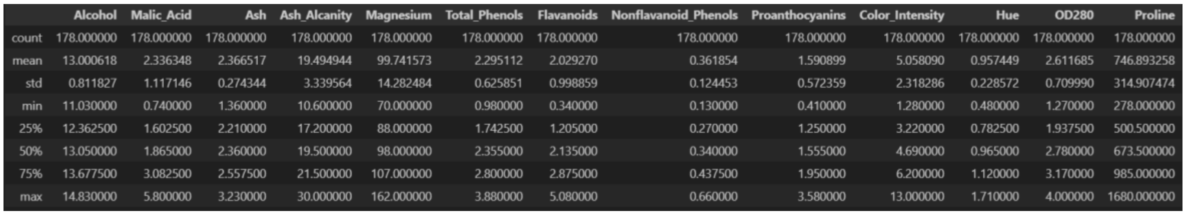
I started with questions about the file based solely on my knowledge of wine. What compounds determine the alcohol content of the beverage? Will there be any reaction of some compound with alcohol? What determines the intensity of the color? Why is it said that wine is good for the heart?

I. Exploratory Analysis

Due to my limited knowledge in wines, it was important to conduct an exploratory analysis before establishing a main hypothesis. To achieve this, I utilized various Python visualization tools such as box plots, scatter plots, and Pearson coefficient graphs

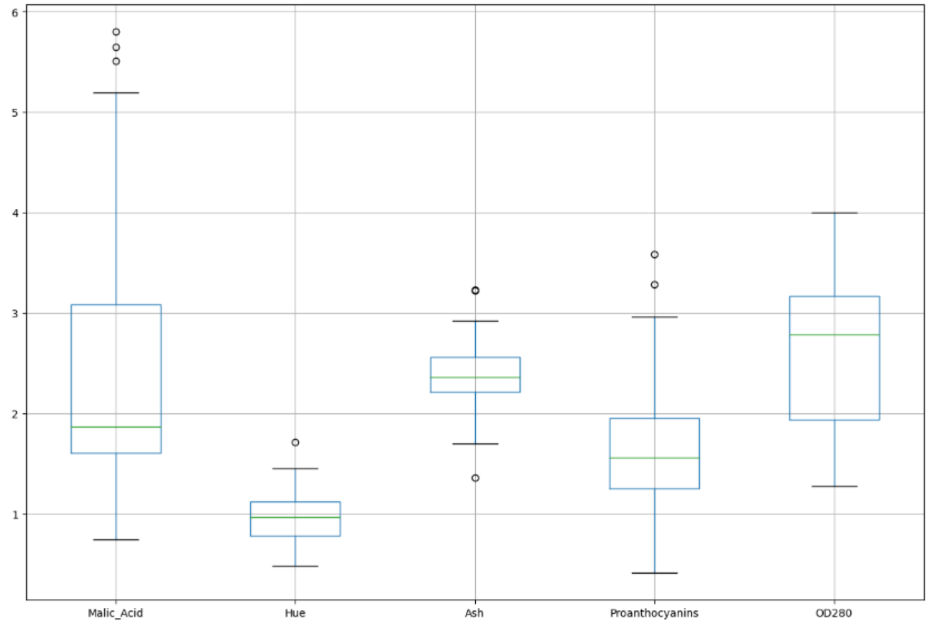
i. Boxplots

Box plots visualize data with respect to their quartiles. The green line represents the second quartile (median), the box represents the interquartile range, the ends represent the minimum and maximum non-outlier values, and small circles represent outliers. To organize the plots, I used a table with statistical tools to group components with values within the same range into a single graph.

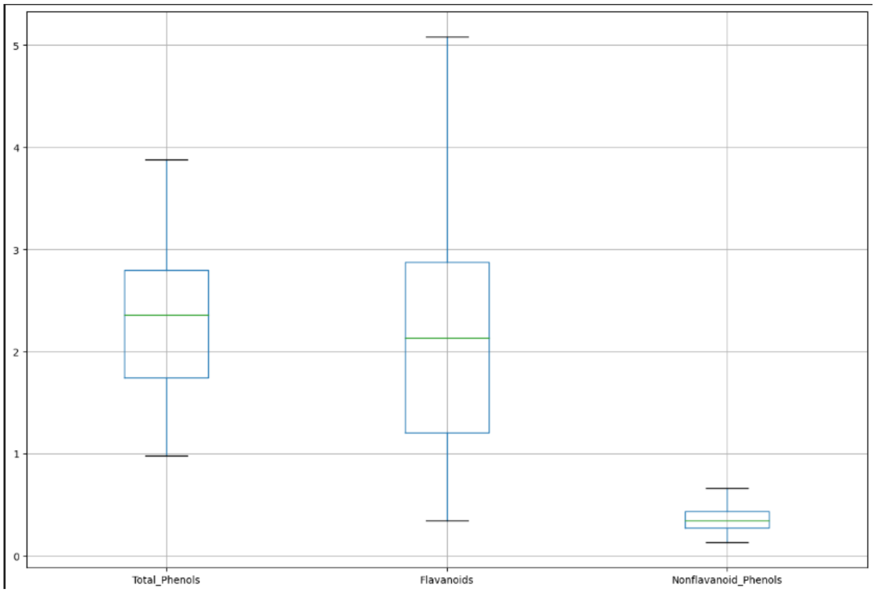


	Alcohol	Malic_Acid	Ash	Ash_Alcanity	Magnesium	Total_Phenols	Flavanoids	Nonflavanoid_Phenols	Proanthocyanins	Color_Intensity	Hue	OD280	Proline
count	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000
mean	13.000618	2.336348	2.366517	19.494944	99.741573	2.295112	2.029270	0.361854	1.590899	5.058090	0.957449	2.611685	746.893258
std	0.811827	1.117146	0.274344	3.339564	14.282484	0.625851	0.998859	0.124453	0.572359	2.318286	0.228572	0.709990	314.907474
min	11.030000	0.740000	1.360000	10.600000	70.000000	0.980000	0.340000	0.130000	0.410000	1.280000	0.480000	1.270000	278.000000
25%	12.362500	1.602500	2.210000	17.200000	88.000000	1.742500	1.205000	0.270000	1.250000	3.220000	0.782500	1.937500	500.500000
50%	13.050000	1.865000	2.360000	19.500000	98.000000	2.355000	2.135000	0.340000	1.555000	4.690000	0.965000	2.780000	673.500000
75%	13.677500	3.082500	2.557500	21.500000	107.000000	2.800000	2.875000	0.437500	1.950000	6.200000	1.120000	3.170000	985.000000
max	14.830000	5.800000	3.230000	30.000000	162.000000	3.880000	5.080000	0.660000	3.580000	13.000000	1.710000	4.000000	1680.000000

