# Análisis de componentes de los vinos

El vino es una bebida alcohólica con la que la mayoría estamos familiarizados, pues a todos nos ha acompañado en un momento especial, desde la cena de un día cualquiera después de una larga jornada laboral, hasta el brindis de celebración de tu matrimonio. Sabemos que el vino es una bebida muy versátil y elegante, pero no creo la fórmula secreta sea de dominio público. Tuve la oportunidad de hacer un profundo análisis de los 13 componentes de 177 vinos en tres viñedos y me gustaría compartir mis descubrimientos.

Inicié con preguntas sobre el archivo basándome únicamente en mis conocimientos sobre el vino. ¿Qué compuestos determinan la cantidad de alcohol que tendrá la bebida? ¿Siquiera habrá reacción de algún compuesto con el alcohol? ¿De qué depende la intensidad del color? ¿Por qué se dice que el vino ayuda al corazón?

## I. Análisis Exploratorio

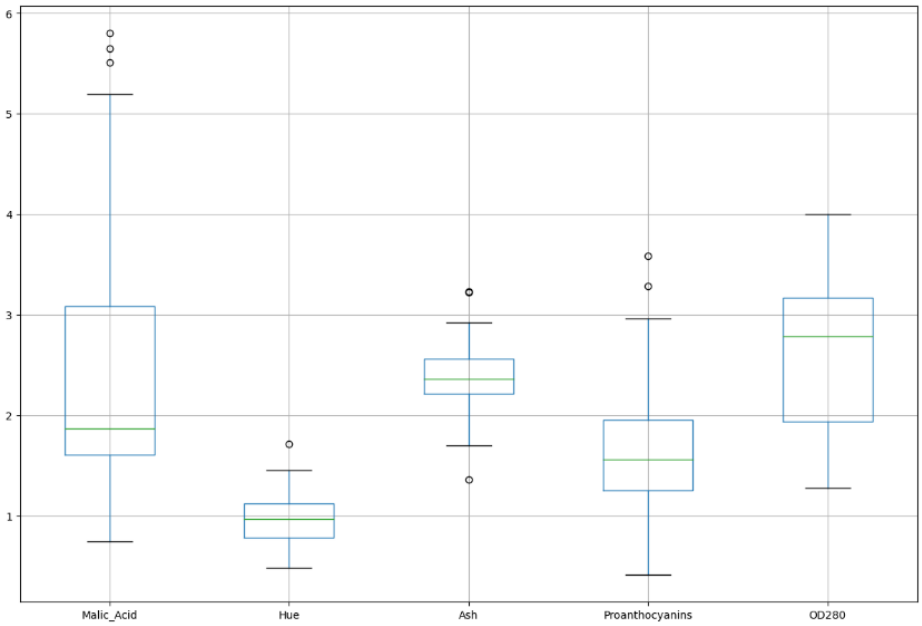
Debido a mi escaso conocimiento en vinos, fue importante realizar un análisis exploratorio antes de fijar una hipótesis principal. Para ello, utilicé distintas herramientas visuales de Python como los gráficos de caja, tablas de dispersión y gráfica de coeficientes de Pearson.

### i. Gráficos de Caja

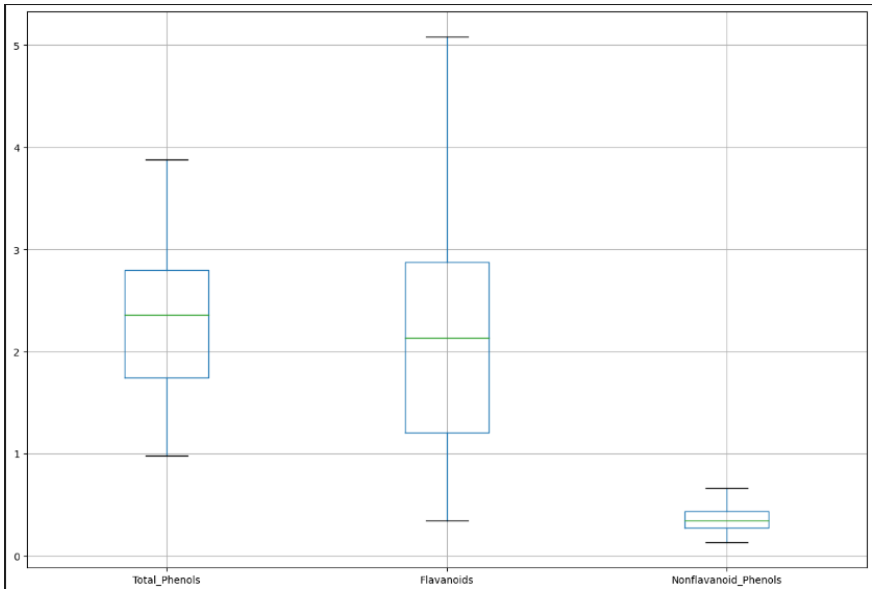
Los gráficos de caja visualizan los datos con respecto a sus cuartiles. La raya verde representa el cuartil 2 (mediana), el cuadrado (o caja) representa el rango intercuartílico, los extremos representan el mínimo y máximo valor no atípico, y los pequeños círculos representan los valores atípicos, también conocidos como *outliers*. Para acomodar los gráficos, utilicé una tabla con herramientas estadísticas para agrupar en un mismo grafo os componentes con valores dentro del mismo rango.

