Red Wine Categories

Proyecto de Ciencia de Datos

Realizado por:
Tadeo Raffaeli

Año de Estudio: 2023

-Abstract-

El propósito fundamental de este proyecto es determinar cuáles son las variables más significativas que influyen en la calidad de un vino. Nuestra base de datos contiene una amplia muestra de más de 1,500 vinos, con exclusión de información sobre precios y viñedos para evitar sesgos en nuestra evaluación. Nuestro objetivo principal es crear un recurso que permita evaluar la calidad de los vinos sin depender exclusivamente de la etiqueta o el precio. Este enfoque tiene aplicaciones valiosas tanto para los consumidores como para los vendedores de vinos. La presentación ejecutiva se centrará en nuestros objetivos, metadatos, hipótesis, visualizaciones clave y hallazgos más destacados.

-Etapas del Proyecto-

1. **Análisis de la Base de Datos**: En esta fase inicial, exploraremos la base de datos de más de 1,500 muestras de vinos. Excluiremos información sobre precios y viñedos para evitar sesgos. Nuestro objetivo es comprender la composición de los datos y definir cómo afectan la calidad del vino.

2. **Análisis de Variables**: Llevaremos a cabo un análisis detallado de cada variable en la base de datos. Esto implica examinar sus distribuciones, estadísticas descriptivas y relaciones con la calidad del vino. Investigaremos correlaciones y tendencias para descubrir qué variables podrían ser determinantes en la calidad del vino.

3. **Limpieza de Datos**: En esta etapa, identificaremos y eliminaremos datos irrelevantes o ruidosos que puedan afectar negativamente nuestro análisis. La limpieza nos permitirá trabajar con datos más confiables y precisos.

4. **Modelos Estadísticos de Predicción**: En la fase final, emplearemos modelos estadísticos para predecir la calidad del vino. Utilizaremos los conocimientos adquiridos en las etapas anteriores para desarrollar modelos que puedan evaluar y clasificar la calidad de nuevos vinos, sin depender únicamente de etiquetas o precios.

Estas etapas son fundamentales para alcanzar nuestro objetivo de proporcionar un recurso útil tanto para consumidores como para vendedores de vinos, permitiéndoles tomar decisiones informadas basadas en la verdadera calidad de los vinos.

-Objetivo del Proyecto-

Nuestro principal objetivo es utilizar el análisis de datos y técnicas de aprendizaje automático para desarrollar un sistema de recomendación y toma de decisiones basado en datos para nuestra cadena de tiendas de vinos boutique. Este sistema permitirá:

1. **Selección de Vinos de Alta Calidad**: Identificar las características cruciales que influyen en la calidad de un vino, sin verse influenciados por marcas, precios o bodegas reconocidas. Esto habilitará a la tienda para mantener una oferta de vinos excepcionales que satisfagan las preferencias de sus clientes.

2. **Optimización de la Gestión de Inventario**: Mejorar la gestión del inventario al reconocer qué vinos tienen una demanda más elevada y cuáles requieren ajustes en sus niveles de existencias. Esto garantizará que la tienda siempre cuente con los vinos más populares y evite el exceso de existencias.

3. **Establecimiento de Precios Competitivos**: Determinar precios competitivos para cada vino, basados en su calidad percibida. Esto ayudará a la tienda a maximizar sus márgenes de beneficio y atraer a los clientes que valoran la relación calidad-precio.

En resumen, nuestro proyecto tiene como objetivo empoderar a nuestra cadena de tiendas de vinos boutique con herramientas analíticas que les permitan tomar decisiones informadas y estratégicas en la selección de vinos, la gestión de inventario y la fijación de precios. Esto fortalecerá su posición en la industria del vino, brindando a los clientes una experiencia excepcional y una selección de vinos de alta calidad.

-Contexto Empresarial-

Nuestro proyecto de Data Science está diseñado para satisfacer las necesidades de una cadena de tiendas de vinos boutique que se enfrenta a un entorno empresarial altamente competitivo y en constante evolución. La industria del vino es conocida por su dinamismo y la creciente sofisticación de los consumidores. Nuestro enfoque es proporcionar soluciones que permitan a nuestro cliente destacarse en este mercado.

Hemos compilado una base de datos extensa que contiene información detallada de más de 1,500 vinos tintos, incluyendo su respectiva calidad. A través de un riguroso análisis de datos, nuestro objetivo es discernir las características cruciales que influyen en la calidad del vino. Este conocimiento permitirá a nuestro cliente tomar decisiones estratégicas y basadas en datos.

En términos prácticos, proponemos desarrollar un sistema de recomendación basado en datos que ayudará a nuestra cadena de tiendas a lograr tres metas esenciales:

1. **Selección de Vinos de Alta Calidad**: Identificar las características intrínsecas que definen la calidad de un vino, sin verse condicionados por factores como marcas, precios o bodegas de renombre. Esto permitirá que la tienda ofrezca una selección de vinos excepcionales, que se ajusten a las preferencias de sus clientes.