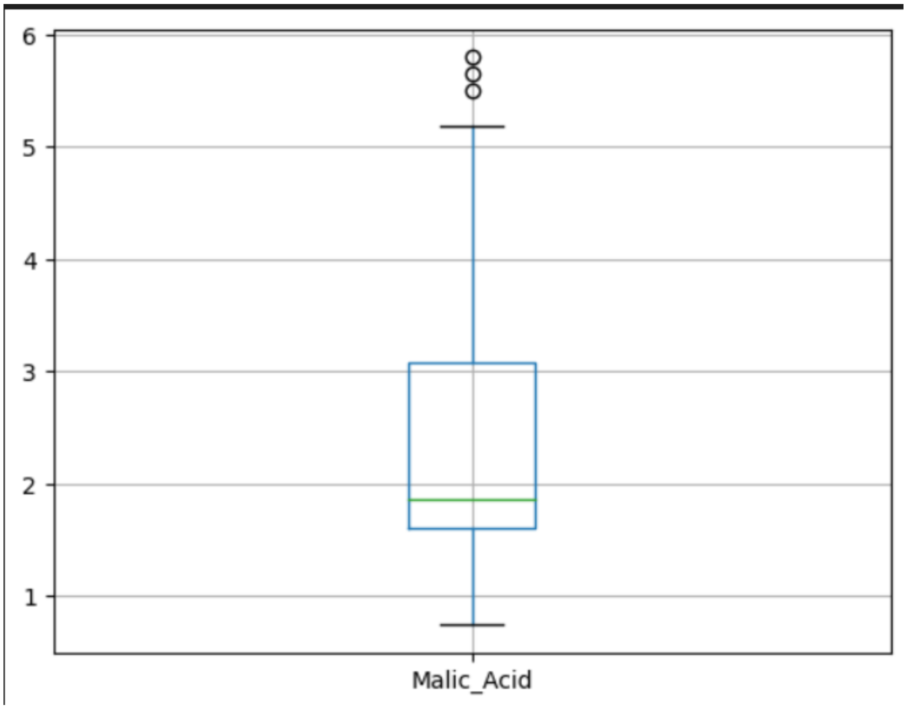
Once determined, I printed my box plots



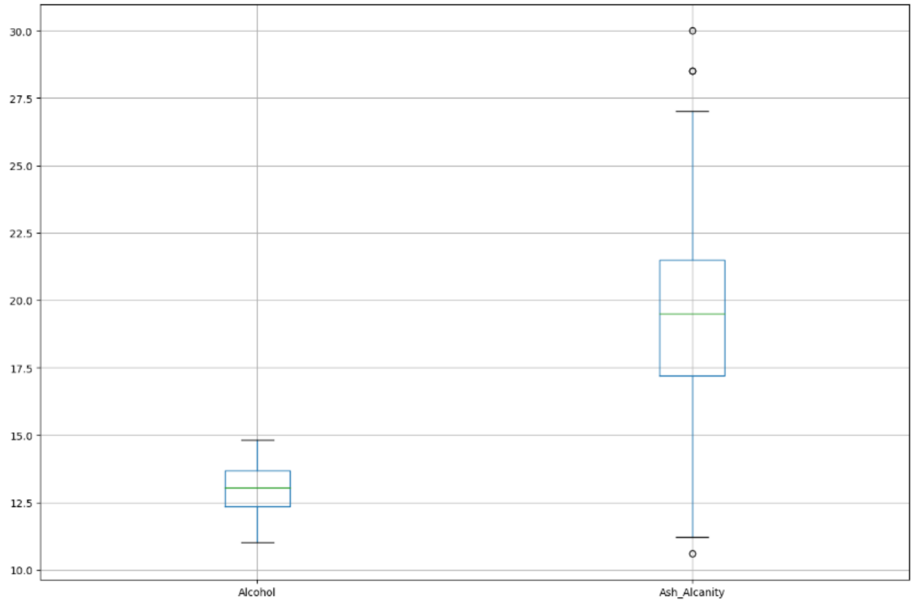
- With malic acid and proanthocyanidins, we observe that the median is very close to the first quartile and has three outliers, although this is less noticeable in the case of proanthocyanidins.
- Hue and ash seem to have an equal interquartile range for all three quartiles, except for a few outliers.
- OD280 does not have outliers and its median is close to its third quartile.



- The situation for flavonoids, non-flavonoids, and total phenols is quite similar; I may assume that they behave similarly. In all three, there is a similar distance between quartiles, and no outliers were detected.

