

Universidad Nacional del Centro de la
Provincia de Buenos Aires

FACULTAD DE CIENCIAS EXACTAS

Ingeniería en Sistemas



**Trabajo Práctico Especial: Análisis de Calidad de vino
Bodega La esperanza**

Fundamentos de la Ciencia de Datos

GRUPO 21

Ortega Mateo: **Mateo Ortega Peratta**

Soutrelle Axel: **Axel Soutrelle**

Lo Fiego Agustín: **Agustín**

ÍNDICE

ÍNDICE.....	1
Introducción.....	3
Limpieza de datos.....	3
Análisis Exploratorio de Datos.....	4
Análisis univariado.....	4
Type.....	4
Citric Acid.....	4
Fixed Acidity.....	5
Volatile Acidity.....	5
Chlorides.....	6
Density.....	7
Alcohol.....	7
Sulphates.....	8
Ph.....	8
Residual Sugar.....	10
Free Sulfur Dioxide.....	11
Total Sulfur Dioxide.....	12
Quality.....	13
Análisis Multivariado.....	14
Clustering.....	16
Conclusión de Clusters.....	18
Planteamiento de hipótesis.....	19
Introducción.....	19
Análisis de correlación de la variable quality.....	19
Hipótesis.....	21
Método para elección de test y validación de hipótesis.....	22
Relación entre las muestras.....	22
Normalidad de las muestras.....	22
Homocedasticidad.....	23
Conclusión.....	23
Analisis de hipotesis.....	24
Hipótesis 1.....	24
Análisis gráfico.....	24
Validación de hipótesis.....	25
Hipótesis 2.....	25
Análisis gráfico.....	25
Validación de hipótesis.....	26



Hipótesis 3.....	26
Análisis gráfico.....	26
Validación de hipótesis.....	27
Hipótesis 4.....	27
Análisis gráfico.....	27
Validación de hipótesis.....	28
Predicción.....	29
Conclusión de la predicción.....	29
Conclusión general.....	30



Introducción

En este informe vamos a analizar un conjunto de datos que contiene información sobre diferentes características de dos tipos de vinos provenientes de una bodega. La idea principal es encontrar patrones y entender qué factores influyen más en la calidad del vino, buscamos proporcionar información valiosa para quienes trabajan en su producción.

Nos enfocamos en variables como la acidez, el contenido de alcohol y otros compuestos químicos presentes en el vino, comparando los efectos en dos tipos de uva: Merlot y Viognier. Para lograr resultados confiables, realizamos una limpieza de datos y aplicamos análisis estadísticos y de clustering, evaluando cómo cada característica podría estar vinculada a la percepción de calidad en el producto final.

Limpieza de datos

Para que el análisis fuera confiable, realizamos una limpieza de datos para corregir posibles problemas en el dataset. Este proceso nos permitió eliminar errores y datos innecesarios que podrían afectar los resultados.

Como primer paso, se utilizaron las herramientas de pandas para obtener información general del dataset. Esto nos ayudó a identificar algunas variables con características inusuales, como **Alcohol** y **Type**.

En el caso de la variable Alcohol, notamos que su tipo de dato estaba definido como object, lo cual nos llamó la atención, ya que, según la información proporcionada por el DataSet, esta variable representa la concentración de alcohol en el vino y, por lo tanto, debería ser numérica (como float o int) en lugar de un objeto. Al intentar convertir el tipo de dato de esta variable usando el método `astype('float64')`, obtuvimos el error “*ValueError: could not convert string to float: '100.333.333.333.333'*”. Esto claramente indicaba un valor erróneo, ya que una concentración de alcohol en esos valores no tiene sentido. Para solucionar el problema, se utilizó una función lambda para localizar todos los valores incorrectos y eliminar las filas que los contenían. Una vez eliminados estos valores, se cambió el tipo de dato de la variable a float para que pudiera ser utilizada correctamente en el análisis.

Para la Variable Type se encontraron dos problemas a la hora de analizarla, la primera era que el nombre de las muestras de un tipo de uva estaba erróneamente escrito, por lo que tuvimos que modificar el nombre de “Viogner” a “Viognier”. Por otro lado se realizó un cambio en el tipo de dato de object a String.



Por último se realizó un chequeo de las filas duplicadas, donde se encontraron casi 600 muestras duplicadas sin sentido, por lo tanto se eliminaron pero manteniendo la primera tupla, para no perder información valiosa.

Análisis Exploratorio de Datos

Análisis univariado

En esta Parte realizamos un análisis detallado de cada variable presente en el dataset.

Type

Es una variable que indica el tipo de vino como string. Toma dos valores:

- **Viognier**, La cepa Viognier es una variedad de uva blanca proveniente de Condrieu, una pequeña comuna del norte del Valle del Ródano, en Francia.
- **Merlot**, El Merlot es un vino tinto suave y versátil, tiene acidez moderada y un tremendo carácter frutal. Es un vino seco, de cuerpo medio a robusto, aunque también varía dependiendo del clima de donde provenga. [\[link\]](#)

Citric Acid

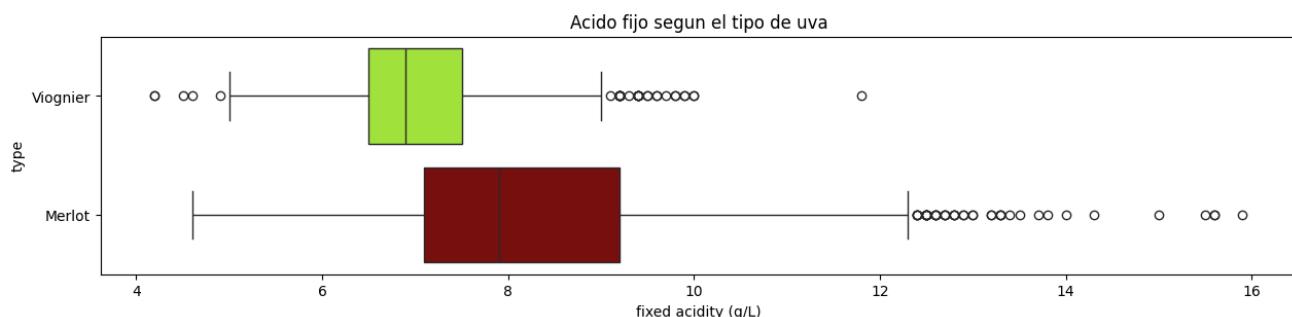
El ácido cítrico (E-330) es un acidificante para corregir la acidez y además posee una acción estabilizante. El Ácido cítrico aporta sensación de frescura, contribuyendo al equilibrio gustativo del vino. **La máxima legal en vino es de 1 g/l.**

Encontramos que uno de los vinos puede ser ilegal, pero decidimos dejarlo ya que no tenemos suficiente información para afirmar que efectivamente es ilegal.



Fixed Acidity

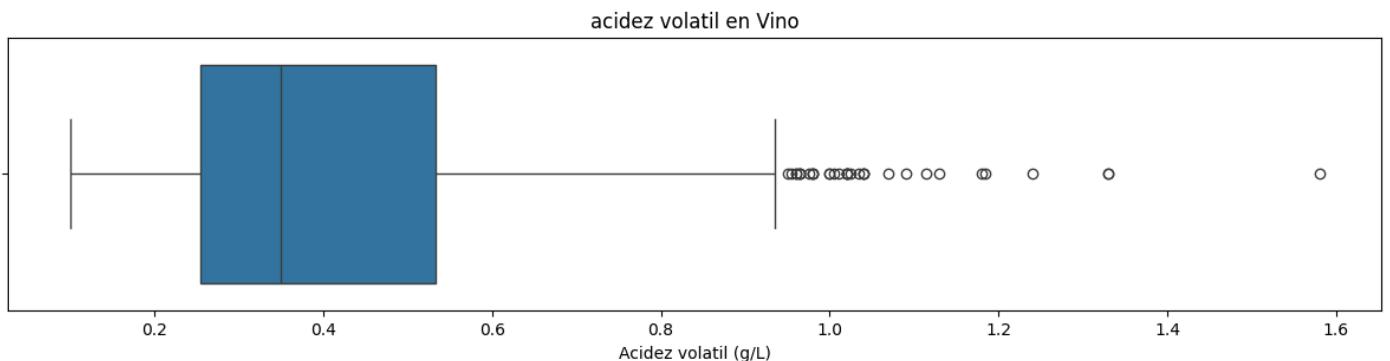
La acidez fija es un parámetro crucial en el análisis de vinos, refiriéndose a los ácidos no volátiles presentes, principalmente ácido tartárico, málico y cítrico. Se mide en gramos por litro (g/L) de ácido tartárico y generalmente oscila entre 4 y 8 g/L ,aunque es preferible que nunca se llegue a ambos extremos, depende del tipo de vino. Este factor es importante porque influye en el sabor, la estabilidad y el pH del vino, además de contribuir a su conservación.