| | 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 |

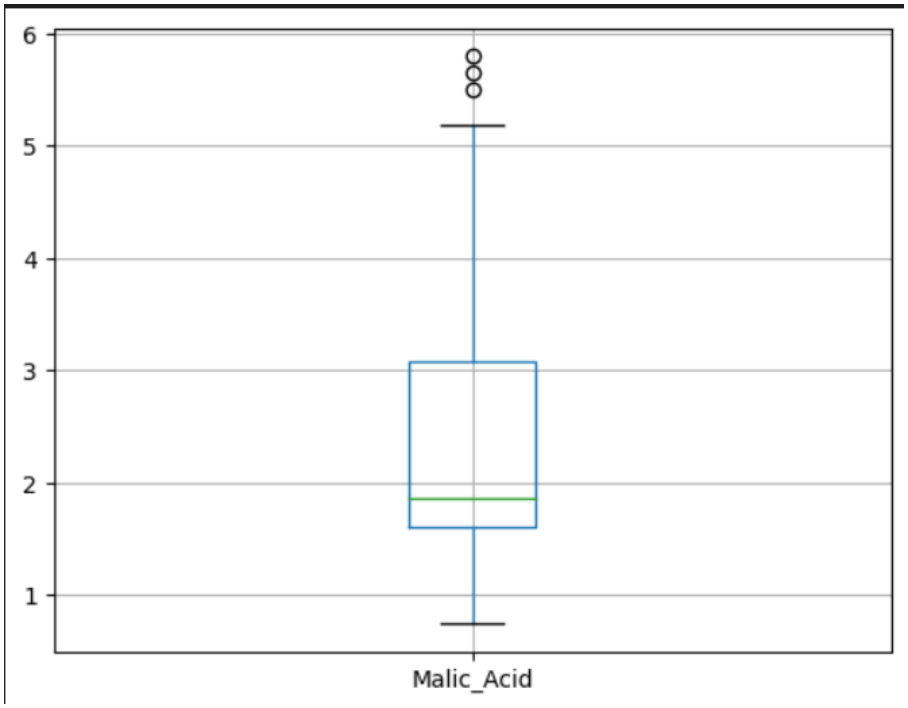
Una vez determinado, imprimí los gráficos de caja.



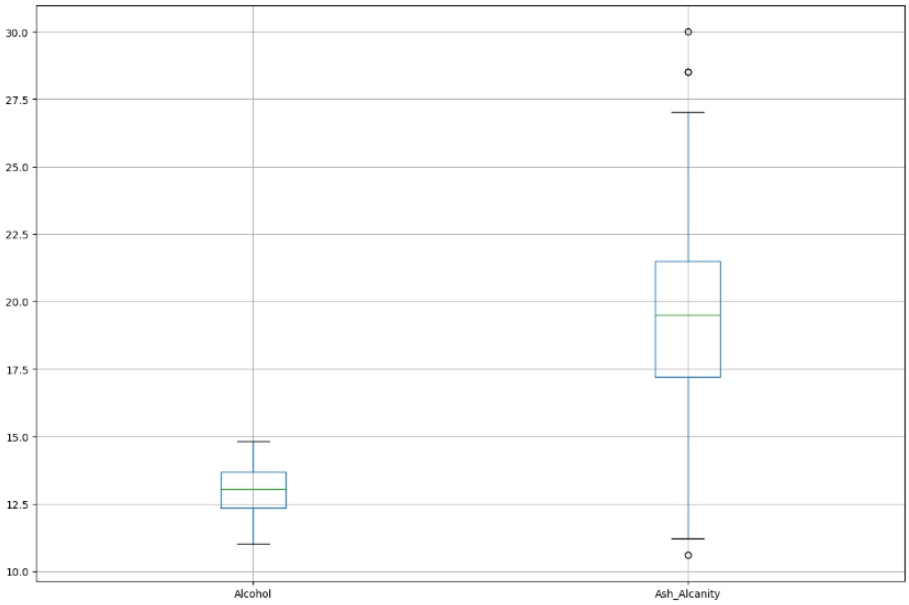

- Con el ácido málico y las proantocianidas vemos que la mediana se acumula muy cerca de su primer cuartil y posee tres *outliers*, aunque esto es menos notorio para el caso de las proantocianidas.
- Hue y las cenizas parecen tener una distancia intercuartílica igual para sus tres cuartiles, a excepción de pocos *outliers*.
- OD280 no reconoce *outliers* y su mediana se acerca a su 3er cuartil.



- El caso para los flavonoides, no flavonoides y fenoles totales es bastante similar, puedo suponer que se comportan similar. En los tres se encuentran una distancia similar entre cuartiles y no se detectaron *outliers*.