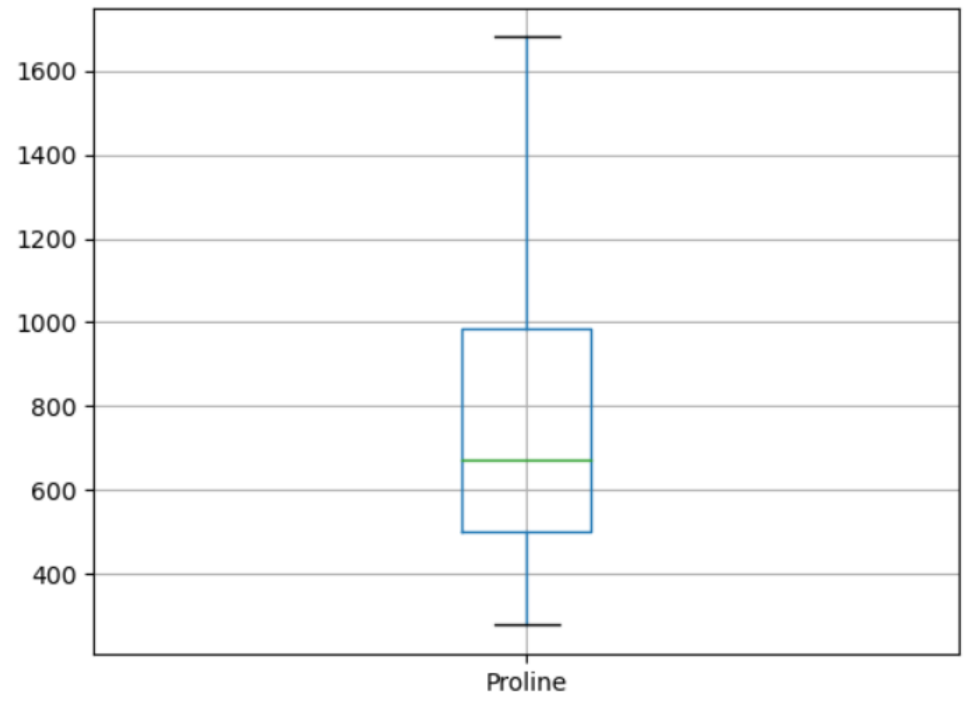


- Most of the data falls between 0.5 and 2, with few data points accumulating in the third quartile, and data points above 5 are considered outliers.

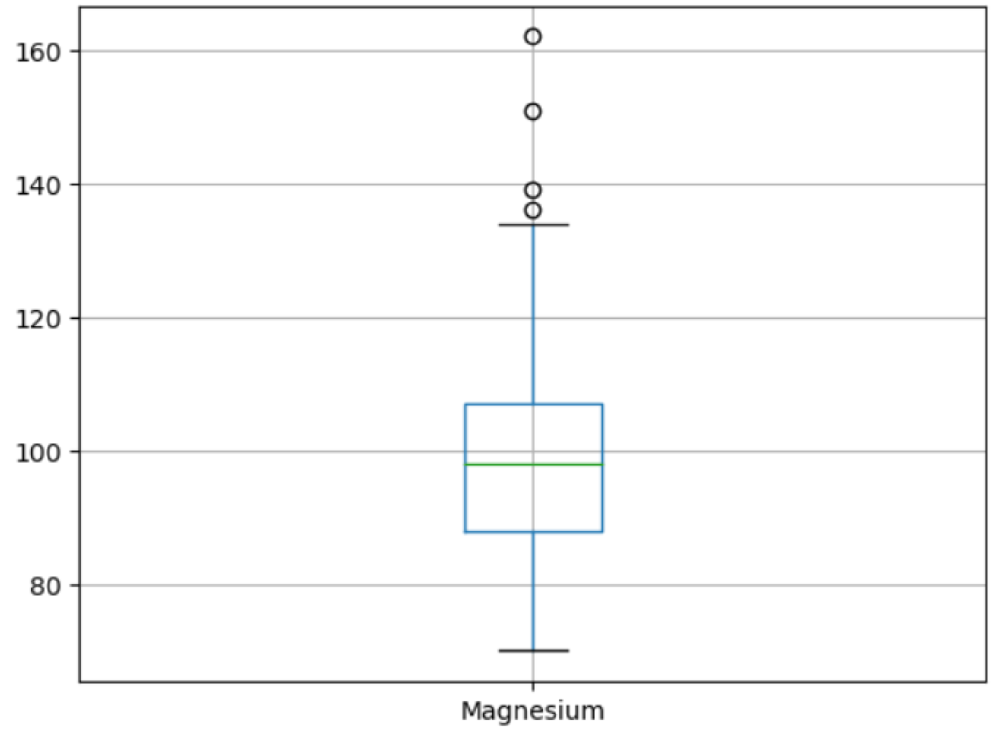


-

- The behavior of alcohol and ash alkalinity appears similar, as the interquartile range is almost equal for both cases (although the maximum and minimum of ash alkalinity deviate significantly from the quartiles), but outliers were detected in ash alkalinity.



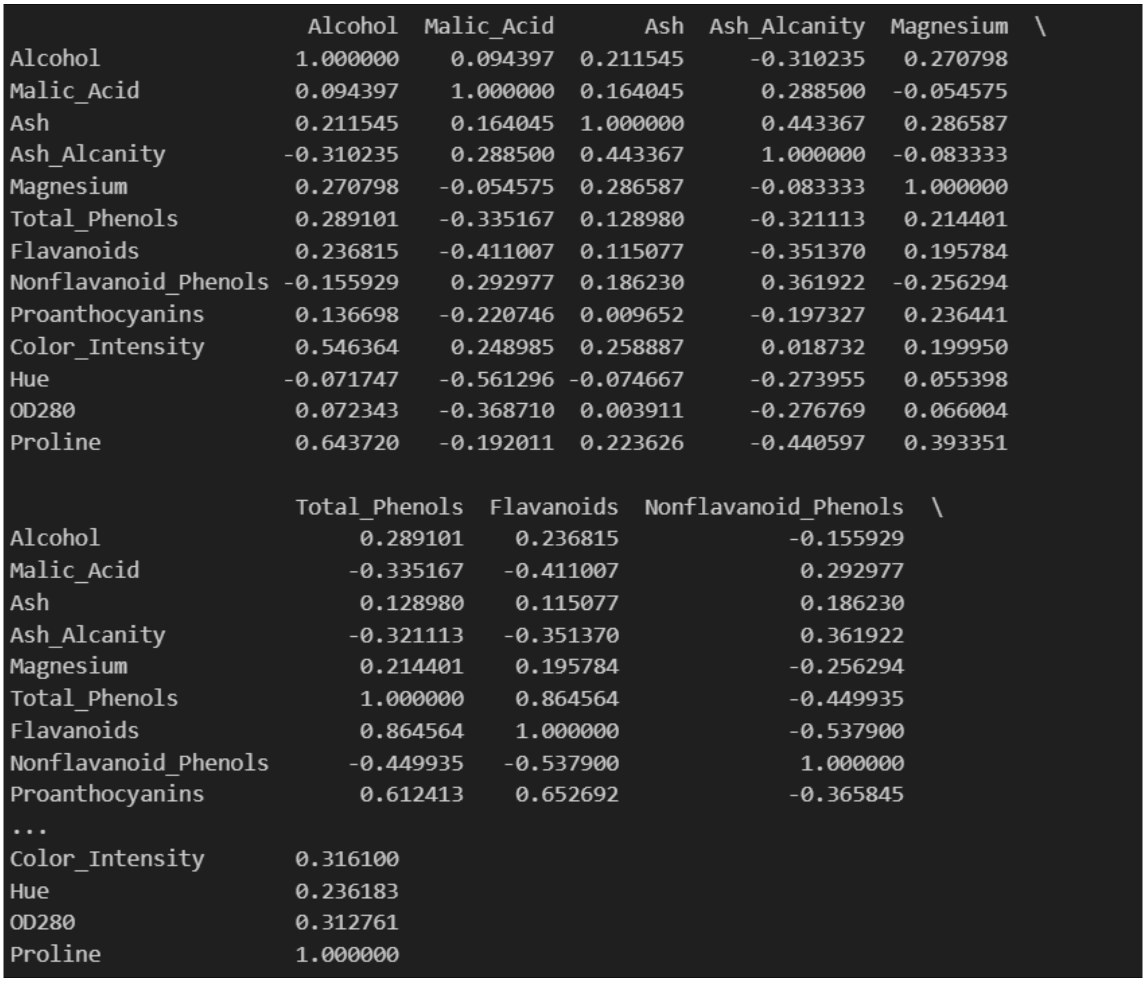
- For proline, a similar pattern is observed as with malic acid and proanthocyanidins, but no outliers are detected.