Según el tipo de uva que se utiliza se puede apreciar que la cantidad de ácido fijo cambia. Los vinos a base de uvas Viognier tiene una acidez baja, mientras que los vinos a base de uvas Merlot tienen una acidez moderada, pero no hay información detallada acerca de estos tipos de uvas, por lo que optamos por tomar la información que nos brinda el DataSet y asumir que los Outlier vistos en el gráfico "Ácido fijo en Vino" son valores a no tomar en cuenta.

Volatile Acidity

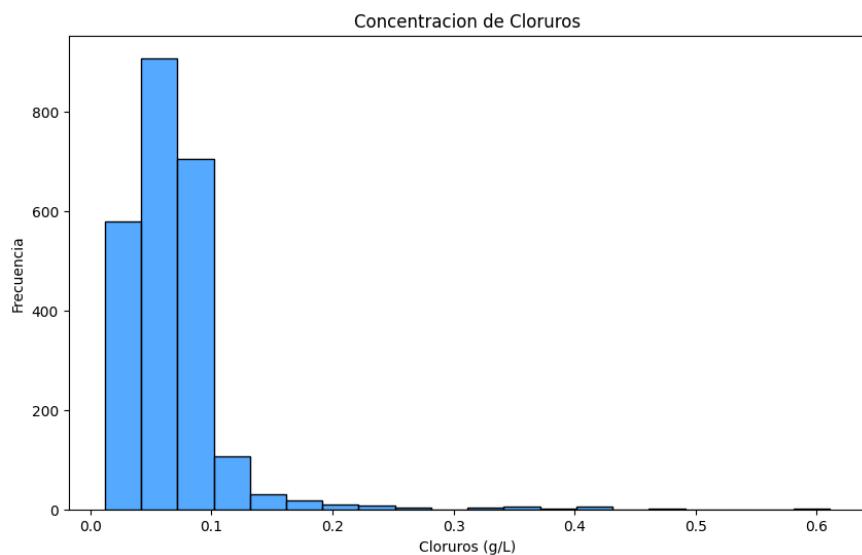
La acidez volátil refleja la cantidad de ácido acético en el vino, lo cual puede impactar en su calidad. Siempre que la acidez volátil no supere los 0,55 o 0,60 gr/litro, el sabor del vino no se verá afectado demasiado, pero hay que tener claro que la calidad de un vino es mucho mayor cuanto más baja sea su acidez volátil. Los valores normales van de 0,30 a 0,60 g/l. Cuando un vino está “picado” presenta una acidez volátil por encima de 1 g/l y aromas que recuerdan al vinagre y al barniz. (fuentes: [link](#), [link2](#))



Utilizamos el Boxplot para analizar los posibles Outliers, donde identificamos que 21 vinos tenían acidez volátil por encima de 1 g/L, lo cual consideramos como outliers o vinos defectuosos. Para que estos vinos no distorsionen los resultados, decidimos eliminarlos del dataset.

Chlorides

Los cloruros en el vino, representan principalmente la concentración de sales de cloruro, como el cloruro de sodio (sal común). Los cloruros afectan el sabor del vino, en pequeñas cantidades, puede mejorar el sabor, pero niveles altos de cloruros no son deseables, ya que pueden darle al vino un sabor salado o incluso amargo, lo cual afecta negativamente su calidad. El Instituto Nacional de Vitivinicultura (INV) establece un límite de cloruros. En cloruro de sodio es de 0,80 g/L para otorgar certificado de análisis para libre circulación y exportación, y de 1 g/L sin ese certificado.(fuentes:[link](#))



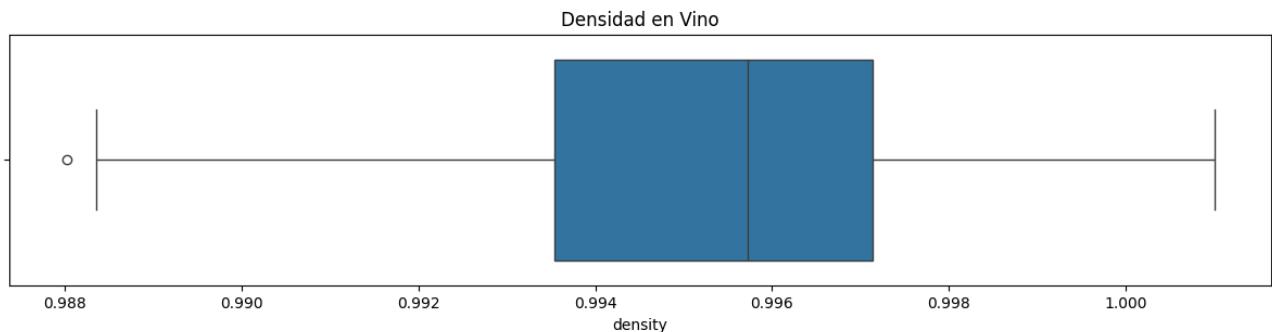
El cloruro en vinos puede tener valores muy altos, por lo tanto no consideramos eliminar los posibles Outlier.

Density

El vino es una mezcla que contiene principalmente sólidos disueltos, los cuales aumentan su densidad por encima del valor de la densidad del agua pura. Sin embargo, también contiene alcohol, cuya densidad es menor que la del agua. La densidad de los vinos blancos y tintos es comparable y normalmente varía entre 0.9912 y 1.0138 g/cm³.(fuentes: [link](#)).

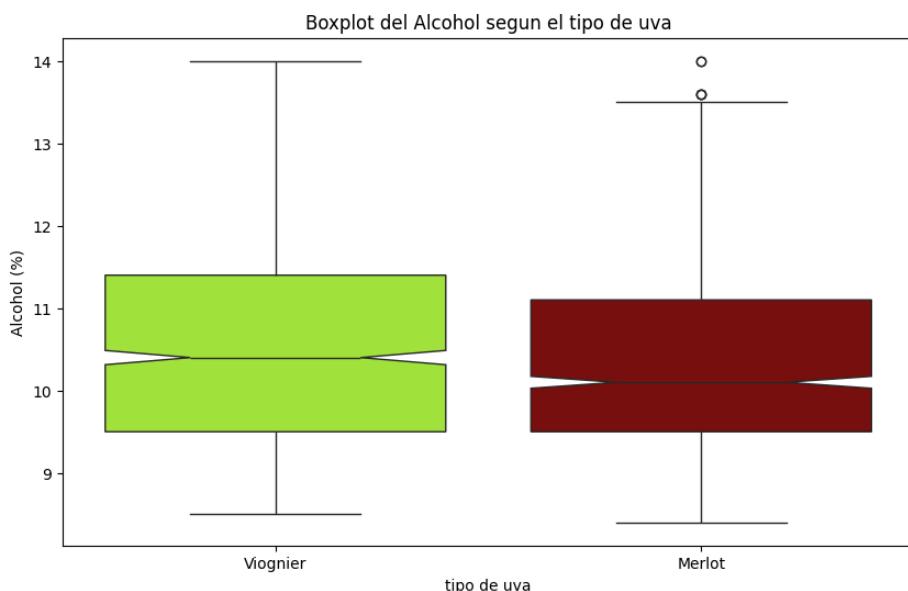
En el DataSet vimos densidades con valores mayores a 10 g/cm³ los cuales no tiene sentido ya que el elemento más denso del mundo es el Osmio que tiene una densidad de 22.6 g/cm³, por lo tanto eliminamos todas las mientras que tiene una densidad mayor a 1.1 gm/cm³.

BoxPlot después de eliminar Outliers:



Alcohol

En general, los vinos tienen un porcentaje que oscila entre el 5% y el 20% y varía según el tipo de uva que se utilizó para su elaboración (La cantidad de azúcar presente en las uvas determinará cuánto alcohol se producirá durante la fermentación). [\[link\]](#)



Podemos ver que los notches no se superponen, eso sugiere que hay una diferencia estadísticamente significativa entre las medianas de esos grupos, al menos con un nivel de confianza del 95%, afirmando que el alcohol depende de la uva.

Sulphates

Contribuyen a los niveles de dióxido de azufre, que actúa como antimicrobiano y antioxidante. Esta sustancia está presente de manera natural en el vino (procedentes de la uva) o añadidas en la práctica enológica.