2. **Optimización de la Gestión de Inventario**: Refinar la gestión del inventario al reconocer cuáles vinos tienen una demanda más fuerte y cuáles requieren ajustes en sus niveles de existencias. Esto asegurará que la tienda mantenga un suministro constante de los vinos más solicitados y evite el exceso de inventario.

3. **Establecimiento de Precios Competitivos**: Determinar precios competitivos para cada vino en función de su calidad percibida. Esto ayudará a la tienda a maximizar sus márgenes de beneficio y atraer a clientes que valoran la relación calidad-precio.

En resumen, nuestro proyecto de Data Science tiene como objetivo empoderar a nuestra cadena de tiendas de vinos boutique con herramientas analíticas que les permitan tomar decisiones informadas y estratégicas. Esto no solo mejorará la satisfacción de los clientes, sino que también fortalecerá su rentabilidad y su posición competitiva en un mercado vitivinícola en constante cambio. Nuestra misión es contribuir a mantener y elevar su reputación como un destino preferido para los amantes del vino, brindando una experiencia excepcional y una selección de vinos de alta calidad.

-Hipótesis-

Hipótesis 1: Relación entre Sulfitos y Calidad del Vino

- **Hipótesis 1a:** Creemos que a medida que aumenta la cantidad de sulfitos en los vinos, la calidad del vino tiende a mejorar. En otras palabras, vinos con más sulfitos se correlacionarán positivamente con calificaciones de calidad más altas.
- **Hipótesis 1b:** Sospechamos que la influencia de la cantidad de sulfitos en la calidad del vino no es uniforme y puede variar según otros factores que influyen en la calidad. Esto incluye el nivel de alcohol, pH, densidad y otros atributos. Por lo tanto, anticipamos que la relación entre sulfitos y calidad será mediada por estas otras características del vino.

Hipótesis 2: Influencia de Factores Químicos en la Calidad del Vino

- **Hipótesis 2a:** Consideramos que ciertos factores químicos, como el nivel de alcohol por litro, pH, densidad, total sulfur dioxide, cloruros y azúcar residual, desempeñarán un papel significativo en la determinación de la calidad del vino. Esto sugiere que, a medida que estos factores químicos varíen, la calidad del vino también variará.
- **Hipótesis 2b:** Proponemos que la calidad del vino no está determinada por una única variable química, sino que es una interacción compleja de múltiples factores químicos. Por lo tanto, esperamos que una combinación de estas variables proporcione una predicción más precisa de la calidad del vino en comparación con el enfoque en una sola variable química.

Hipótesis 3: Relación entre Características Sensoriales y Calidad del Vino

- **Hipótesis 3a:** Suponemos que las características sensoriales, como citric acid, volatile acidity y fixed acidity, también se relacionarán con la calidad del vino. Esto significa que las calificaciones de calidad más altas estarán asociadas con niveles específicos de estas características sensoriales.
- **Hipótesis 3b:** Consideramos que las características sensoriales pueden agregar una dimensión adicional en la determinación de la calidad del vino, y pueden capturar aspectos que no son completamente abordados por las variables químicas. Esto implica que las características sensoriales podrían proporcionar información única sobre la calidad del vino.

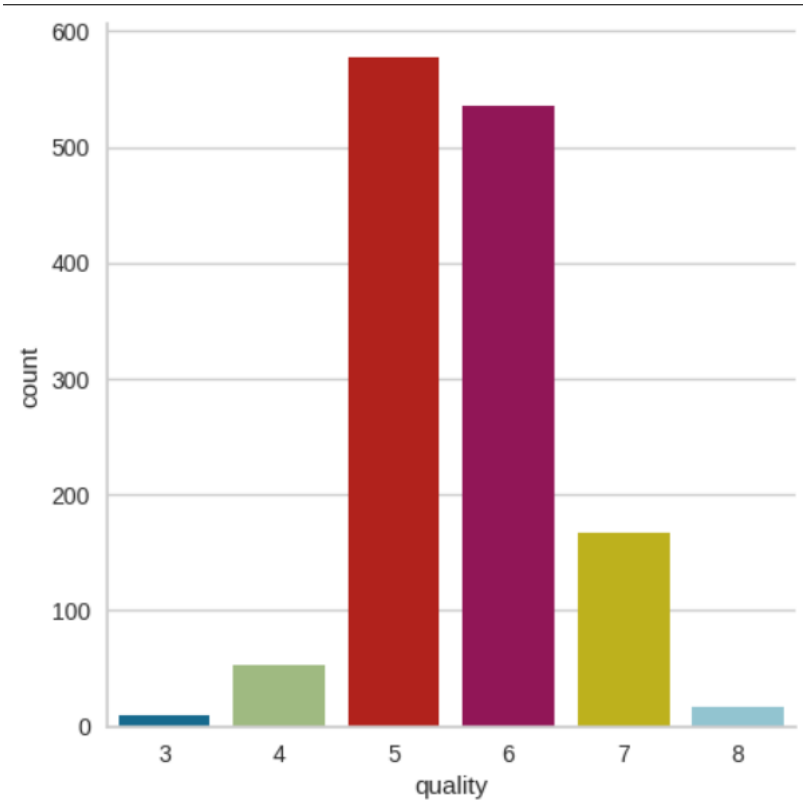
Exploración de Datos (EDA) y Limpieza de Datos Inicial

El proceso de Exploración de Datos (EDA) comenzó con una fase crucial: la limpieza de datos. Esta etapa es fundamental para garantizar la calidad y confiabilidad de nuestros análisis. A continuación, describiremos cómo abordamos esta fase inicial del EDA.