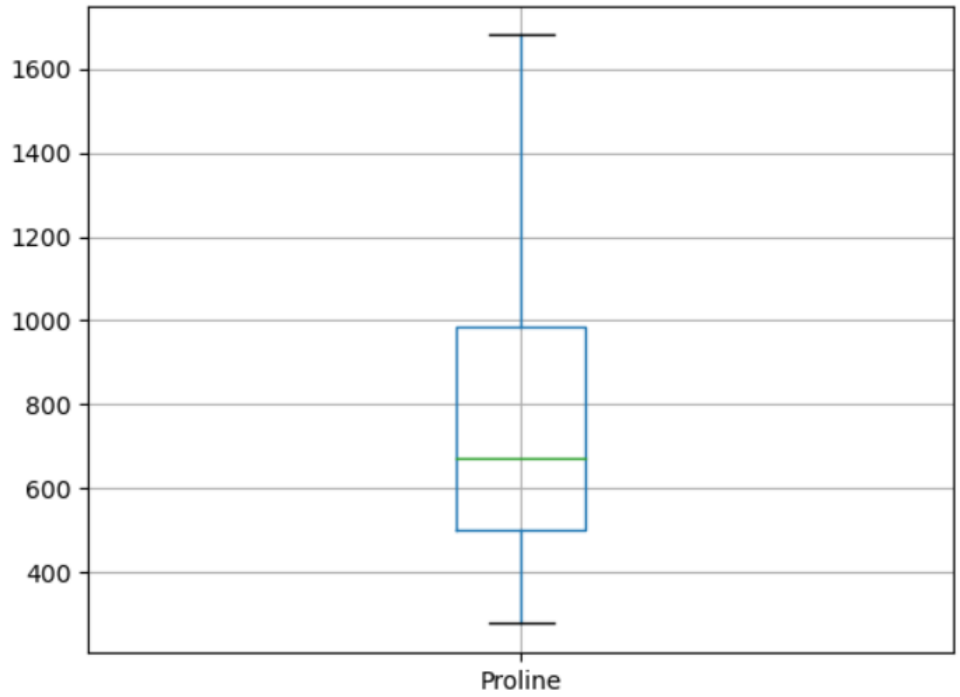



- La mayoría de los datos se encuentran entre 0.5 y 2, pocos datos se acumulan en el tercer cuartil y datos por encima de 5 se consideran *outliers*.

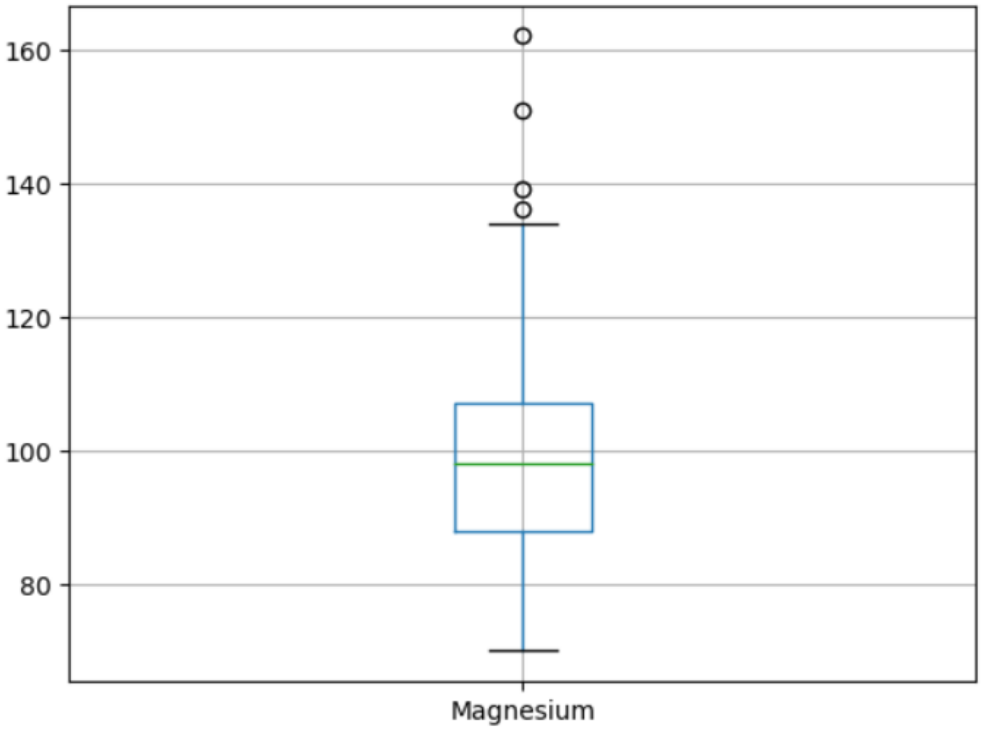



- El comportamiento del alcohol y alcalinidad de las cenizas se ve similar, pues la distancia intercuartílica es casi igual para ambos casos (aunque el máximo y

mínimo de alcalinidad de las cenizas se distancia mucho de los cuartiles), pero se detectaron *outliers* en alcalinidad de las cenizas.




- Para la prolina se observa un caso similar que el ácido málico y las Proantocianidas, pero no se detectan *outliers*.



- El magnesio acumula datos en su media y primer cuartil, pero existe una distancia significativa entre tercer cuartil y el valor máximo, además de que se detectaron algunos *outliers*.

Como podemos notar, en la base se encuentran datos atípicos para casi la mitad de los componentes. Sin embargo, no eliminé ninguno, pues no conozco la importancia de dichos datos específicos y eliminarlos puede afectar demasiado al presentar resultados.

ii. Matriz de correlación

La matriz de correlación utiliza el coeficiente de Pierson para determinar valores coincidentes entre dos componentes, es decir, si existe un impacto directa o inversamente proporcional entre dos componentes.