El vino oscila entre aproximadamente 0.005 g/L y aproximadamente 0.2 g/L. El límite legal máximo en los Estados Unidos es 0.35 g/L. Un vino tinto seco bien elaborado suele tener unos 0.05 g/L de sulfitos.

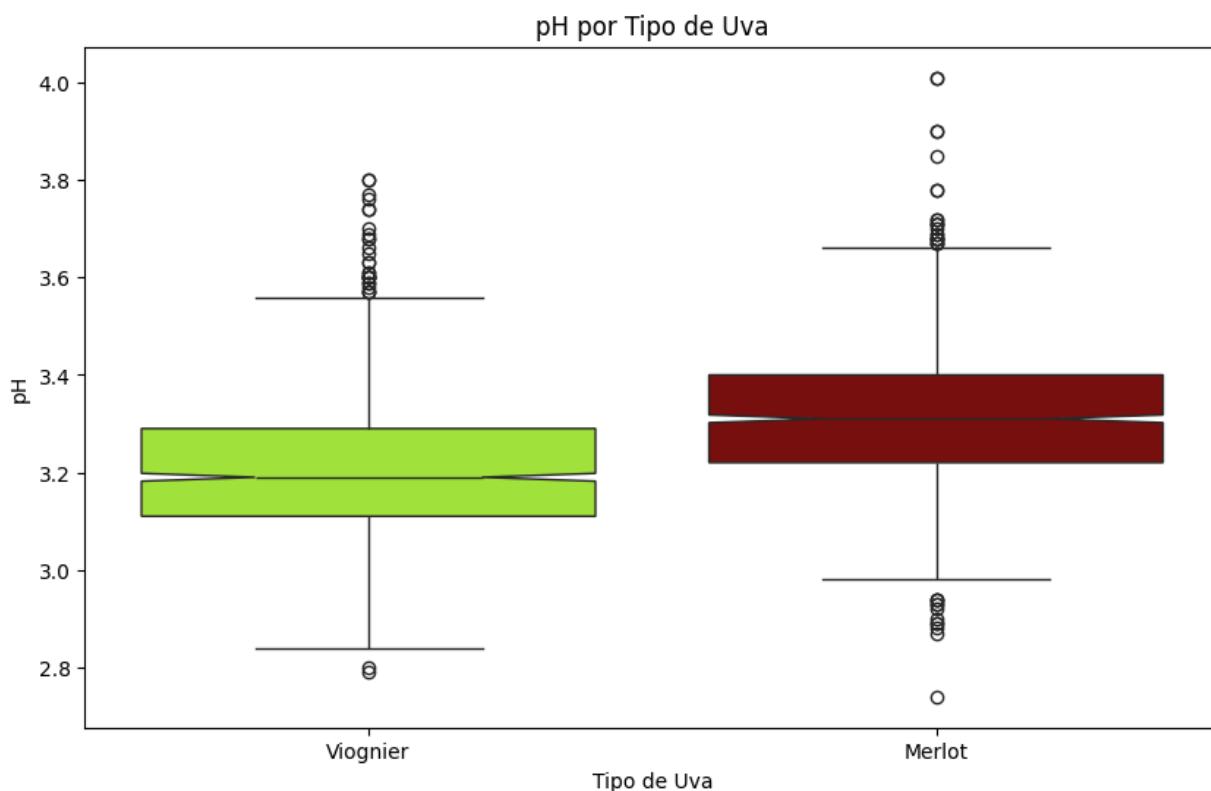
Encontramos posibles Outliers pero como la información que hay sobre la cantidad de sulfato en los vinos no es certera, no podemos validar los supuestos Outliers, por lo tanto los dejamos en el DataSet.(fuentes: [link1](#), [link2](#)).

Ph

Mide el grado de acidez o de alcalinidad de una disolución obtenida de cualquier elemento o sustancia.

El pH de la mayoría de los vinos se encuentra en el intervalo de 2,8 a 4, lo que lógicamente recae en el lado ácido de la escala. Un vino con un pH de 2,8 es extremadamente ácido mientras que uno con un pH en torno a 4 es plano, carente de acidez. Los vinos blancos suelen estar entre 3 y 3,3 y la mayoría de los tintos entre 3,3 y 3,6

El pH de un vino merlot puede variar según la marca y el tipo de vino, pero en general, se encuentra entre 3,3 y 3,6:



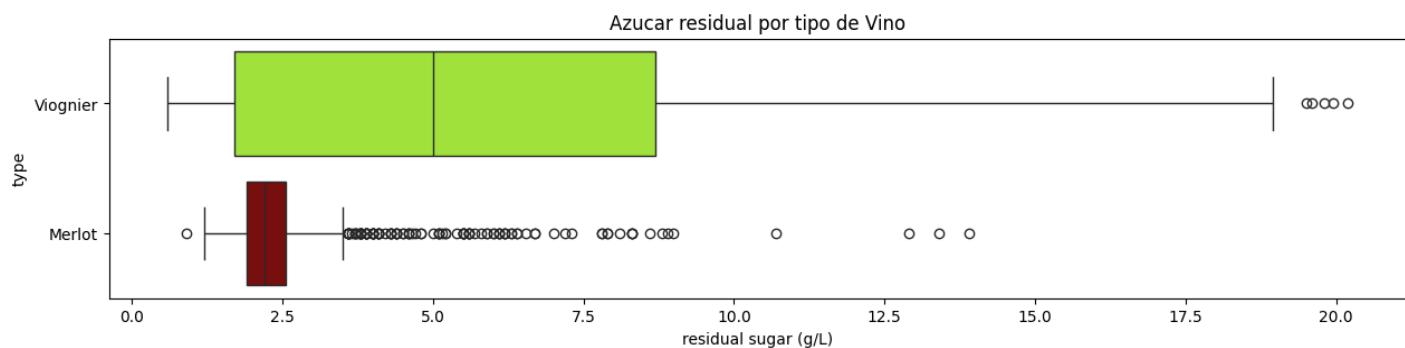
El Merlot tiene un pH mediano más alto en comparación con el Viognier. Esto sugiere que, en promedio, el Merlot es ligeramente menos ácido que el Viognier, ya que un pH más alto indica menor acidez.

Decidimos eliminar esos posibles Outliers ya que no están en los niveles normales de pH.(fuente:[link](#))

Residual Sugar

El azúcar residual proviene de los azúcares naturales de la uva que quedan en el vino después de que finaliza la fermentación alcohólica. Se mide en gramos por litro. Dependiendo de la cantidad de azúcar residual, es posible clasificar el vino como seco, dulce, etc.

Por ende, todos los valores que tenemos tienen sentido aunque predominan los vinos secos.



Vemos que hay una diferencia notoria de azúcar residual según el tipo de uva.

- No consideraremos eliminar los posibles Outliers ya que los vinos pueden tener cantidades de azúcar mucho más altas que las presentes en el BoxPlot. (fuente:[link](#))

Free Sulfur Dioxide

Se refiere a la cantidad de dióxido de azufre en el vino que no está químicamente unido y permanece activo. Esta forma de SO₂ actúa como conservante y antioxidante en el vino, protegiendo contra la oxidación y el crecimiento microbiano.

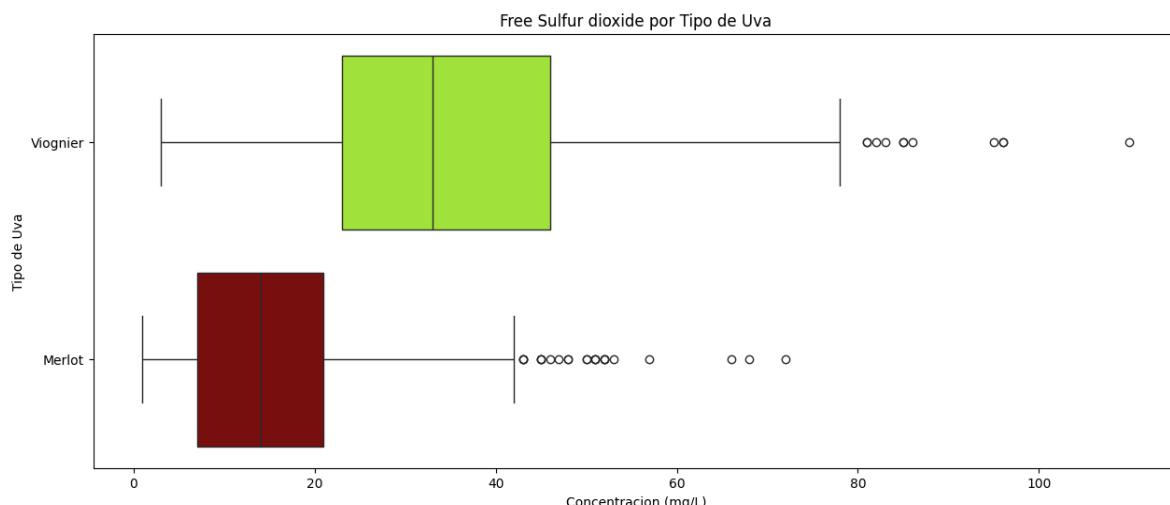
Una concentración de SO₂ activo de 0,35 mg/L garantiza una protección mínima, y un valor de 0,6 mg/L de protección máxima.