Eliminación de Datos Duplicados

Uno de los primeros pasos fue identificar y eliminar los datos duplicados en nuestra base de datos. La presencia de datos duplicados podría generar sesgos y afectar negativamente la precisión de nuestros análisis. Tras este procedimiento, estábamos seguros de que estábamos trabajando con datos únicos y sin repeticiones.

Análisis de la Distribución de Categorías de Calidad



Para comprender mejor la distribución de las categorías de calidad de los vinos en nuestro conjunto de datos, realizamos un conteo de cada categoría. Esto nos proporcionó información valiosa sobre cómo estaban representadas las diferentes calidades de vino en nuestra muestra. Los resultados revelaron que teníamos una variada gama de calidades, desde la categoría 3 hasta la categoría 8.

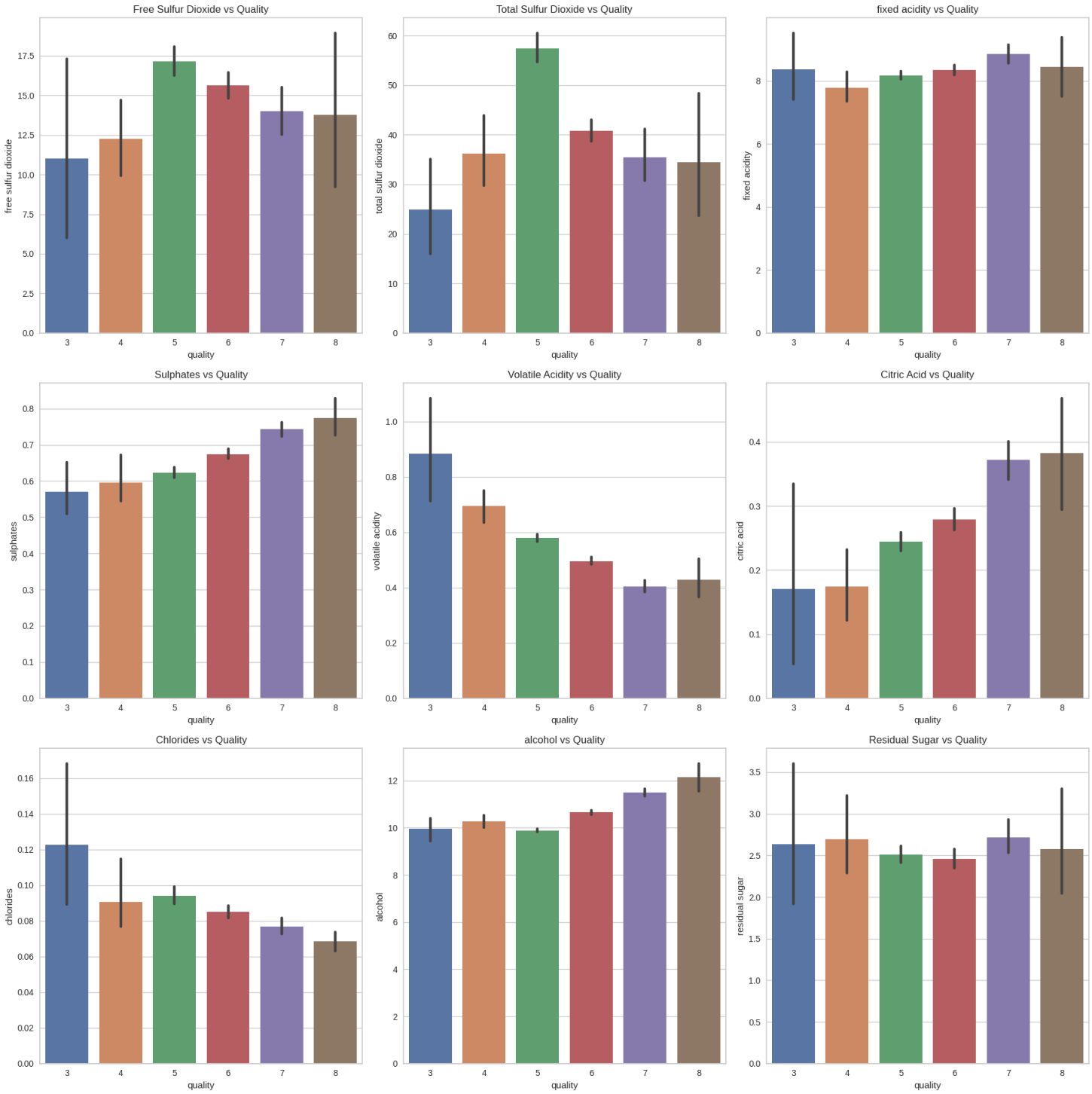
Los hallazgos de este análisis mostraron que no todas las categorías estaban igualmente representadas. Por ejemplo, las categorías 5 y 6 tenían una mayor cantidad de muestras en

comparación con otras categorías. Esto nos indicó que teníamos una distribución desigual de datos en términos de calidad del vino, y que debíamos abordar este desequilibrio al avanzar en nuestro análisis.