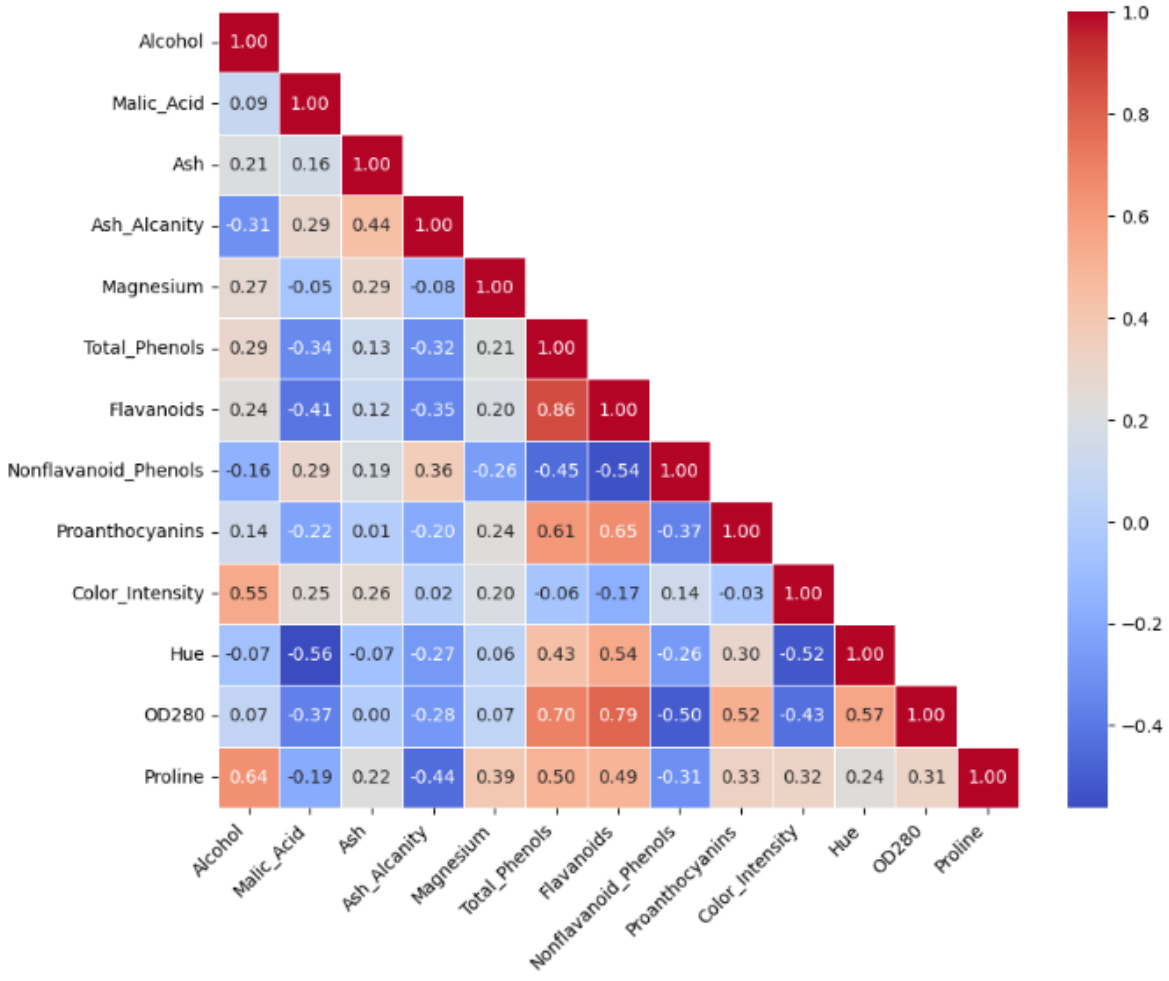
| | Alcohol | Malic_Acid | Ash | Ash_Alcanity | Magnesium | \ |
|---|---|---|---|---|---|---|
| Alcohol | 1.000000 | 0.094397 | 0.211545 | -0.310235 | 0.270798 | |
| Malic_Acid | 0.094397 | 1.000000 | 0.164045 | 0.288500 | -0.054575 | |
| Ash | 0.211545 | 0.164045 | 1.000000 | 0.443367 | 0.286587 | |
| 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 | | | |

La matriz compara cada componente y otorga valores entre -1 y 1 indicando lo siguiente:

- Si el valor se encuentra entre 0 y 1, el impacto directo se intensifica al acercarse a 1. Es decir, si tenemos un valor cercano al 1, significa que existe una correlación directa alta entre ambos componentes estos. Por el contrario, si el valor se acerca al 0, existe un bajo impacto directamente proporcional entre ambos componentes.
- De la misma manera, si el valor se encuentra entre -1 y 0, el impacto inversamente proporcional se intensifica al acercarse a -1. Es decir, si el valor correspondiente a dos componentes se acerca al -1, significa que existe una relación inversamente proporcional de un componente a otro. Por el contrario, existe poca relación al acercarse a 0.

La siguiente relaciona todos de rojo y azul a los valores, siendo un rojo intenso los valores cercanos a 1 y -1, y azul claro los valores cerca del 0.