Vinos blancos:

- 20 a 50 mg/L para la mayoría de los vinos blancos.
- En algunos vinos dulces blancos o vinos de postre, los niveles pueden ser aún mayores (hasta 60-70 mg/L).

Vinos tintos:

- 15 a 30 mg/L para la mayoría de los vinos tintos.



Decidimos eliminar esos posibles Outliers ya que no encontramos información suficiente para validar esos valores. (fuentes: [link1](#), [link2](#))

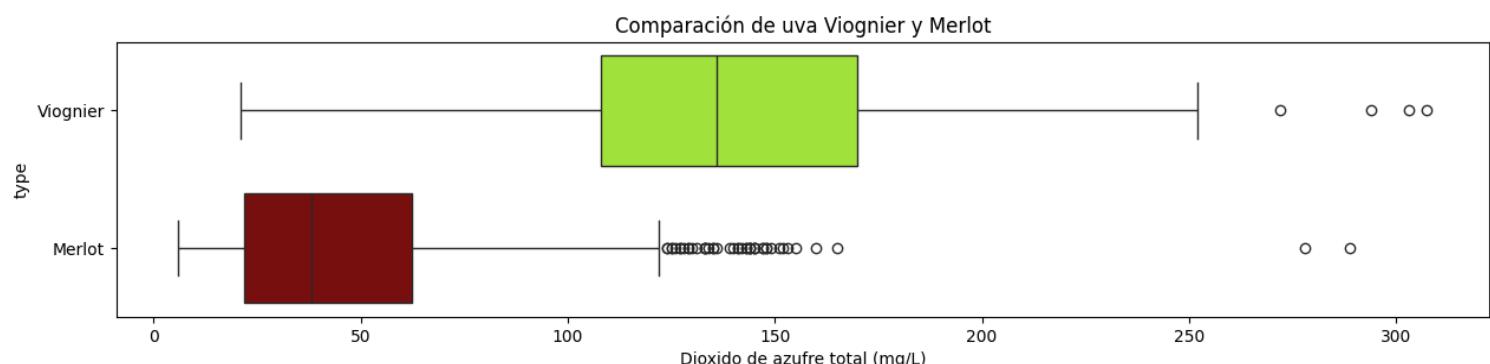


Total Sulfur Dioxide

Incluye tanto el SO₂ libre(Free Sulfur Dioxide) como el que está unido a otras moléculas en el vino, como azúcares y ácidos.

el mayor valor alcanzable es 400 mg/L y el mínimo de 0,5 mg/L. Lo cual indica que la variable tiene valores razonables.

Dióxido de Azufre total según el tipo de uva:



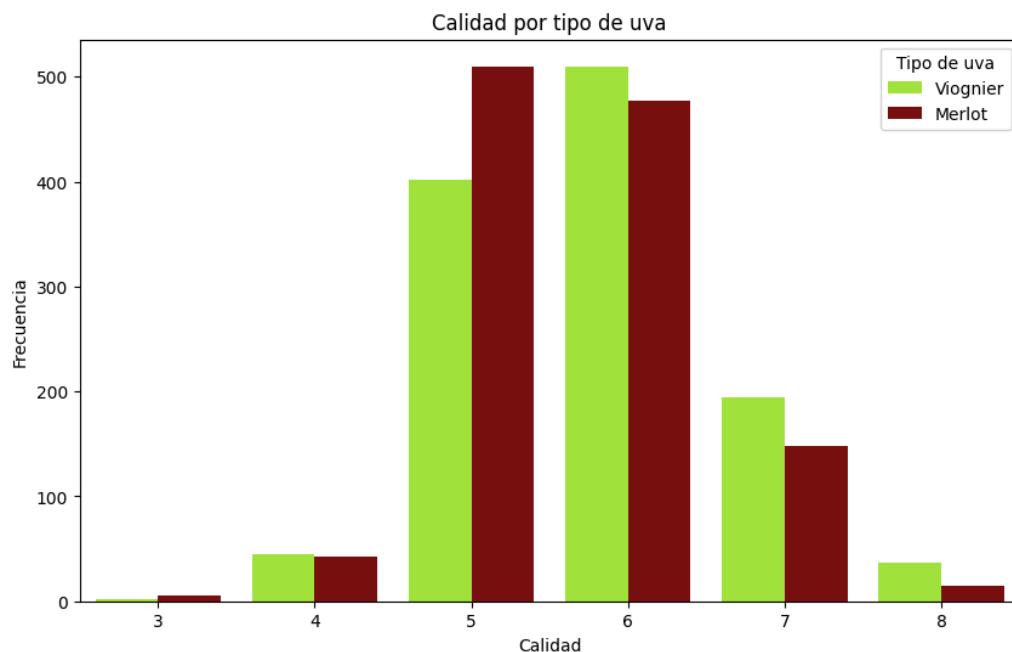
- Los vinos blancos, como los Viognier, tienen una cantidad de dióxido de azufre total de hasta 200 mg/l.
- Los vinos tintos, como los Merlot, tienen una cantidad menor que 150 mg/l.

Para los blancos y tintos con más de 4 g/l de azúcar, su máximo es 300 mg/l de dióxido de azufre total.

Tanto los vinos Merlot como los Viognier con valores de dióxido de azufre totales mayores a 150 y 200 mg/l respectivamente, y con una azúcar residual menor a 4 g/L se van a considerar como Outliers, por ende procedemos a eliminarlo. (fuentes: [link1](#), [link2](#))

Quality

Puntuación del vino, basada en evaluaciones sensoriales, con una escala de 0 a 10



Para los niveles de calidad de 5 y 6, los vinos Viognier tienen una frecuencia un poco mayor que los Merlot, especialmente en la calidad 6.

En la calidad 7, el Viognier también tiene una mayor frecuencia, lo que indica que los vinos Viognier tienden a calificar un poco mejor que los Merlot en este conjunto de datos.

Los niveles de calidad más bajos (3 y 4) y más altos (8) son poco frecuentes para ambos tipos de vino.

Los pocos vinos con calidad de 3 parecen ser exclusivamente Merlot, mientras que en calidad 8 hay una pequeña representación de ambos tipos de uva pero predominan los Viognier.

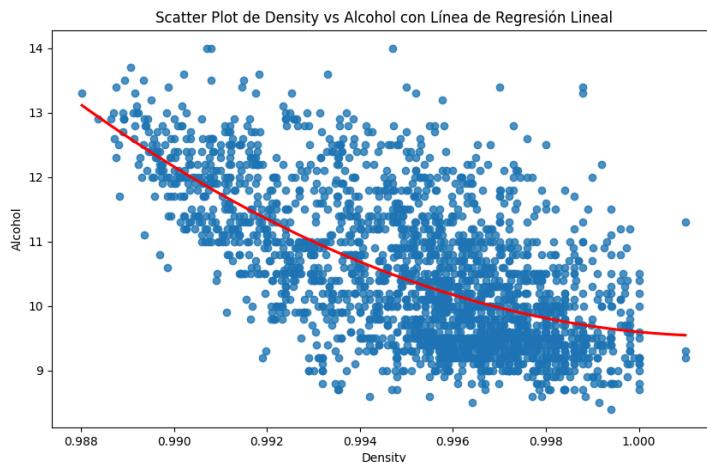
La calidad del vino no depende estrictamente de la uva, todas las variables del dataset influyen en la calificación del vino.

Análisis Multivariado

Analizando la matriz de correlación pudimos obtener información sobre qué variables pueden ser interesantes estudiar para intentar predecir la calidad de los vinos.