- Magnesium accumulates data around its mean and first quartile, but there is a significant distance between the third quartile and the maximum value, and some outliers were detected.

i. Correlation Matrix

The correlation matrix uses the Pearson coefficient to determine matching values between two components, i.e., whether there is a direct or inversely proportional impact between two components.



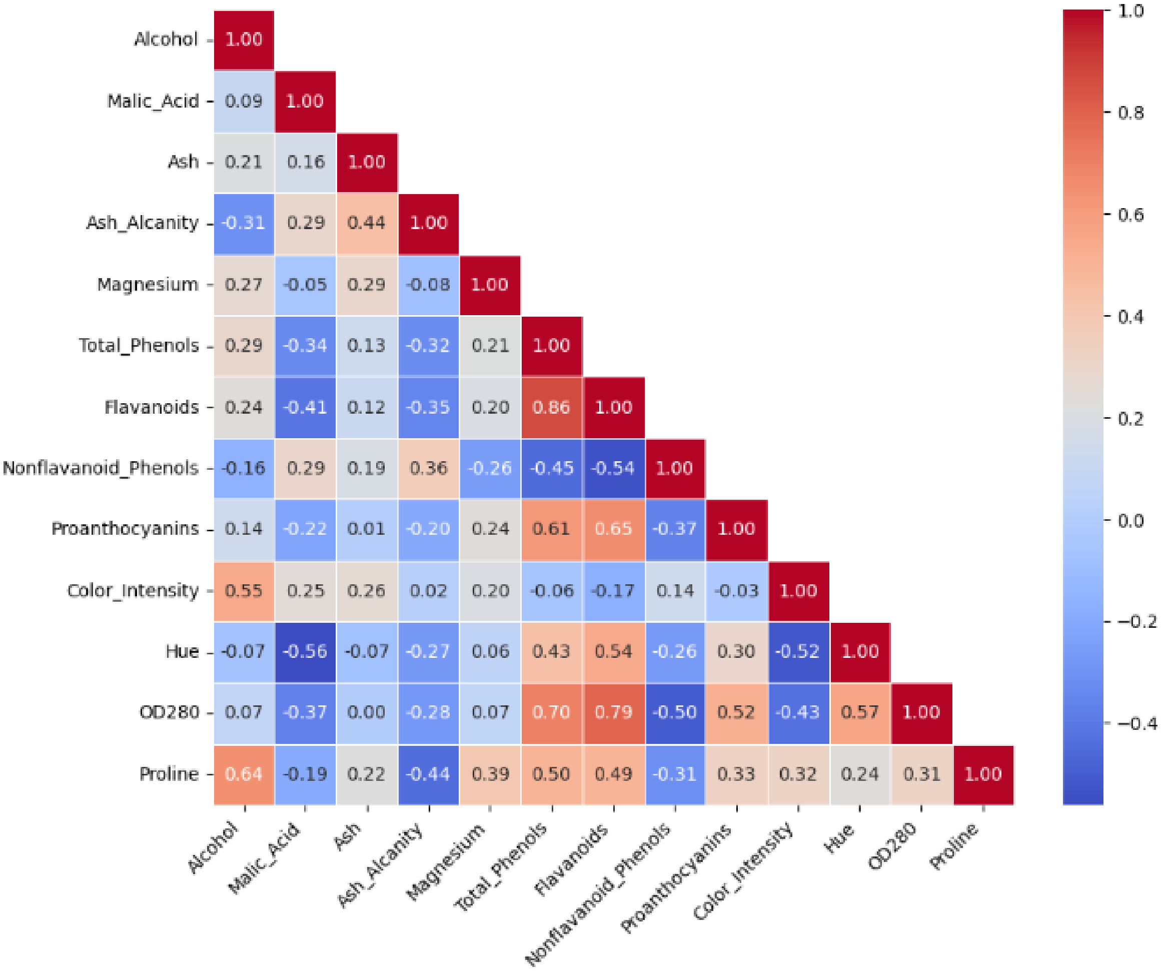
```
                       Alcohol  Malic_Acid        Ash  Ash_Alcanity  Magnesium  \
Alcohol               1.000000    0.094397   0.211545     -0.310235   0.270798
Malic_Acid            0.094397    1.000000   0.164045      0.288500  -0.054575
Ash                   0.211545    0.164045   1.000000      0.443367   0.286587
Ash_Alcanity         -0.310235    0.288500   0.443367      1.000000  -0.083333
Magnesium             0.270798   -0.054575   0.286587     -0.083333   1.000000
Total_Phenols         0.289101   -0.335167   0.128980     -0.321113   0.214401
Flavanoids            0.236815   -0.411007   0.115077     -0.351370   0.195784
Nonflavanoid_Phenols -0.155929    0.292977   0.186230      0.361922  -0.256294
Proanthocyanins       0.136698   -0.220746   0.009652     -0.197327   0.236441
Color_Intensity       0.546364    0.248985   0.258887      0.018732   0.199950
Hue                  -0.071747   -0.561296  -0.074667     -0.273955   0.055398
OD280                 0.072343   -0.368710   0.003911     -0.276769   0.066004
Proline               0.643720   -0.192011   0.223626     -0.440597   0.393351

                     Total_Phenols  Flavanoids  Nonflavanoid_Phenols  \
Alcohol                   0.289101    0.236815             -0.155929
Malic_Acid               -0.335167   -0.411007              0.292977
Ash                       0.128980    0.115077              0.186230
Ash_Alcanity             -0.321113   -0.351370              0.361922
Magnesium                 0.214401    0.195784             -0.256294
Total_Phenols             1.000000    0.864564             -0.449935
Flavanoids                0.864564    1.000000             -0.537900
Nonflavanoid_Phenols     -0.449935   -0.537900              1.000000
Proanthocyanins           0.612413    0.652692             -0.365845
...
Color_Intensity       0.316100
Hue                   0.236183
OD280                 0.312761
Proline               1.000000
```