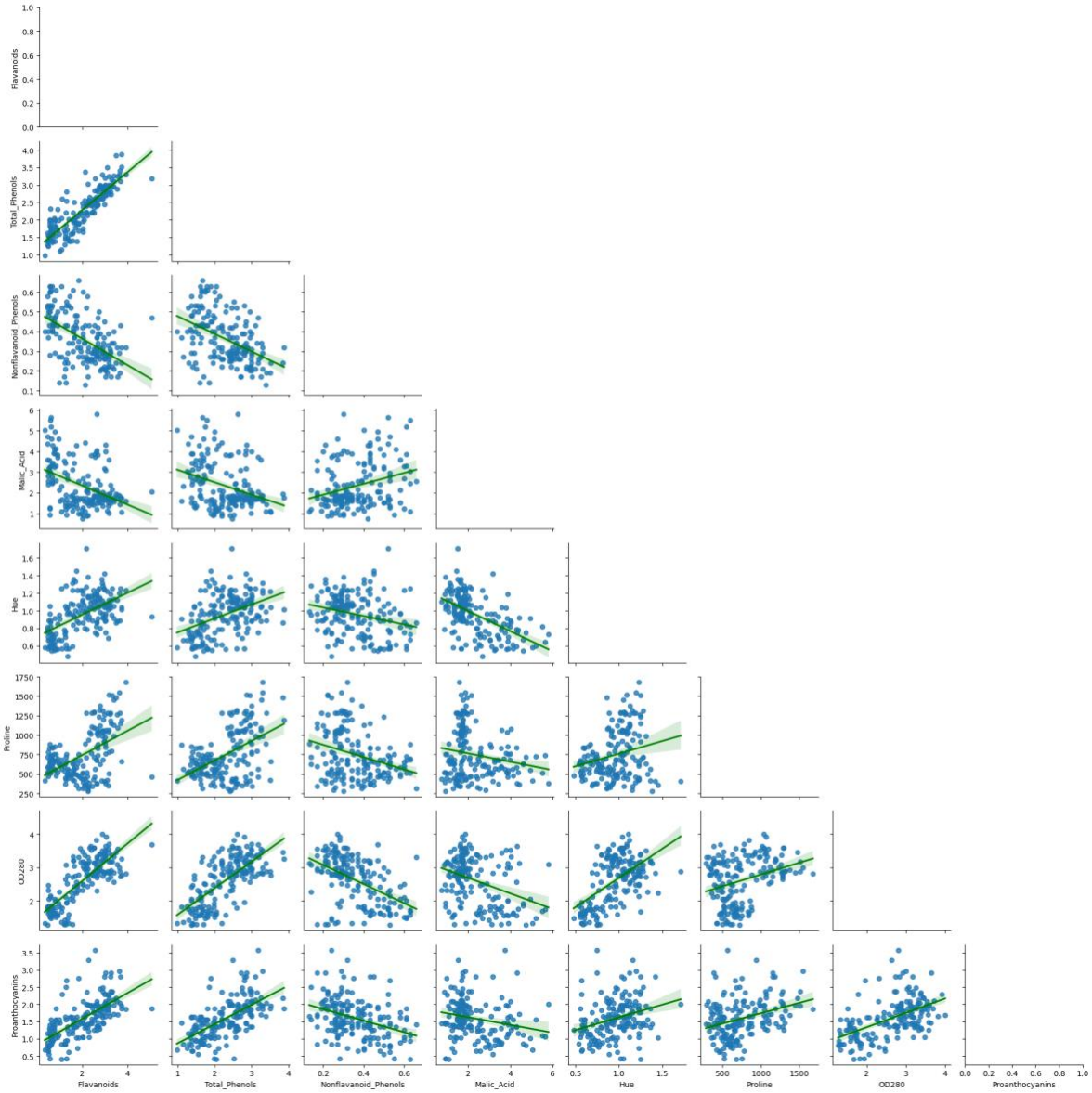



Para este ejercicio, decidí partir de los valores mayores o iguales a 0.55 y menores o iguales a -0.55.

Es necesario mencionar que los componentes flavonoides no suelen reaccionar con componentes no flavonoides. En esta gráfica podemos notar que aquellos componentes que tienen alto impacto directamente proporcional con los flavonoides tienen al mismo tiempo impacto inversamente proporcional con los no flavonoides.

iii. Gráficos de dispersión

Por último, pero no menos importante, realicé gráficos de dispersión que utilizan regresiones lineales para determinar correlación lineal entre dos componentes. La cercanía de los puntos a una recta que con pendiente 1 o -1 nos indica lo cerca que están dos componentes de relacionarse linealmente. Por otro lado, el histograma nos indica la distribución estadística a la que convergen los datos. Estas gráficas se realizaron con base en los resultados de la gráfica de coeficientes de Pearson.



Podemos notar que las regresiones de los flavonoides con las proantocianidas, el OD280, la prolina y los fenoles totales, así como el hue con el ácido málico son bastante claras, pues en todos los casos se nota la tendencia lineal de los datos.

Además, se incluyó un *clustermap* para todas las variables y un histograma por variable. Un *clustermap* es una gráfica que clasifica los datos según su correlación. Esta gráfica también representa la agrupación jerárquica de los datos por sus filas y columnas. Un histograma demuestra visualmente si las variables convergen a cierta distribución, esto sirve para determinar su comportamiento. Sin embargo, ninguno de las dos herramientas fue determinante para el análisis final.

## II. Hipótesis

Dada la información previa, podemos considerar relevante la relación de los flavonoides y no flavonoides hacia otros compuestos, pero especialmente la relación entre los flavonoides con los componentes hue, OD280, proantocianidas y prolina. Además, acorde a la matriz de correlación, se puede considerar la relación del Alcohol con la prolina y la intensidad el color. Estos últimos no poseen una relación tan significativa, pero se probarán para descartar (o no) las preguntas iniciales.

Es aquí cuando mis nuevas preguntas surgen: ¿Con qué reaccionan los flavonoides? ¿Qué tanto mejoran o empeoran la calidad del vino? ¿Tendrá alguna relación no lineal con el alcohol o la intensidad del color?