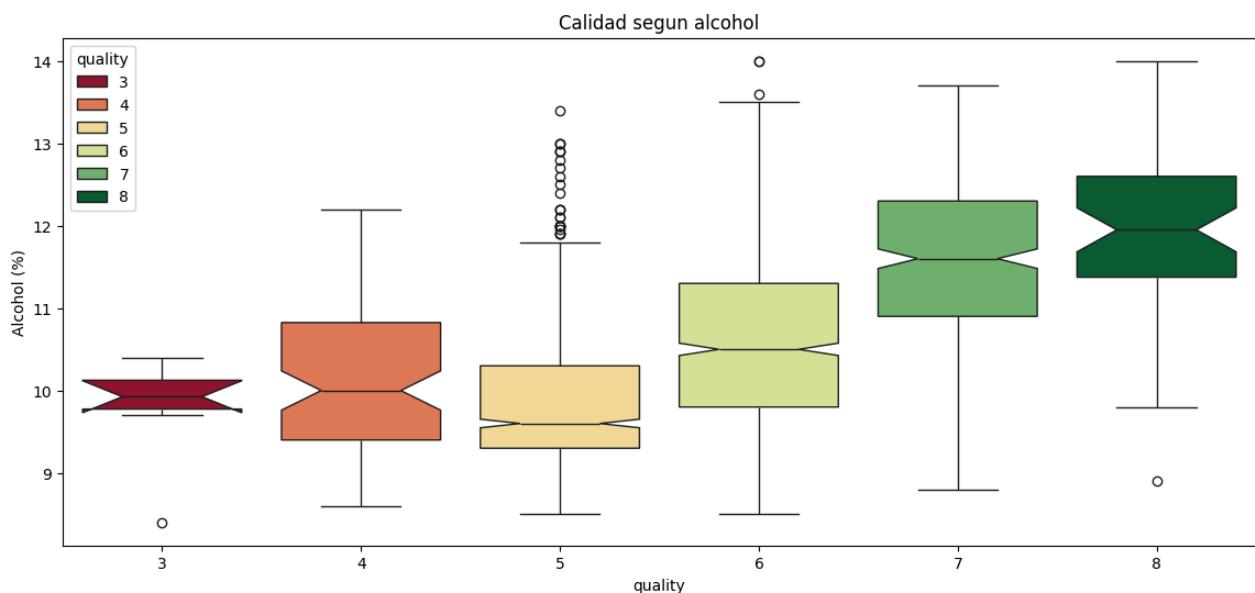
Si bien las variables Free Sulfur dioxide y Total Sulfur dioxide tienen una correlación muy alta (más de 0.7) decidimos no analizar ya que una está formada a partir de la otra por lo que la relación es más que obvia.

La relación que existe entre la calidad del vino y la cantidad de alcohol que tiene es un punto interesante para intentar predecir la calidad de los vinos. Pero podemos ver que la densidad también es una variable interesante ya que tiene una relación relativamente alta con el alcohol.



Este gráfico nos muestra que efectivamente hay una relación entre densidad y alcohol, sin embargo los puntos están muy dispersos, por lo que podemos concluir que hay otras variables que afectan a esta relación.

Analizando la relación entre la calidad y el alcohol (ver gráfico “Calidad según alcohol”), se puede apreciar una tendencia a tener mejor calidad cuanto mayor sea el porcentaje de alcohol, sin embargo este DataSet no tiene suficientes muestras tanto de calidad buena (4 o menos) como de calidad alta (7 o más) para poder dar una conclusión sólida sobre dicha relación, pero más adelante retomaremos esto a través de la [Hipótesis 2](#).

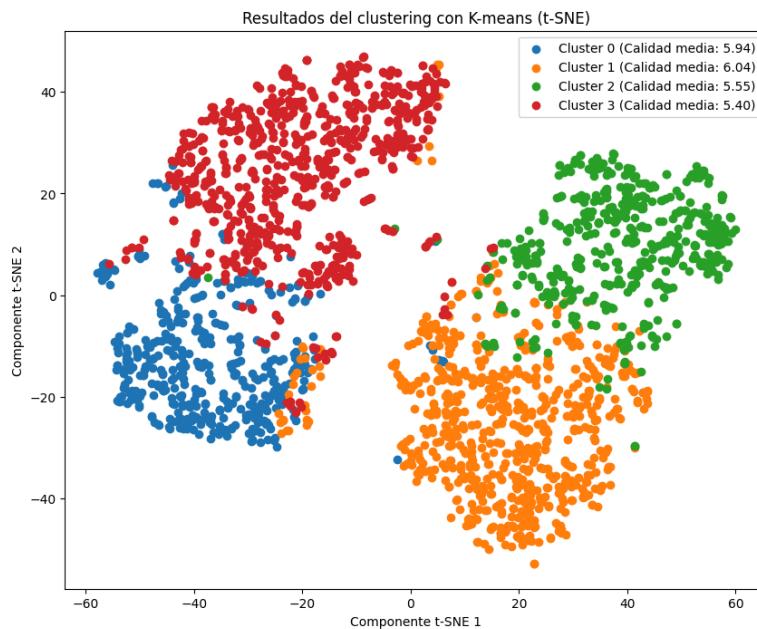


También analizamos las correlaciones según el tipo de uva, pero solo pudimos deducir que las uvas Viognier tienen una relación muy fuerte entre su densidad y la cantidad de azúcar residual, mientras que las uvas Merlot no tienen relaciones fuertes, sin embargo notamos que existe una diferencia de correlaciones entre calidad y alcohol, ya que en uvas Merlot la correlación entre estas variables es un poco más fuerte que en las uvas Viognier, pero eso lo analizamos en profundidad en la Hipótesis 1.

En conclusión, con respecto a las correlaciones que existen, la mayoría no son extremadamente fuertes, lo que sugiere que la calidad del vino está influenciada por otros factores y posiblemente por otros elementos que no están incluidos en este dataset, por ende analizaremos a partir de clusters para obtener mayor información sobre los factores más relevantes que afectan la calidad del vino.

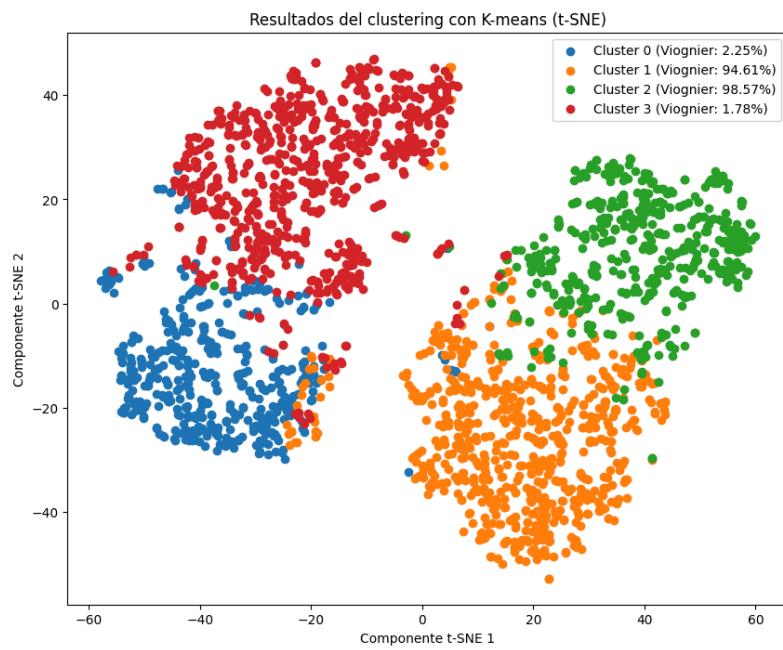
Clustering

Este gráfico muestra cómo se agrupan los datos en función de sus características usando K-means y la reducción de dimensionalidad con t-SNE.



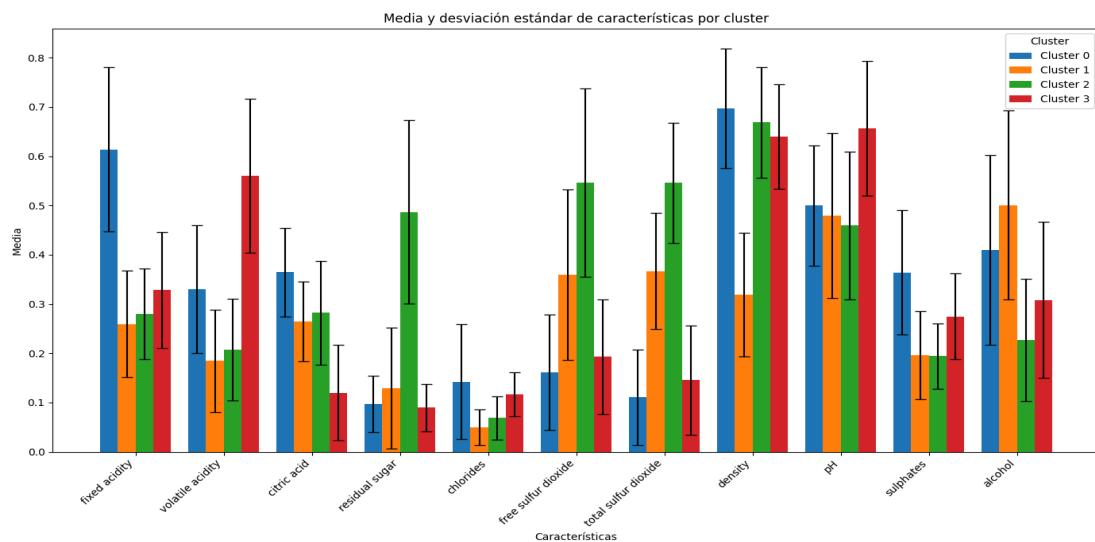
La calidad promedio de los clusters (de 5.40 a 6.04) indica diferencias pequeñas para distinguir claramente la calidad.