Esta información también nos permitió comprender que algunas categorías podrían tener una cantidad insuficiente de datos para realizar análisis significativos, lo que requeriría una estrategia para equilibrar las categorías y garantizar resultados más robustos.

-Comparación de Variables y Calidad de Vino-

La siguiente fase de nuestro Análisis de Datos Exploratorio (EDA) implicó una comparación detallada de cada variable con las categorías de calidad de vino. Aquí, presentamos un resumen de los resultados obtenidos de esta evaluación visual:



- **Niveles de Dióxido de Azufre**: Tanto el dióxido de azufre libre como el total no muestran una correlación directa con la calidad de los vinos. En otras palabras, no es evidente que una mayor o menor cantidad de dióxido de azufre sea un indicador claro de calidad.
- **Acidez Fija**: La acidez fija parece tener valores bastante similares en vinos de diversas calidades, lo que sugiere que no es un factor determinante en la diferenciación de la calidad del vino.
- **Sulfitos**: Los sulfitos muestran una clara relación positiva con la calidad del vino, lo que significa que, en general, a medida que aumenta la calidad, también lo hace la cantidad de sulfitos. Sin embargo, no son la única variable influyente en la calidad.
- **Acidez Volátil**: La acidez volátil tiende a disminuir en vinos de mejor calidad. Esta disminución es evidente, aunque hay algunas excepciones que podrían deberse a la falta de datos en la categoría de calidad 8. Esto sugiere que la acidez volátil podría ser una variable relevante para determinar la calidad del vino.
- **Ácido Cítrico**: El ácido cítrico muestra una relación positiva con la calidad, lo que tiene sentido ya que esta característica aporta sabores frutales y frescos que se asocian con vinos de mayor calidad.
- **Cloruros de Sodio**: La cantidad de cloruros de sodio tiende a disminuir a medida que aumenta la calidad del vino. Esta disminución contribuye a evitar un sabor salado no deseado, lo que es coherente con la búsqueda de vinos de alta calidad.
- **Cantidad de Alcohol**: La cantidad de alcohol en los vinos tiene una leve tendencia al alza a medida que aumenta la calidad. Esto sugiere que, en general, los vinos de mejor calidad pueden tener un contenido alcohólico ligeramente mayor.

Estos análisis preliminares nos proporcionan información valiosa sobre cómo ciertas variables pueden influir en la calidad del vino. Sin embargo, continuaremos explorando y analizando más a fondo nuestros datos para obtener conclusiones más sólidas y reveladoras. Este es solo el primer paso en nuestro viaje hacia la comprensión y mejora de la calidad de los vinos.

-Evaluación de la Relación entre la Cantidad de Sulfitos y la Calidad del Vino-

Para evaluar la relación entre la cantidad de sulfitos en los vinos y su calidad, llevamos a cabo un análisis de correlación utilizando el coeficiente de correlación de Pearson. Nuestras hipótesis iniciales planteaban la suposición de que una mayor cantidad de sulfitos estaría asociada con una calificación de calidad más alta, y estas hipótesis se formularon como sigue:

Hipótesis 1a: A mayor cantidad de sulfitos en los vinos, se observará una tendencia hacia una calificación de calidad más alta.

Hipótesis 1b: La influencia de la cantidad de sulfitos en la calidad del vino variará según las características del vino, como su nivel de alcohol, pH, densidad y otros factores.

El análisis arrojó un coeficiente de correlación de Pearson de aproximadamente -0.4 entre la cantidad de sulfitos y la calidad del vino. Este valor indica una correlación negativa, pero no particularmente fuerte, entre ambas variables. Si bien la correlación es estadísticamente significativa ($p < 0.05$), su magnitud es moderada.