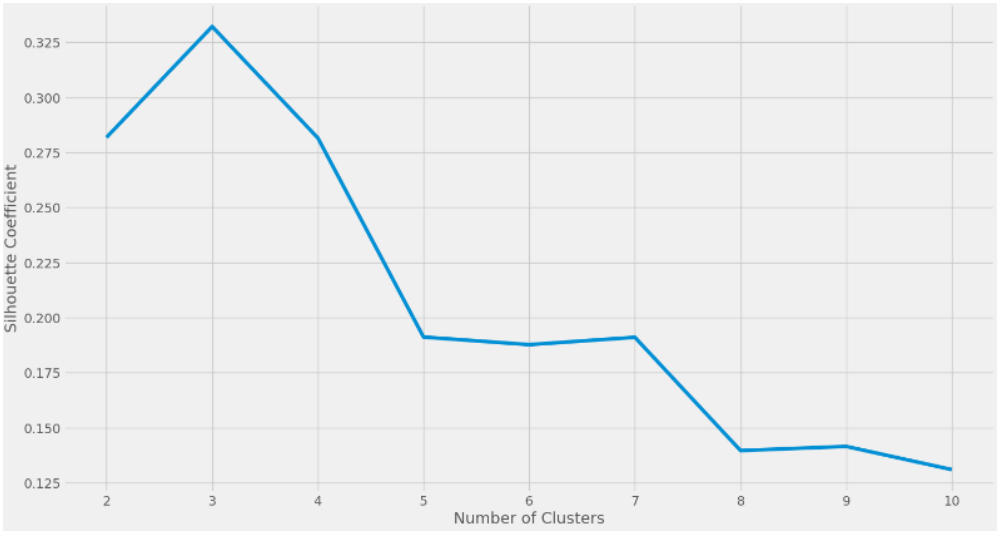
## III. Clustering

Se dice que no existe una definición exacta de lo que es un *cluster* o de lo que es hacer *clustering*. Personalmente, me atrevo a definirlo como una manera de clasificar datos con características y comportamientos similares en k-grupos.

i. Paso uno: determinar k utilizando distintas métricas.

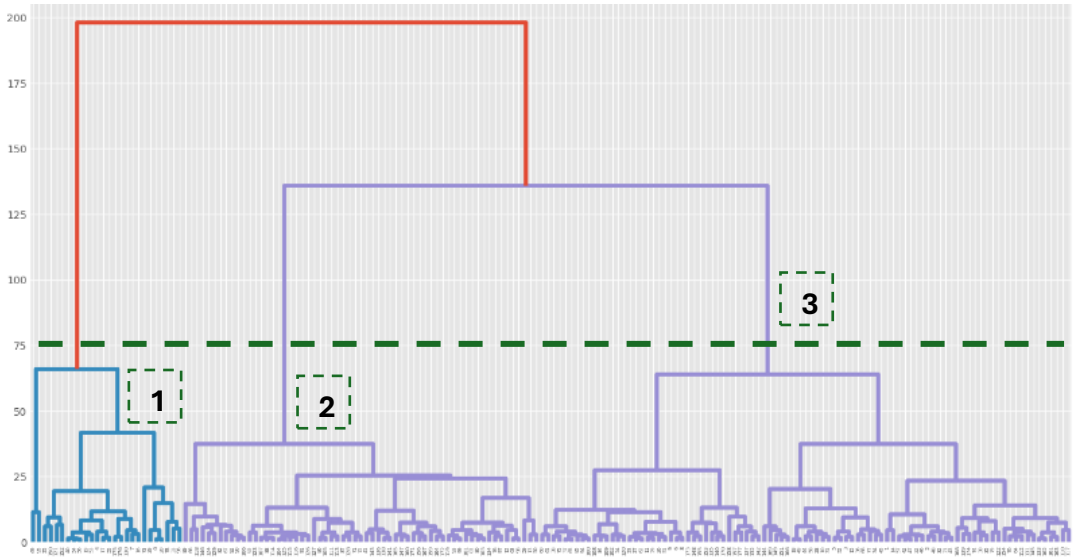
El número de *clusters* se representa con la letra k, siendo k un número entero estrictamente positivo. Hay distintas maneras de determinar la cantidad de *clusters* a usar, pero el hecho de que esta recolección de datos se tomó de 3 cavas de vinos distintas nos da una buena pista. Bastó con confirmarlo a través de dos métodos: un dendrograma y el coeficiente de Silhouette.

COEFICIENTE DE SILLOUHETTE




El coeficiente de Silhouette determina la certeza para clasificar en un *cluster* cada punto basándose en dos cosas: qué tan cerca está el punto de otros en el cluster y qué tan lejos está de los puntos en otros clusters. El coeficiente se encuentra entre -1 y 1, pero la siguiente gráfica expone el promedio de los coeficientes para todos los puntos. El punto máximo global de la gráfica determina la cantidad óptima de *clusters* para el conjunto de datos. Esta gráfica coincide con nuestro pensar para el número de *clusters*, pues el punto máximo se encuentra en x=3.

DENDROGRAMA




En dendrograma es una herramienta visual que organiza los datos en subcategorías, las cuales se dividen de dos en dos hasta llegar al nivel deseado. Es un diagrama en forma de árbol que usa un modelo “de arriba para abajo”. Para determinar el óptimo de clusters, debemos localizar la distancia más larga de una subdivisión a la primera división y trazar

una raya sobre el cluster. La cantidad k será igual a la cantidad de rayas bajo la línea trazada. Una vez más, coincide con la cantidad de cavas de vino que se registró en la base de datos.

ii. Paso dos: realizar el *clustering* con los datos escalados.