Además, los clusters tienen una alta concentración de uno de los tipos de uva, lo cual sugiere que las características en el dataset están bien diferenciadas entre Viognier y Merlot en ciertos clusters.



El gráfico de Silhouette muestra valores bajos, lo que indica que los clusters no son compactos y bien definidos. Esto sugiere que los grupos formados no son sólidos y no hay una estructura de clustering fuerte en los datos.

El índice de Davies-Bouldin también es alto, lo que refuerza la idea de que los clusters no son bien definidos y pueden no ser útiles para predecir la calidad. Las grandes desviaciones estándar en varias características indican que los datos dentro de cada cluster son bastante variados y, por lo tanto, los clusters no representan bien los datos. Además, las medias de muchas características no varían significativamente entre los clusters, lo que implica que no existen diferencias claras en las características que definen la calidad del vino.



Conclusión de Clusters

Los clusters generados no son suficientemente fuertes ni consistentes para diferenciar de manera confiable la calidad del vino. Las métricas de validación de clusters sugieren que la estructura de los datos no es adecuada para clustering en relación a la calidad. La variabilidad dentro de los clusters (como lo muestra la desviación estándar) indica que los grupos son heterogéneos y que las características seleccionadas no permiten diferenciar claramente la calidad. Además, dividimos el DataSet en dos partes según el tipo de uva y tampoco obtuvimos resultados convincentes a partir del Clustering.

La limitación de los datos es un factor importante, ya que pueden no ser suficientes para tener una predicción sólida, además otras características pueden ser aún más determinantes para la calidad del vino que las que tenemos.



Planteamiento de hipótesis

Introducción

La Bodega afirma que sus clientes compran vinos basados en su puntuación. Por ende, se planteó una inversión que tienen pensada hacer una mejora de la calidad de sus vinos con el fin de que tengan una mejor puntuación como estrategia de conseguir nuevos clientes. Es así como nos encargaron analizar la relación de la calidad de un vino con los diversos atributos que lo caracterizan. Por esta razón, nuestro análisis se centrará en la variable quality que indica la puntuación de cada vino y otras variables de interés para la Bodega.

Análisis de correlación de la variable quality

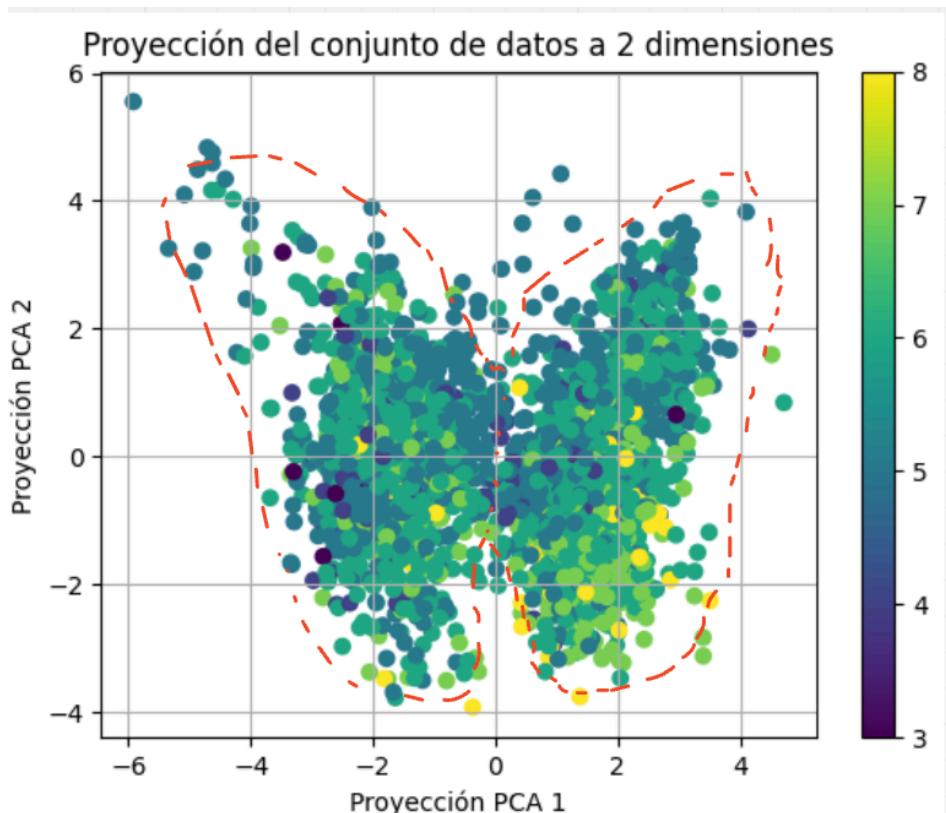
Comenzamos estudiando la correlación entre la variable quality con las distintas variables del dataset porque buscamos indicios sobre qué variables pueden tener mayor incidencia en el puntaje de calidad de los vinos. Para este análisis se utilizaron dos métodos distintos: Matriz de correlación y PCA.

A partir de la matriz de correlaciones (ver gráfico 4.1.1) observamos que en un principio solo la variable alcohol que parece tener una correlación medianamente alta(0,5) con la variable quality y luego todas las demás variables tienden a tener una correlación cercana a 0. Luego, se planteó la idea de que sería interesante plantear una hipótesis formal sobre la relación entre alcohol y quality.

Para usar el método de PCA primero se repasaron las variables contenidas en el dataset y ver cuales de ellas son dicotómicas ya que es un requerimiento quitarlas para realizar este análisis de componentes principales. Con este objetivo, encontramos que la variable type era la única variable que cumplía esta restricción así que se procedió a quitarla del dataset con el que se hará PCA. Además, como se buscó estudiar la correlación de la variable quality con las demás variables también debimos quitar a quality del dataset.

A partir del análisis usando PCA (ver gráfico 4.1.2) no se obtuvieron conclusiones muy certeras. Los datos parecen estar distribuidos de manera continua, sin divisiones claras o grupos evidentes, aunque notamos que podría haber una estructura subyacente con forma de “mariposa”. Donde cada una de las dos alas podrían representar uno de los dos tipos distintos de vinos.





(gráfico 4.1.2 alterado para indicar la estructura de mariposa que nombramos)

Más allá de esta última idea, se observó que los datos se dispersan principalmente a lo largo de ambas dimensiones del PCA, con una ligera tendencia a agruparse en estas dos áreas que se señalaron anteriormente. Aunque no hay una clara separación de la variable *quality* en el espacio proyectado, parece que los valores de *quality* tienden a ser más frecuentes a medida que incrementa el componente 1 y disminuye el componente 2, provocando que los puntos amarillos y verdes claros se concentren más en estas áreas específicas del gráfico.

Luego de este análisis, nos quedamos con la idea de plantearnos formalmente la hipótesis de que la calidad pueda estar relacionada o no con algunos de estos grupos que en un principio identificamos como los dos tipos distintos de vinos.

Hipótesis

En base a la información que recolectamos de la correlación de la variable quality con el resto de variables del dataset planteamos las hipótesis. Algunas de estas hipótesis también fueron motivadas por charlas con la dirección de la bodega ya que había preguntas puntuales que querían resolver. Así, las hipótesis planteadas fueron:

1. Los vinos elaborados con el tipo de uva Merlot tienen una puntuación promedio igual a la de los vinos elaborados con el tipo de uva Viognier. (ver sección 4.2)
2. Los vinos de la bodega con un alto nivel de alcohol tienen una puntuación mayor que los vinos con un bajo nivel de alcohol. (ver sección 4.3)
3. Los vinos con mayor acidez tienen menor puntuación que aquellos con una menor acidez. (ver sección 4.4)
4. Los vinos de la bodega con niveles de ácido volátil más altos tienen una puntuación de calidad más baja. (ver sección 4.5)
5. La puntuación de calidad de los vinos que tienen niveles de ácido volátil bajos y mayor porcentaje de alcohol es superior a los vinos que no tienen estas características.
6. La densidad, el azúcar residual y la acidez fija vienen dadas por el tipo de uva con el que el vino fue hecho.