The matrix compares each component and assigns values between -1 and 1, indicating the following:

- If the value is between 0 and 1, the direct impact intensifies as it approaches 1. That is, if we have a value close to 1, it means there is a high direct correlation between these components. Conversely, if the value approaches 0, there is a low directly proportional impact between the two components.

- Similarly, if the value is between -1 and 0, the inversely proportional impact intensifies as it approaches -1. That is, if the value corresponding to two components approaches -1, it means there is an inversely proportional relationship between one component and another. Conversely, there is little relationship as it approaches 0.

The visualization assigns shades of red and blue to the values, with intense red indicating values close to 1 and -1, and light blue indicating values close to 0.

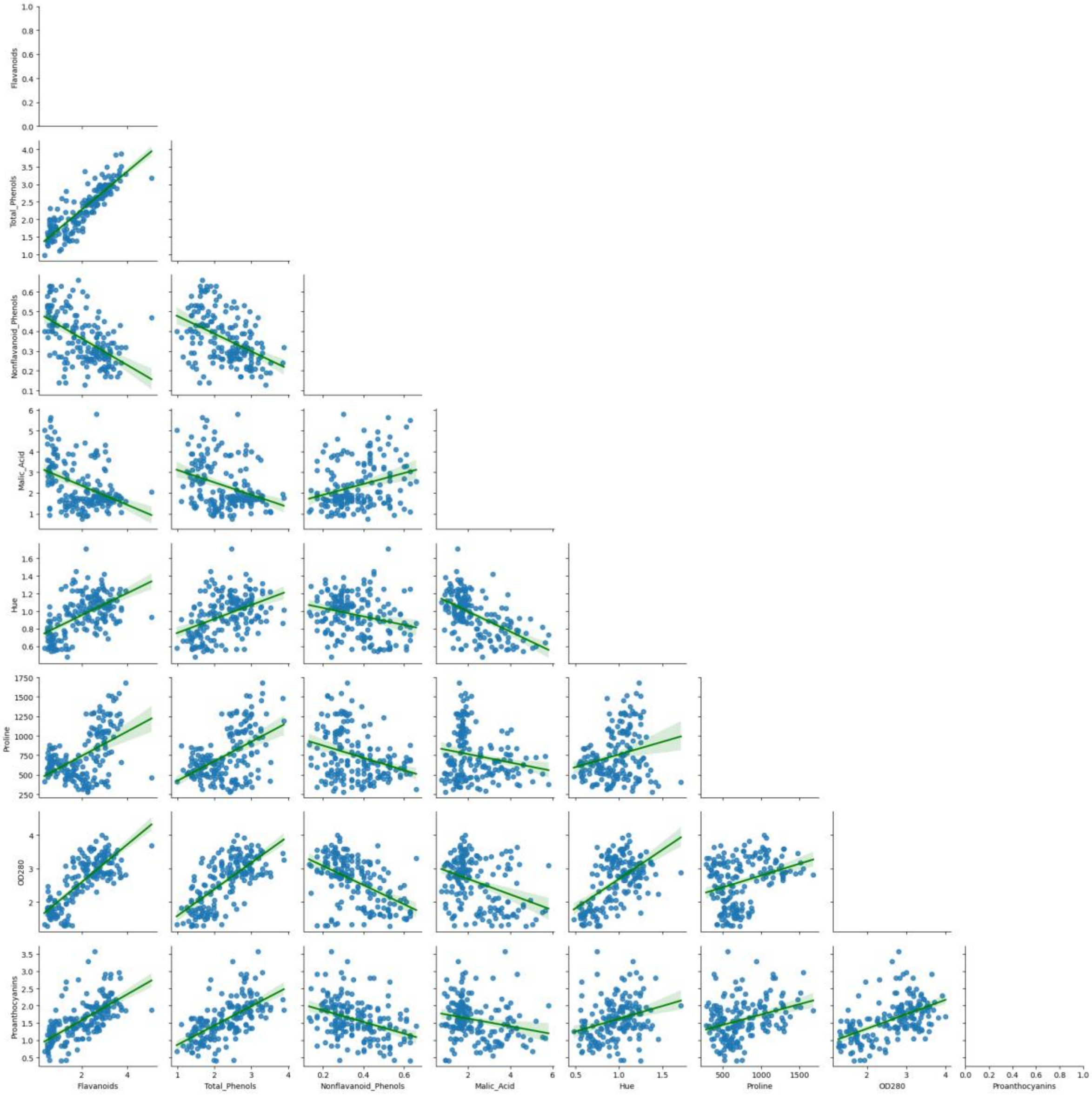


For this exercise, I decided to focus on values greater than or equal to 0.55 and less than or equal to -0.55. It is necessary to mention that flavonoid components usually do not react with non-flavonoid components. In this graph, we can observe that components with a high directly proportional impact on flavonoids also have an inversely proportional impact on non-flavonoids.

ii. Scatterplots

Last, but not least, I created scatter plots using linear regressions to determine linear correlation between two components. The proximity of points to a line with a slope of 1 or

-1 indicates how closely two components are linearly related.