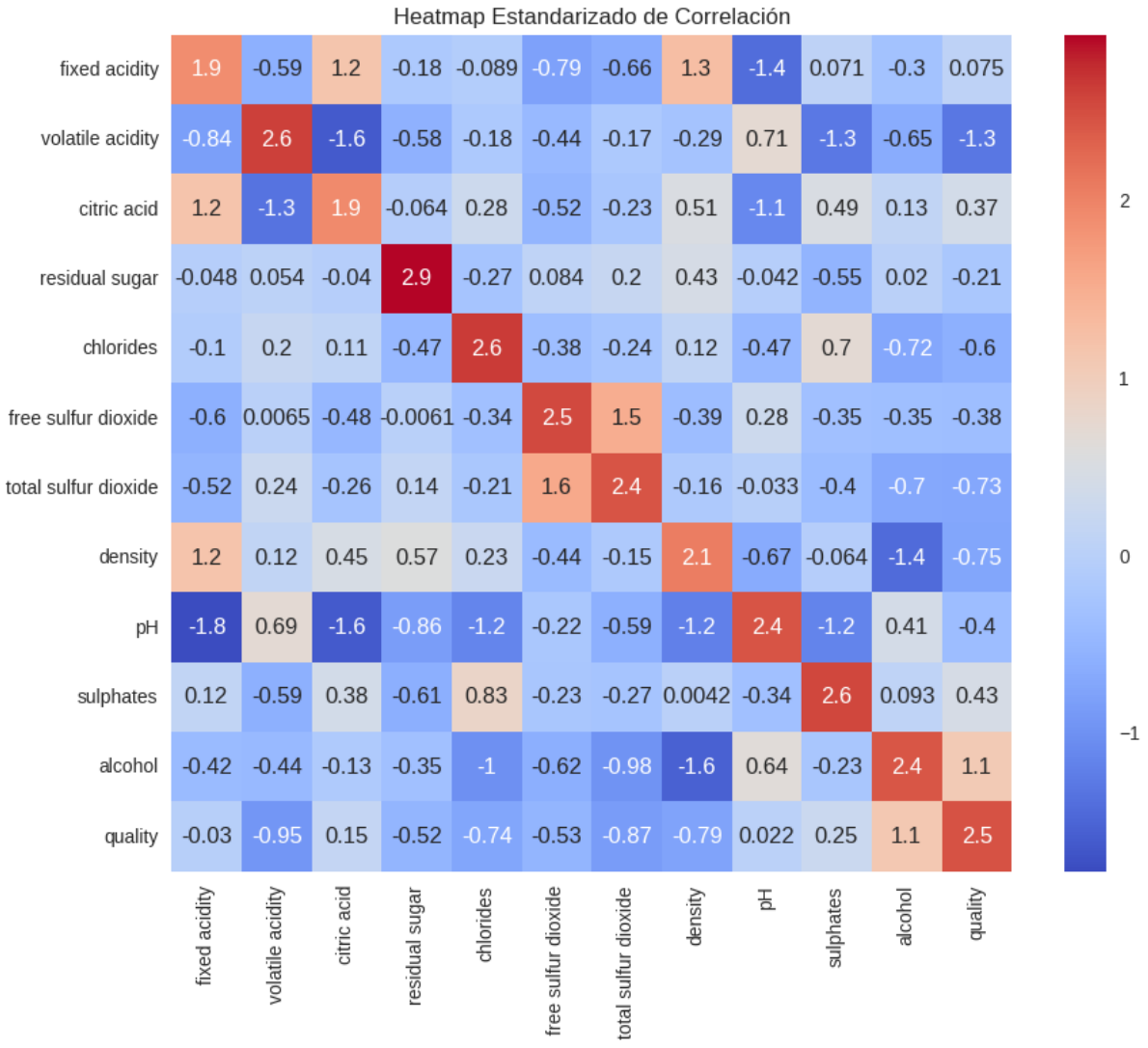
En resumen, nuestros hallazgos no respaldan por completo la primera hipótesis, ya que la correlación no es muy fuerte. Esto sugiere que la cantidad de sulfitos en los vinos puede estar relacionada con la calidad, pero no es el único factor determinante. La segunda hipótesis, que plantea una interacción con otras características del vino, sigue siendo válida y requerirá un análisis más detallado en futuras etapas de este estudio. Nuestro enfoque se centrará en la exploración de cómo otros atributos del vino pueden influir en esta relación.

-Heatmap con Datos Estandarizados: Explorando las Correlaciones-

En nuestro análisis, utilizamos un heatmap que representa las correlaciones entre las variables, pero con un enfoque especial: todos los datos se han estandarizado. Esto implica que todas las variables tienen una media de cero y una desviación estándar de uno. Este proceso de estandarización nos permite comparar las correlaciones entre las variables sin verse afectadas por sus diferentes escalas y unidades de medida.

Los colores en el heatmap juegan un papel crucial. En este contexto, los tonos más oscuros indican una correlación negativa entre las variables, lo que significa que cuando una variable aumenta, la otra tiende a disminuir. Por otro lado, los tonos más claros en el heatmap sugieren una correlación positiva, lo que implica que cuando una variable aumenta, la otra también tiende a aumentar.

Este análisis nos proporciona una visión general de las relaciones y patrones de correlación entre las diferentes características de los vinos, lo que es fundamental para entender qué variables pueden influir significativamente en la calidad de los mismos. Este conocimiento nos ayudará a tomar decisiones informadas en etapas posteriores del proyecto.