Método para elección de test y validación de hipótesis

En esta sección explicaremos el método seguido para tomar las decisiones previas a la elección del test de hipótesis y posteriormente el método de test de hipótesis.

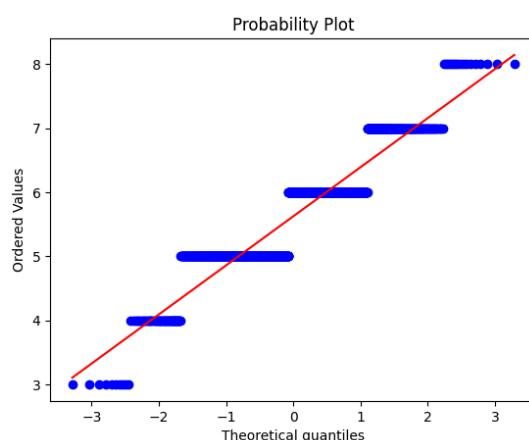
Relación entre las muestras

Primero, comenzamos preguntándonos si las muestras obtenidas de la población eran independientes o dependientes. Llegamos a la conclusión de que todas nuestras muestras resultan ser independientes.

Seguimos el siguiente razonamiento: Nuestra población son todos los vinos provenientes de la bodega. Las características de los vinos las cuales utilizamos para separar la población en muestras hacen que los grupos sean conceptualmente diferentes. La selección de un vino para una muestra no influye ni reduce la cantidad de posibles selecciones de vinos para la otra muestra con la que se quiere comparar.

Normalidad de las muestras

En un principio, teníamos la intención de realizar un test de normalidad para las cuatro hipótesis. Al comenzar con la primera hipótesis, realizamos un test de Shapiro-Wilk, obteniendo un p-valor igual a cero para ambas muestras, lo que nos llevó a concluir que las muestras no siguen una distribución normal. Sin embargo, para confirmar esta observación, generamos un QQ plot, un método gráfico que evalúa si un conjunto de datos sigue una distribución normal. Este gráfico compara los cuantiles teóricos de una distribución normal con los cuantiles observados de nuestros datos. Como esperábamos, no encontramos una relación lineal al analizar el gráfico.



Lo que observamos derivó en darnos cuenta de que la distribución de la variable no es continua sino que es una variable categórica ordinal, mostrando saltos en lugar de una línea continua. En este punto, comprendimos que las muestras de la variable *quality* siempre tendrán una distribución no normal debido a su naturaleza discreta. Por tanto, determinamos que realizar un test de normalidad no sería necesario en adelante, dado que la estructura de la variable garantiza una distribución no normal.

Homocedasticidad

La homocedasticidad (homogeneidad de varianzas) es un supuesto fundamental en el test de Mann-Whitney (test usado para probar hipótesis). Se refiere a que las varianzas de los grupos que se están comparando deben ser aproximadamente iguales. El test que usamos fue el test de Levene que es robusto para distribuciones que no son normales. Como en este caso trabajamos con datos que no son normales, usaremos el test de levene en todos los casos. En Levene la hipótesis nula es que las varianzas son significativamente diferentes entre sí (hay heterocedasticidad).

Conclusión

Primero resolvimos cual era relación de dependencia/independencia entre las muestras, luego cuantas muestras tenemos, si las muestras siguen una distribución normal o no y por último la validación de homocedasticidad para posteriormente determinar qué test usamos según el marco (ver tabla 1).

Distribución	Muestras independientes	
	2 muestras	≥ 2 muestras
Normal (paramétrico)	Test t (corregido o no)	Anova
Libre (no paramétrico)	Wilcoxon o (Mann-Whitney U)	Kruskal - Wallis

(Tabla 1: Tests)



Analisis de hipótesis

Hipótesis 1

La bodega tiene tan solo 2 tipos de uva con la que hacen vinos. El tipo de uva Merlot y el tipo Viognier. Desde el análisis hecho usando PCA ya nos preguntamos ¿Cuál será la uva que obtenga mejores puntuaciones? En un principio consultamos con la bodega y si bien ellos creían que los vinos hechos con la uva Merlot generalmente tienen mejor puntuación, no estaban del todo seguros debido a la gran cantidad de vinos que tienen.

Por otro lado, nosotros en base a nuestro análisis previo creímos que no habría grandes diferencias en la puntuación de calidad de los vinos hechos con estos dos tipos de uva. Para este análisis vamos a usar las variables quality y type (ya que indica el tipo de uva). A raíz de esta pregunta planteamos la siguiente hipótesis:

- **Los vinos elaborados con el tipo de uva Merlot tienen una puntuación promedio igual a la de los vinos elaborados con el tipo de uva Viognier.**

Análisis gráfico

Comenzamos por hacer un análisis gráfico de la hipótesis con un gráfico de boxplot para cada una de las muestras (ver gráfico 4.2.1). El boxplot compara la distribución de la puntuación de calidad para los vinos de tipo Viognier y Merlot. Este gráfico resultó ser muy similar para ambos tipos de vino, por lo que optamos por añadir la media para poder observar mayores diferencias. A partir de este gráfico, podemos deducir las siguientes observaciones: La mediana es similar, ambas variedades de vino tienen una mediana de calidad cercana, esto indica que los valores centrales de calidad son bastante similares para los dos tipos de vino. Además, en ambos casos el IQR es parecido, lo que sugiere que la variabilidad en la puntuación de calidad es similar entre ambas clases de vino. También, notamos que los puntos rojos (que representan la media) parecerían indicar que los vinos Viognier tienen un puntaje levemente superior a los vinos Merlot lo cual en un principio nos indica que nuestra hipótesis puede ser rechazada.



Validación de hipótesis

Dado que las muestras no son normales y se prueba la homocedasticidad de las muestras, elegimos como test el test de Mann-Whitney. El test nos dio como resultado que hay que rechazar la hipótesis nula y por ende concluimos que hay diferencia entre la puntuación promedio de los vinos Merlot y Viognier. Luego de realizar la prueba de Mann-Whitney a dos colas nos dimos cuenta que esta respuesta sólo nos indica que un tipo de vino tiene mayor puntuación que otro pero no sabíamos cual, aunque nos dábamos una idea viendo el boxplot. Por ende, usamos el test de Mann-Whitney a una cola y verificamos que los vinos hechos con tipo de uva Merlot tienen mayor puntuación promedio que los vinos hechos con el tipo de uva Viognier.

Hipótesis 2

En el mapa de correlaciones hecho en la sección de análisis de la variable quality habíamos observado como la variable quality tenía una correlación considerable con la variable alcohol. Cuando consultamos con la bodega nos hablaron de que no siempre era el caso que los vinos con mayor graduación alcohólica eran más vendidos y que ellos suponían que no era algo que tenga mucha relación ya que creen que sus clientes no buscan un vino por su porcentaje de alcohol. Nosotros no teníamos clara esta cuestión y decidimos plantear la siguiente hipótesis:

- **Los vinos de la bodega con un alto nivel de alcohol tienen una puntuación mayor que los vinos con un bajo nivel de alcohol.**

Para decidir qué consideramos como un vino con un alto nivel de alcohol usamos la media como regla. Los vinos con porcentaje de alcohol superior a la media son considerados vinos con un alto nivel de alcohol y los que su porcentaje de alcohol no supera la media son considerados vinos con un bajo nivel de alcohol. Para este análisis usamos la variable quality y la variable alcohol.

Análisis gráfico

Comenzamos por hacer un análisis gráfico de la hipótesis con un gráfico de boxplot para cada una de las dos muestras (ver gráfico 4.3.1). De aquí lo primero que pudimos observar es una clara diferencia entre ambas muestras que nos hablaba de que nuestra hipótesis tiene sentido. Observamos que los vinos con mayor contenido de alcohol tienden a tener una mediana de



calidad más alta comparada con los vinos con menor contenido de alcohol. Además, se puede ver que la media de los vinos con mayor porcentaje de alcohol es superior y eso nos indica que, en promedio, los vinos con niveles altos de alcohol reciben mejores puntuaciones de calidad.

Validación de hipótesis

Como no se cumple normalidad y tampoco homocedasticidad optamos por un test de Kruskal-Wallis. El resultado obtenido fue que había que rechazar la hipótesis nula y por ende, hay una diferencia significativa en la puntuación de calidad entre vinos con alto y bajo porcentaje de alcohol. Con el análisis gráfico previamente realizado resulta bastante obvio llegar a la conclusión que los vinos con un alto porcentaje de alcohol tienen mayor puntuación en promedio que los vinos con un bajo porcentaje de alcohol. Intentamos reafirmar nuestra decisión con un test de Wilcoxon pero no cumplimos el supuesto de que las muestras sean dependientes.