We can observe that the regressions of flavonoids with proanthocyanidins, OD280, proline, and total phenols, as well as hue with malic acid, are quite clear, as the linear trend of the data is evident in all cases.

Additionally, a clustermap for all variables and a histogram per variable were included. A clustermap is a graph that classifies data according to its correlation. This graph also represents the hierarchical clustering of data by its rows and columns. A histogram visually demonstrates whether variables converge to a certain distribution, helping to determine their behavior. However, neither of these tools was decisive for the final analysis.

II. Hypothesis

Given the previous information, we can consider the relationship between flavonoids and non-flavonoids towards other compounds, but especially the relationship between flavonoids and components such as hue, OD280, proanthocyanidins, and proline as relevant. Additionally, according to the correlation matrix, we can consider the relationship between alcohol, proline, and color intensity. The latter two do not have a very significant relationship, but they will be tested to confirm (or not) the initial questions.

This is where my new questions arise: What do flavonoids react with? How much do they improve or worsen the quality of the wine? Is there any non-linear relationship with alcohol or color intensity?

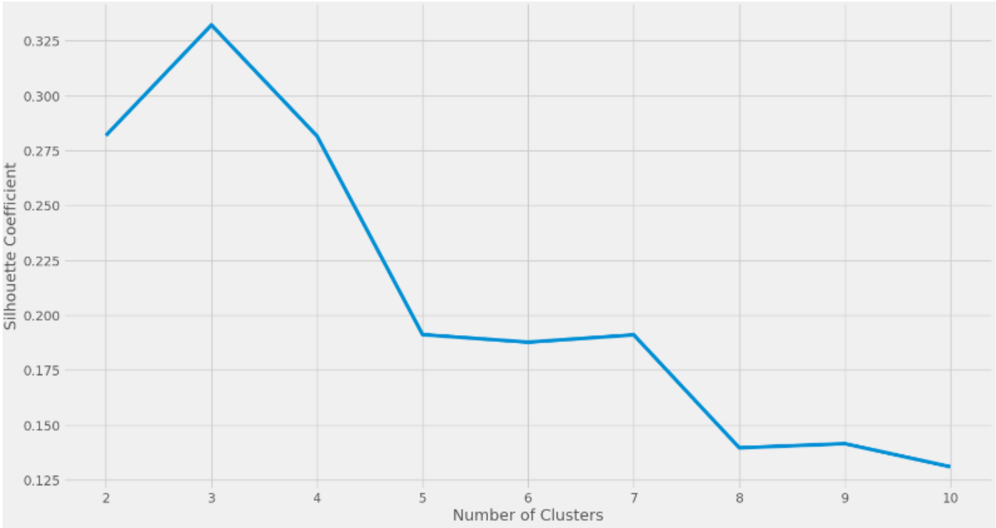
III. Clustering

It is said that there is no exact definition of what a cluster is or what clustering entails. Personally, I dare to define it as a way of classifying data with similar characteristics and behaviors into k-groups.

i. Step one: determine k using different metrics.

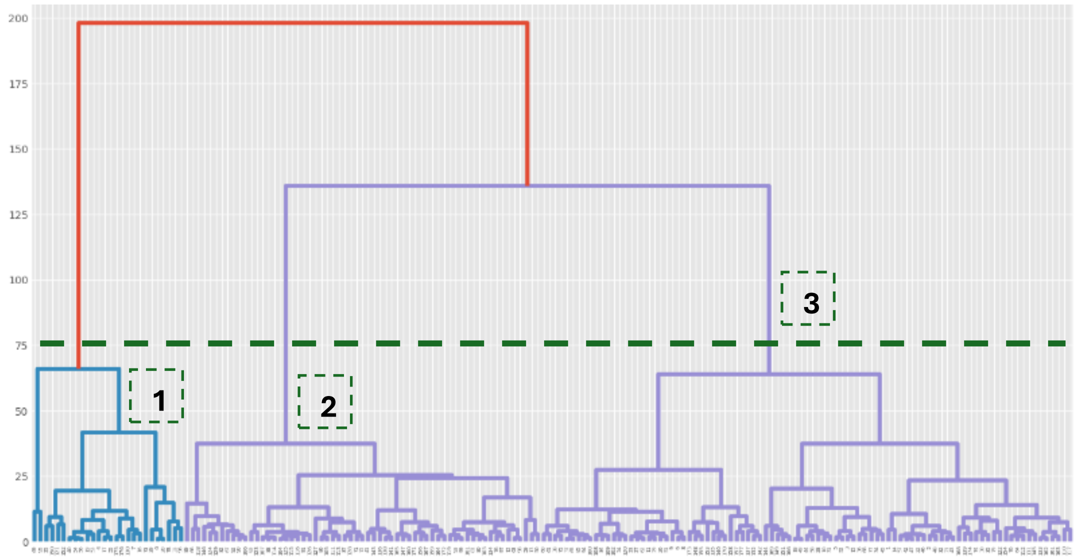
The number of clusters is represented by the letter k, where k is a strictly positive integer. There are different ways to determine the number of clusters to use, but the fact that this data collection was taken from 3 different wine cellars gives us a good clue. It was sufficient to confirm it through two methods: a dendrogram and the Silhouette coefficient.

SILHOUETTE COEFFICIENT



The Silhouette coefficient determines the certainty to classify each point into a cluster based on two things: how close the point is to others in the cluster and how far it is from points in other clusters. The coefficient ranges between -1 and 1, but the following graph shows the average of the coefficients for all points. The global maximum point on the graph determines the optimal number of clusters for the dataset. This graph aligns with our thinking for the number of clusters, as the maximum point is at x=3.