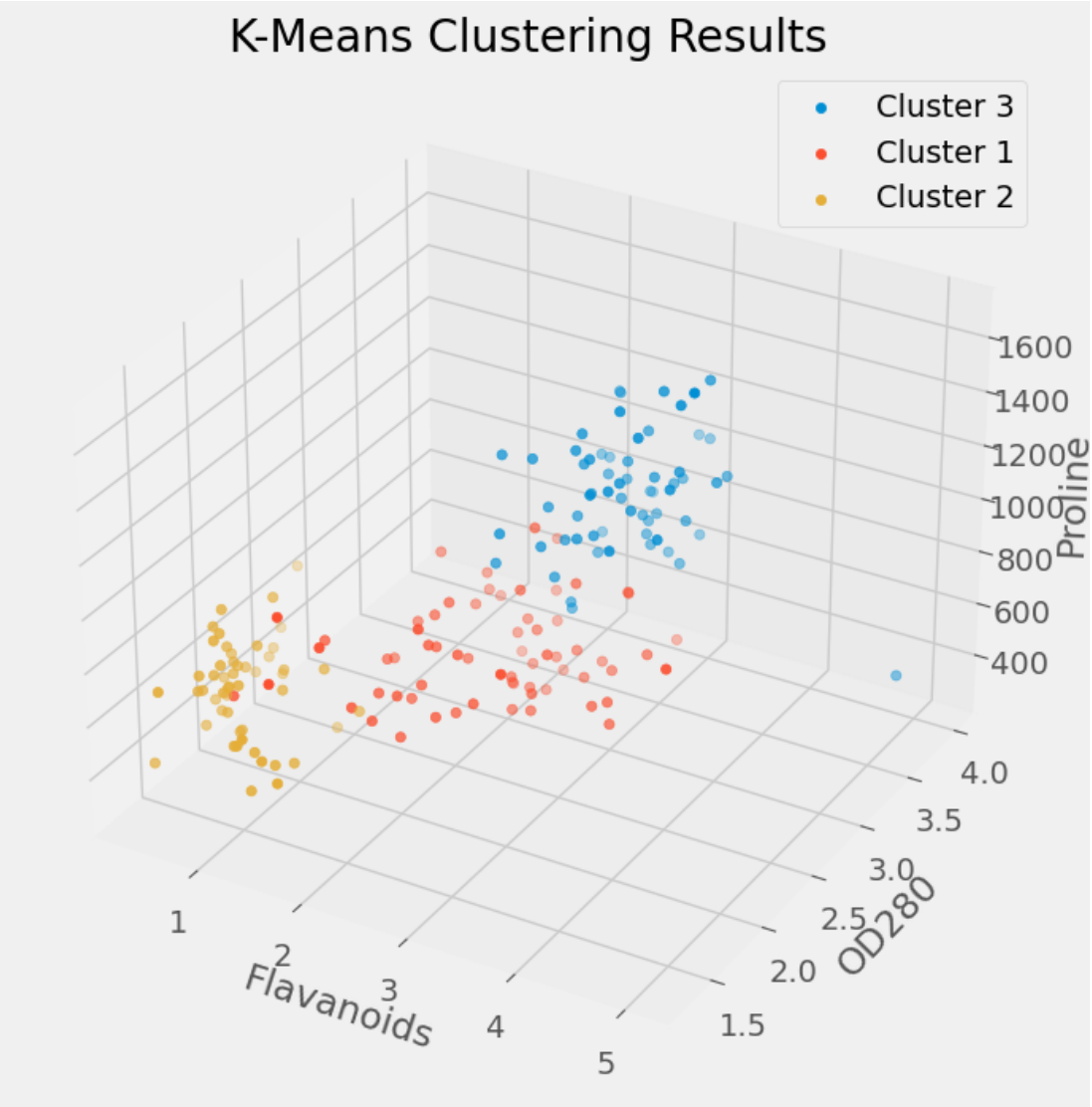
Una vez determinada la cantidad de *clusters*, realicé una prueba con el coeficiente de Silhouette para determinar si era más apto realizar un *clustering* con *KMeans* (Kmedias) O jerárquico aglomerado. Es importante aclarar que este coeficiente no debe ser la única métrica para tomar decisiones, pero es una buena guía.

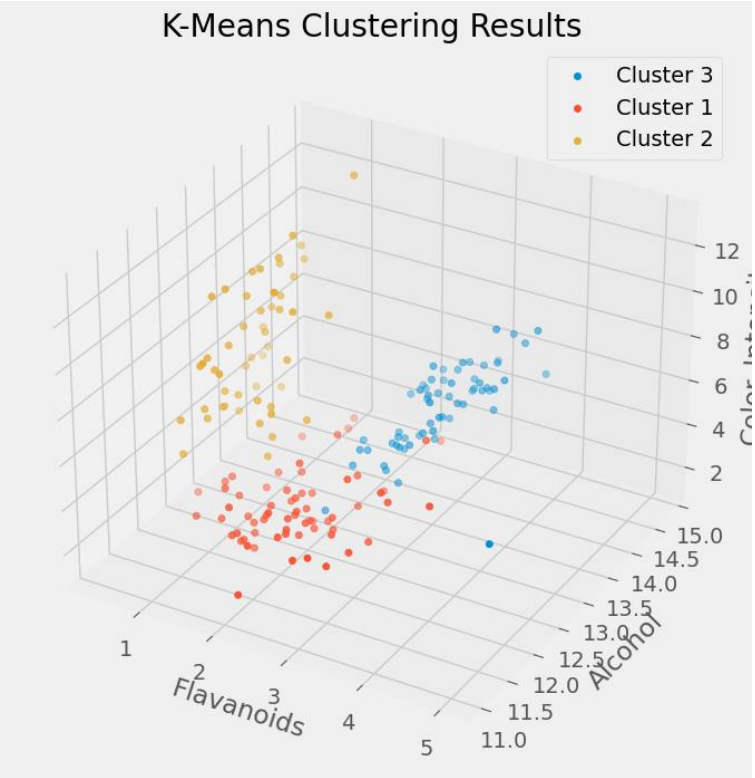
```
Silhouette Score - K-Means: 0.332193624443781
Silhouette Score - Hierarchical: 0.328011617677561
```