Hipótesis 3

La bodega nos transmitió su preocupación acerca de que notan una tendencia de que la composición de las uvas está cambiando a medida que pasan los años y que están seguros de que esta alteración deriva en una mayor acidez en la uva.

Por esta razón, las hipótesis 3 y 4 van centradas en esta advertencia por parte de la bodega. Así es como planteamos la siguiente hipótesis:

- **Los vinos con mayor acidez tienen menor puntuación que aquellos con una menor acidez.**

Siguiendo la metodología de anteriores hipótesis definimos como pH alto o pH bajo a los vinos que tienen pH inferior y superior al promedio, respectivamente. (Nota: mientras más chico sea el valor de pH mayor es la acidez). Para este análisis usamos las variables quality y pH del dataset.

Análisis gráfico

Comenzamos por hacer un análisis gráfico de la hipótesis con un gráfico de boxplot para cada una de las dos muestras (ver gráfico 4.4.1). Lo que observamos es que ambos boxplot nos dieron prácticamente idénticos. Esto en un principio es bueno ya que indica la falta de una



diferencia notable entre los grupos sugiere que, en este dataset, el pH no tiene un efecto significativo sobre la calidad promedio del vino.

La mayoría de los valores de pH están concentrados entre 3.0 y 3.4, con una forma de distribución aproximadamente normal. El pico alrededor de 3.25 indica que esta es la concentración de pH más común en los vinos del conjunto de datos. Este rango relativamente estrecho de pH podría explicar por qué no se observa una fuerte relación entre el pH y la calidad, ya que la variabilidad en pH es limitada.

Validación de hipótesis

Dado que las muestras no son normales y se prueba la homocedasticidad elegimos como test el test de Mann-Whitney. El test nos dio como resultado que no hay que rechazar la hipótesis nula, por ende no existe una diferencia significativa entre la calidad de los vinos con altos valores de pH y los vinos con bajos valores de pH.

Hipótesis 4

Siguiendo el propósito de la hipótesis 3 de buscar la relación entre el incremento de la acidez de uva y calidad del vino, nos propusimos estudiar la incidencia de los niveles de ácido volátil en la puntuación de calidad de los vinos. Es así como planteamos la siguiente hipótesis:

- **Los vinos de la bodega con niveles de ácido volátil más altos tienen una puntuación de calidad más baja.**

Como hemos hecho anteriormente, sepáramos los niveles de acidez volátil en tres categorías: alta, media y baja. Para separar estas 3 categorías usamos cuantiles para no alterar los resultados en base a una decisión arbitraria nuestra. Además, así logramos una distribución más pareja entre las 3 categorías que definimos. Para este análisis usamos las variables quality y volatile acidity.

Análisis gráfico

Comenzamos por hacer un análisis gráfico de la hipótesis con un gráfico de boxplot para cada una de las tres muestras (ver gráfico 4.5.1). Observamos que la acidez volátil tiende a disminuir ligeramente a medida que la calidad aumenta. Sin embargo, la diferencia entre las categorías no es drástica. Este análisis podría indicar que la acidez volátil puede influir en la calidad. Además, tanto la categoría de calidad Media como la Alta presentan outliers, especialmente por encima de



0.65 g/L en la acidez volátil, lo cual indica que hay algunas observaciones con valores bastante altos en comparación con el resto.

Validación de hipótesis

Como no se cumple normalidad y homocedasticidad optamos por un test de Kruskal-Wallis para las tres muestras. El resultado obtenido fue que había que rechazar la hipótesis nula y por ende, hay una diferencia en la acidez volátil entre al menos uno de los grupos de calidad (alta, media y baja).

Como no sabemos en qué sentido se dan estas diferencias y si bien el boxplot nos dio una idea, vamos a analizar con otro boxplot por cada puntaje de calidad para ver si realmente la acidez volátil varía significativamente entre los distintos niveles de calidad de los vinos. (ver gráfico 4.5.2) .

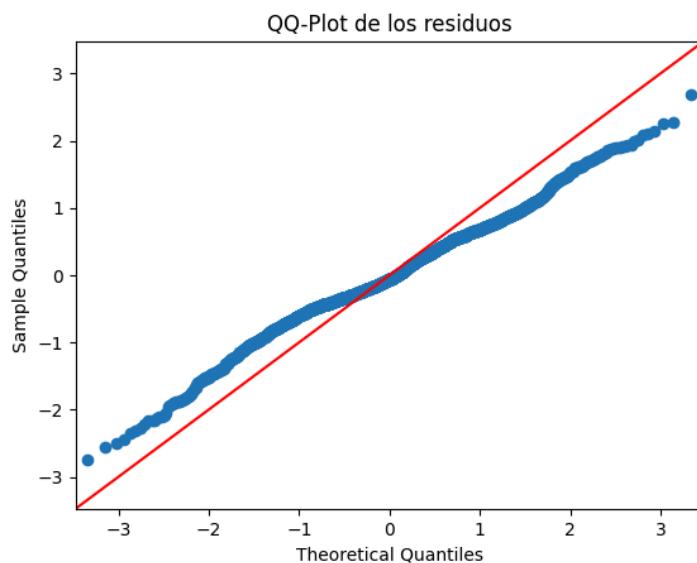
Observando el nuevo gráfico, finalmente concluimos que es evidente la tendencia de que a mayor acidez volátil tenga el vino menor es su puntuación de calidad. Además, también comprobamos esto nuevamente usando otro dataset de vinos que encontramos en Kaggle (ver gráfico 4.5.2).



Predicción

Según las hipótesis 1, 2 y 4, las variables type, alcohol y volatile acidity tienen una relación con la calidad de los vinos, por lo que las utilizamos para intentar predecir dicha calidad mediante un modelo de regresión lineal múltiple. Los resultados no fueron satisfactorios, ya que el valor de R^2 (R-squared) fue bajo, lo cual sugiere que estas variables no son suficientes para explicar la variabilidad en la calidad de los vinos. El valor de R^2 nos indica qué proporción de la variabilidad en la calidad puede ser explicada por las variables seleccionadas. En este caso, el valor bajo de R^2 sugiere que estas variables no son lo suficientemente útiles para realizar predicciones precisas.

Para validar esta conclusión, utilizamos el gráfico QQ-Plot para verificar la normalidad de los residuos. El gráfico mostró que los residuos están bastante cerca con la línea de normalidad, lo que indica que la suposición de normalidad de los errores es razonable.



Conclusión de la predicción

Las variables elegidas (alcohol, acidez volátil y tipo de uva) no son tan útiles para predecir la calidad de los vinos. Este resultado sugiere que podrían faltar variables importantes en el modelo o que la relación entre las variables y la calidad es más compleja y no se ajusta bien con una regresión lineal múltiple.

Conclusión general

En conclusión, este estudio ofrece una visión sobre cómo ciertas características del vino pueden influir en su calidad, pero también revela la necesidad de buscar e incluir variables adicionales y mayor cantidad de muestras en el dataset. Usando los análisis de correlación y las propuestas por parte de la bodega, pudimos establecer conclusiones que fueron de mucha utilidad para dar respuestas a preguntas importantes.

Si bien las Hipótesis 1, 2 y 4 mostraron que las variables planteadas tienen relación con la puntuación de calidad del vino, no pudimos hacer predicciones sobre la calidad del vino a partir de estas.

Los datos incluidos en el DataSet no parecieran ser suficientes para poder hacer un modelo de predicción útil y eficaz. Para lograr un análisis multivariado más preciso, sería ideal contar con más muestras de vinos de calidades extremas: tanto de baja calidad (menores a 3) como de alta calidad (mayores a 8). Actualmente, estos casos son muy pocos (menos del 2.17% del total), lo cual limita nuestra capacidad para captar patrones claros en esos niveles. Tener más datos en estos extremos nos ayudaría a obtener resultados más confiables y representativos.

Además, sería útil agregar al dataset otros tipos de uva y diferentes componentes químicos del vino. Esto permitiría explorar un rango más amplio de características y su relación con la calidad, dándonos una visión más completa. Con un conjunto de datos más equilibrado y variado, podríamos hacer predicciones más precisas y entender mejor qué factores realmente influyen en la calidad de los vinos.

Bibliografía:

Data Science from Scratch, 2nd Edition, Joel Grus

Fundamentals of Data Engineering: Plan and Build Robust Data Systems