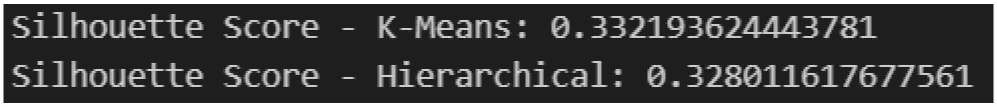
DENDROGRAM



A dendrogram is a visual tool that organizes data into subcategories, which are successively divided into two until reaching the desired level. It is a tree-like diagram that uses a "top-down" model. To determine the optimal number of clusters, we must locate the longest distance from a subdivision to the first division and draw a line across the cluster. The quantity k will be equal to the number of lines below the drawn line. Once again, it aligns with the number of wine cellars recorded in the database.

ii. Step two: work with clustering with scaled data.

Once the number of clusters was determined, I conducted a test with the Silhouette coefficient to determine whether it was more suitable to perform clustering with KMeans or hierarchical agglomerative clustering. It's important to clarify that this coefficient should not be the sole metric for decision-making, but it serves as a good guide.

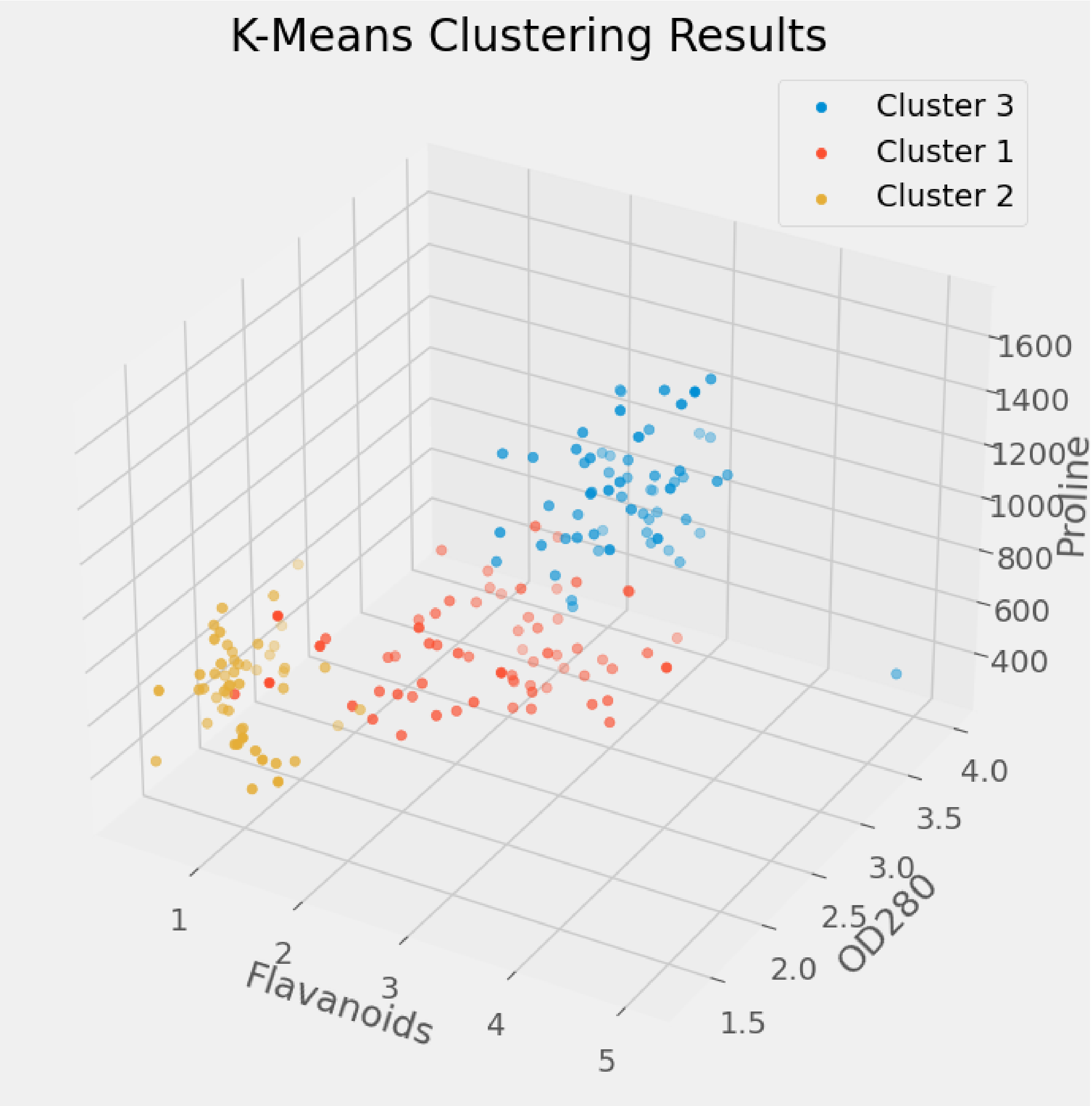


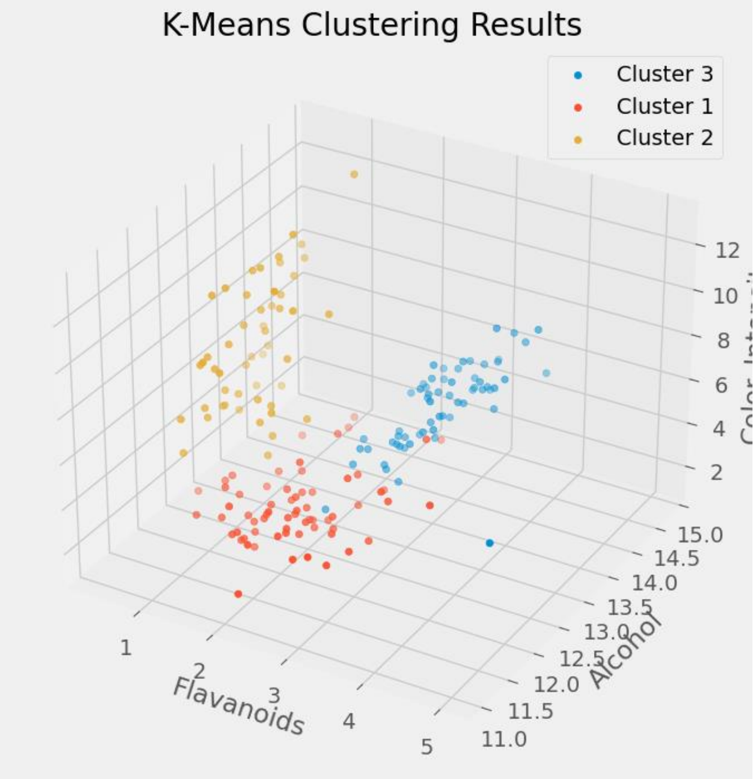
```
Silhouette Score - K-Means: 0.332193624443781
Silhouette Score - Hierarchical: 0.328011617677561
```