-Eliminación de Variable: Fixed Acidity-

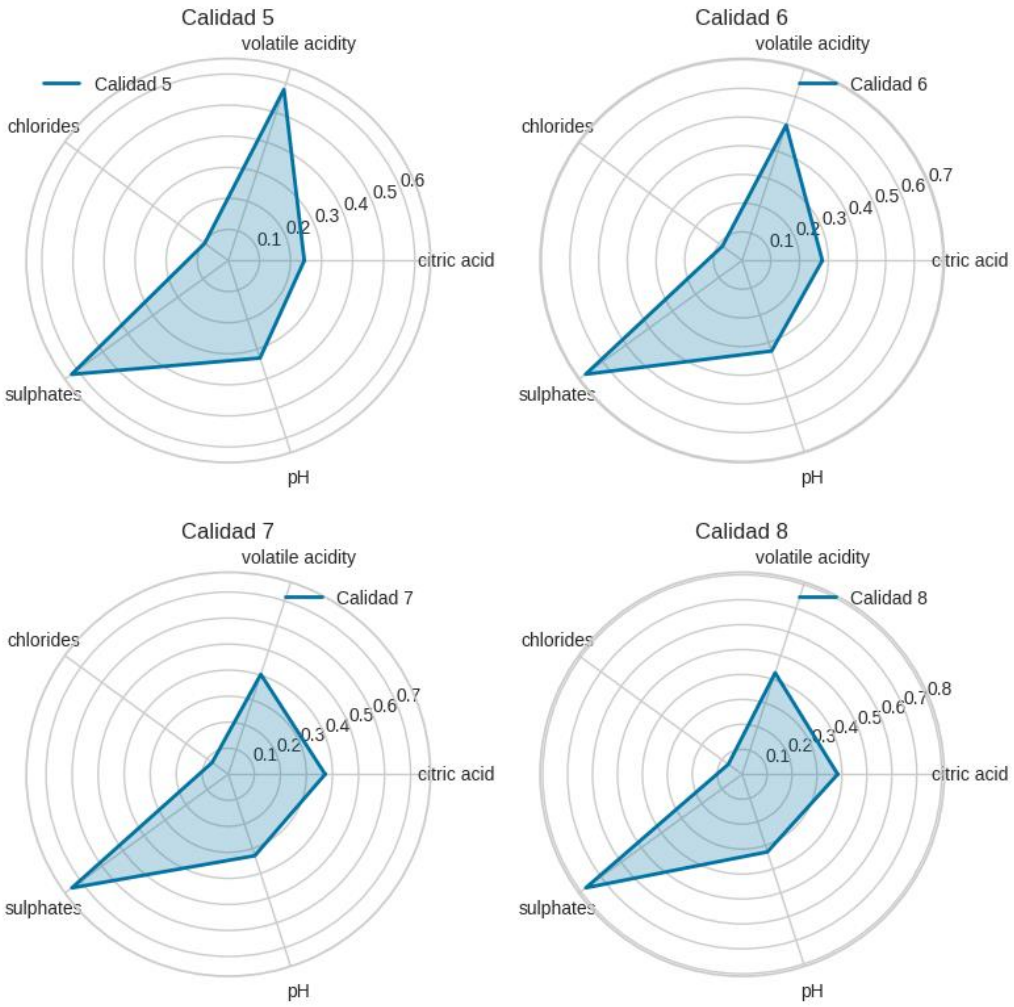
En la exploración de nuestro conjunto de datos, identificamos un desafío común en el análisis de datos: tratar con variables que presentan una fuerte correlación entre sí. La pregunta que se plantea es si deberíamos eliminar algunas de estas variables para simplificar nuestro modelo. Sin embargo, esta decisión no debe tomarse a la ligera, ya que eliminar variables importantes podría sesgar nuestro análisis.

En nuestro caso, examinamos la variable "Fixed Acidity" y notamos que tiene una correlación relativamente baja con otras variables, especialmente en comparación con las correlaciones sólidas observadas entre otras variables. Su correlación más notable es con la densidad, y su valor es de 0.32, lo cual puede considerarse moderado en comparación.

Dado este hallazgo y con el objetivo de evitar sesgos potenciales en nuestro análisis, hemos optado por crear un nuevo DataFrame que excluya la variable "Fixed Acidity." Este paso nos permitirá comparar modelos que incluyen y excluyen esta variable para determinar cuál de ellos ofrece una mejor capacidad de predicción de la calidad del vino. Esta decisión es fundamental en nuestra búsqueda de comprender las variables más influyentes en la calidad del vino.

-Análisis de los Gráficos de Radar-

Los gráficos de radar ofrecen una valiosa perspectiva sobre cómo ciertas variables impactan en la calidad del vino. Estas visualizaciones revelan tendencias significativas que arrojan luz sobre la influencia de diferentes factores.



Observamos que a medida que la calidad del vino mejora, la acidez volátil tiende a disminuir. Este hallazgo tiene sentido, ya que un exceso de acidez volátil puede resultar en un sabor no deseado en el vino. La disminución de la acidez volátil se alinea con la búsqueda de la calidad.