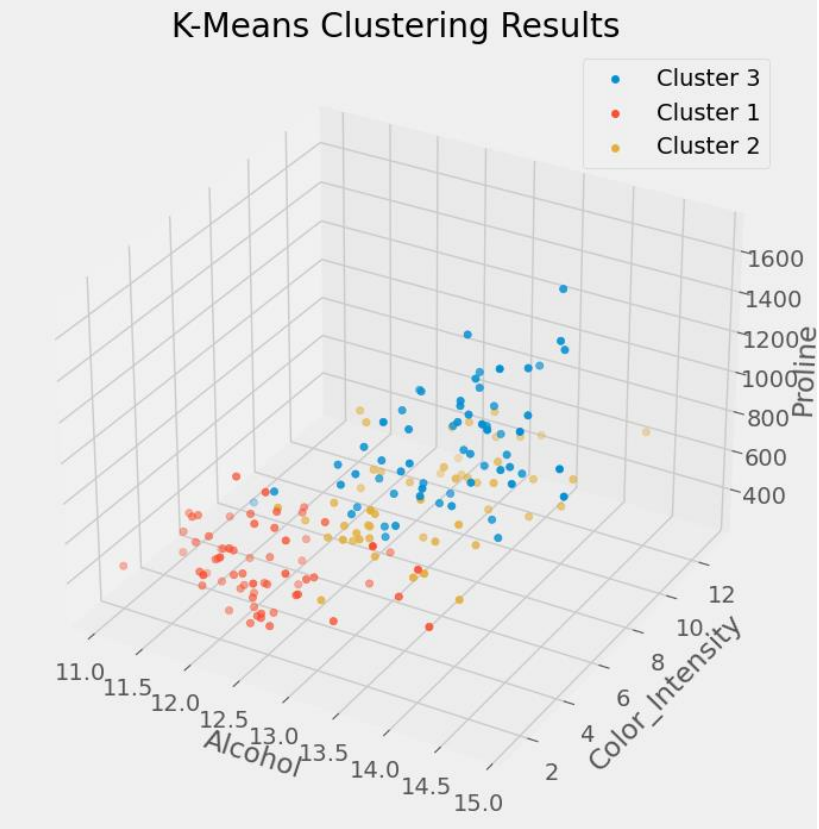
El coeficiente de Silhouette es mayor en valor absoluto. Además, después de revisar dendrograma y algunas pruebas con los datos, concluí que lo mejor sería utilizar *KMeans*, pues es más conveniente por la naturaleza de los datos.

iii. Paso tres: Jugar con las gráficas.

Debo hacer la aclaración de que, por más que intentemos, los datos en el mundo no son perfectos, y se considera bastante difícil encontrar clusters que no se empalmen en lo absoluto. Sin embargo, considero se llegó a una buena aproximación, donde los compuestos que mejor ajustaron al modelo fueron, como se esperaba, los flavonoides con la prolina y OD280.


K-Means Clustering Results
Cluster 3
Cluster 1
Cluster 2
Proline
1600
1400
1200
1000
800
600
400
OD280
4.0
3.5
3.0
2.5
2.0
1.5
Flavanoids
1
2
3
4
5




## IV. Conclusiones

Los flavonoides tienen alta reacción con el total de fenoles, y por ende con los no flavonoides, pues un flavonoide no puede ser al mismo tiempo flavonoide). Los flavonoides fueron una variable de significativa relevancia en los datos, pues diversos componentes comprobaron relacionarse con ellos.

Los flavonoides son compuestos con alta capacidad antioxidante y antiinflamatoria, Además, poseen gran capacidad para prevenir enfermedades crónicas cardiovasculares y cáncer. Se encuentran principalmente en frutas y verduras. Acorde a los datos presentados, la uva es una de las frutas con alta cantidad de flavonoides. Según al análisis presentado, los compuestos OD280, hue, proantocianidas, prolina y fenoles están altamente relacionados con los flavonoides. En otras palabras, son compuestos que reaccionan con los flavonoides o fomentan los flavonoides. Esto implica que, a mayor presencia de dichos componentes en el vino, mayores beneficios brindará la bebida.

Dicho lo anterior, si quisiéramos realizar una fórmula para un vino que tuviera mayores propiedades, debemos buscar que tenga alta concentración de los compuestos. Por otro lado, si yo como comensal buscara un vino de la mejor calidad, revisaría que tuviera estas propiedades.

Curiosamente, ningún componente tiene alta relación con la cantidad de alcohol que posee el vino, ni la intensidad del color. Esto me hace creer que la fermentación de la uva no afecta de casi ninguna manera a los otros componentes, y puedo suponer que lo único que de verdad determina la intensidad del color del vino es la cantidad de uva que se integra a la mezcla.

## V. Bibliografía

Coder, R. (2022, 12 octubre). 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, 15 julio). 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, 4 agosto). 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.