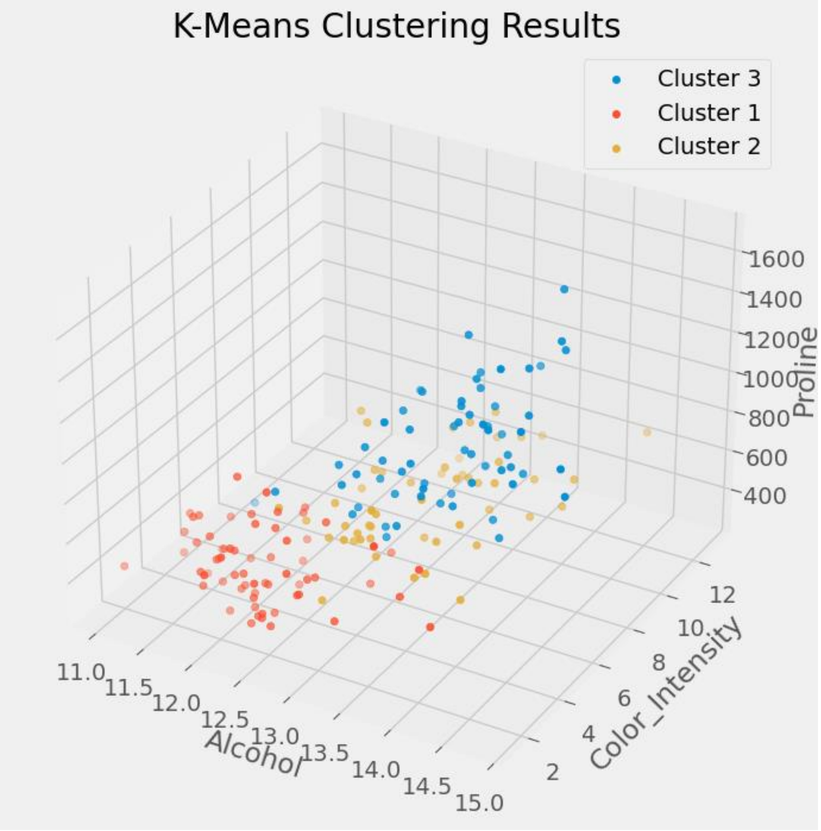
The Silhouette coefficient is higher in absolute value. Additionally, after reviewing the dendrogram and conducting some tests with the data, I concluded that it would be best to use KMeans, as it is more suitable for the nature of the data.

iii. Step three: play with graphics.

It is important to note that, despite our efforts, real-world data is not perfect, and finding clusters that do not overlap at all is considered quite challenging. However, I believe we have reached a good approximation, where the compounds that best fit the model were, as expected, flavonoids with proline and OD280.




K-Means Clustering Results
Cluster 3
Cluster 1
Cluster 2
12
10
8
6
4
2
Color_Intensi
15.0
14.5
14.0
13.5
13.0
12.5
12.0
11.5
11.0
Alcohol
1
2
3
4
5
Flavanoids


K-Means Clustering Results
Cluster 3
Cluster 1
Cluster 2
1600
1400
1200
1000
800
600
400
Proline
12
10
8
6
4
2
Color_Intensity
11.0
11.5
12.0
12.5
13.0
13.5
14.0
14.5
15.0
Alcohol

IV. Conclusions

Flavonoids exhibit a high reactivity with the total phenols and, consequently, with non-flavonoids (since a compound cannot simultaneously be a flavonoid and a non-flavonoid). Flavonoids proved to be a significantly relevant variable in the data, as various components were found to be related to them.

Flavonoids are compounds with high antioxidant and anti-inflammatory capacities. Additionally, they have a significant potential to prevent chronic cardiovascular diseases and cancer. They are primarily found in fruits and vegetables. According to the presented data, grapes are one of the fruits with a high concentration of flavonoids.

Based on the analysis, the compounds OD280, hue, proanthocyanidins, proline, and phenols are highly correlated with flavonoids. In other words, these are compounds that react with flavonoids or promote the presence of flavonoids. This implies that the higher the concentration of these components in the wine, the greater the benefits the beverage will offer.

Given the above, if one were to formulate a wine with enhanced properties, it would be advisable to aim for a high concentration of these compounds. On the other hand, as a consumer seeking the highest quality wine, one might prioritize wines with these properties.

Interestingly, no component has a high correlation with the alcohol content or color intensity of the wine. This leads me to believe that the fermentation process of grapes has minimal impact on the other components. I can assume that the true determinant of the wine's color intensity is the quantity of grapes integrated into the blend.

V. References

Coder, R. (2022). Pairs Plot (gráfico por pares) en Seaborn con la función Pairplot. PYTHON CHARTS | Visualización de datos con Python. https://python-charts.com/es/correlacion/pairplot-seaborn/#diagonal. Consultado el 31 de enero de 2024.

Na, & Na. (2020). K-Means con Python paso a paso | Aprende Machine Learning. Aprende Machine Learning. https://www.aprendemachinelearning.com/k-means-en-python-paso-a-paso/. Consultado el 31 de enero de 2024.

Python, R. (2023). K-Means Clustering in Python: A Practical Guide. https://realpython.com/k-means-clustering-python/. Consultado el 31 de enero de 2024.

Sklearn.Cluster.AgglomerativeClustering. (s. f.). scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html. Consultado el 31 de enero de 2024.