Por otro lado, el ácido cítrico muestra una tendencia opuesta. A medida que la calidad del vino aumenta, se observa un ligero aumento en las cantidades de ácido cítrico. Esto sugiere que una mayor presencia de ácido cítrico, que aporta frescura y sabor, se percibe como una característica positiva por los catadores.

El pH, por otro lado, parece ejercer una influencia limitada en la calidad del vino. Los vinos de alta calidad generalmente mantienen valores de pH dentro del rango estándar de 3 a 4. Solo los valores

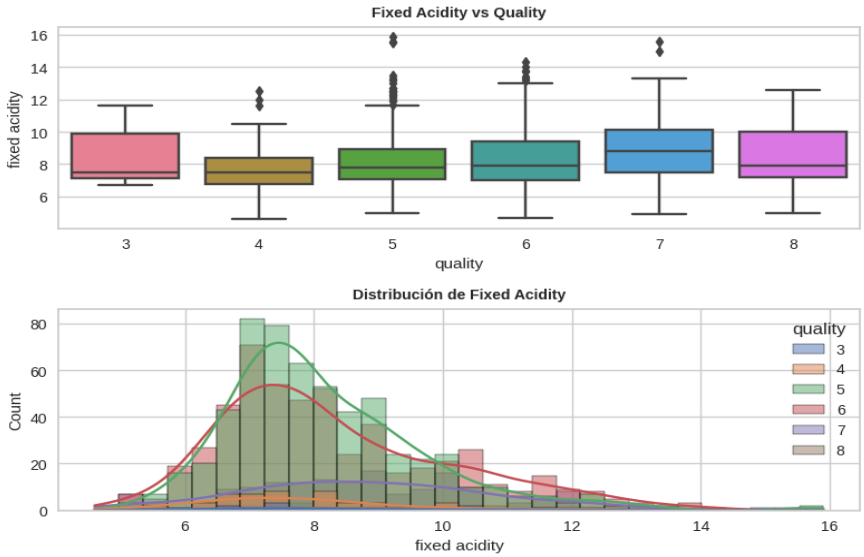
extremos, cercanos a 0-1 o 8-14, afectan significativamente el sabor, mientras que los valores dentro del rango estándar no parecen influir de manera discernible.

En lo que respecta a los sulfitos, los gráficos sugieren que no experimentan cambios drásticos en relación con la calidad del vino. Esto podría deberse a que los sulfitos son un aditivo económico que actúa como antimicrobiano y antioxidante. Su influencia no parece variar significativamente en función de la calidad del vino. Estos hallazgos son fundamentales para comprender cómo los componentes químicos afectan la percepción de calidad del vino.

-Análisis de Outliers y Distribución de Variables-

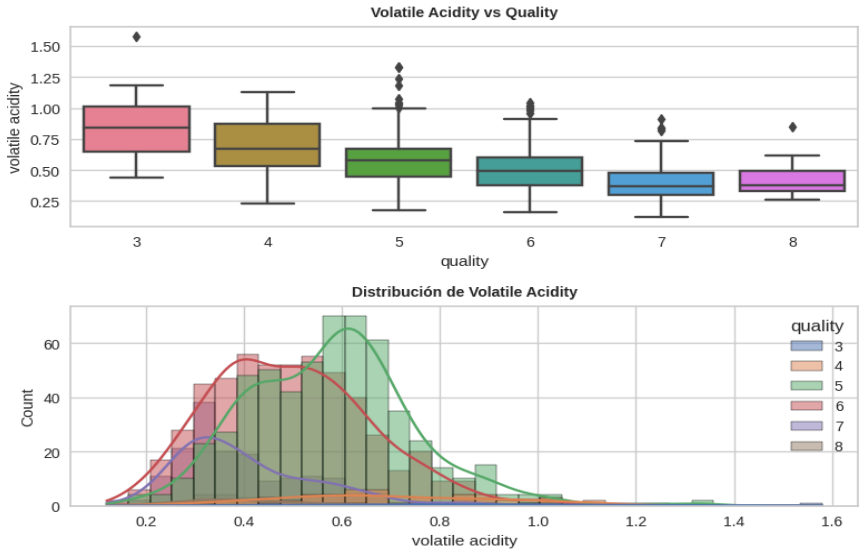
Durante nuestra exploración, hemos evaluado la distribución y presencia de valores atípicos en cada una de las variables del conjunto de datos. Aquí están los hallazgos clave para cada variable:

Fixed Acidity:



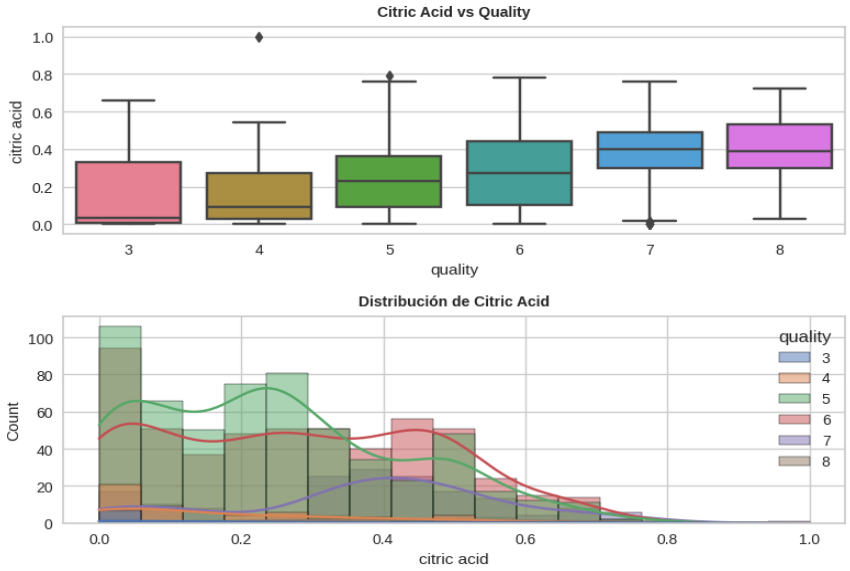
Esta variable muestra una ligera asimetría hacia la izquierda, lo que indica que tiende a concentrarse en valores más altos, pero esta asimetría es muy leve y no afecta significativamente la normalidad de la distribución.

Volatile Acidity:



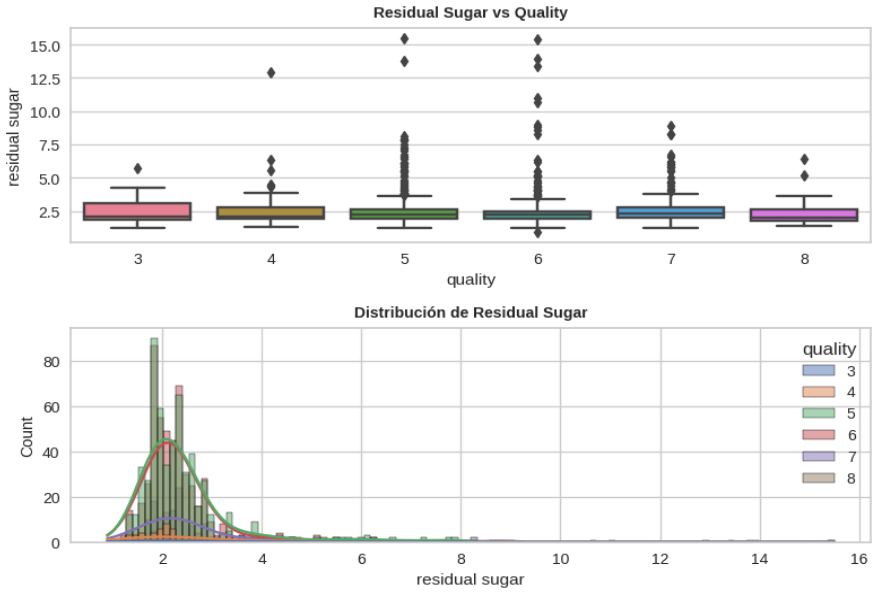
Al igual que Fixed Acidity, Volatile Acidity también presenta una ligera asimetría hacia la izquierda, pero la distribución se acerca a una forma normal.

Citric Acid:



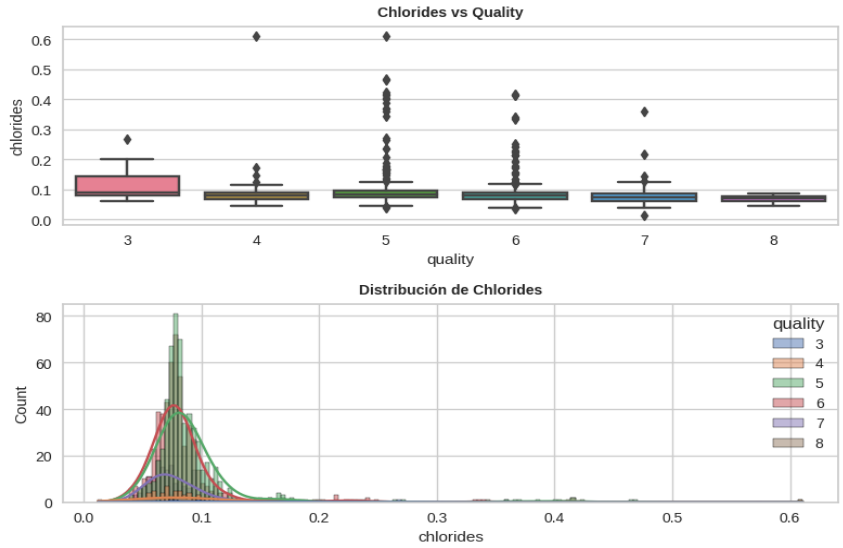
En contraste, el ácido cítrico muestra una clara asimetría hacia la derecha, indicando que la mayoría de las muestras tienen valores más bajos de este componente.

Residual Sugar:



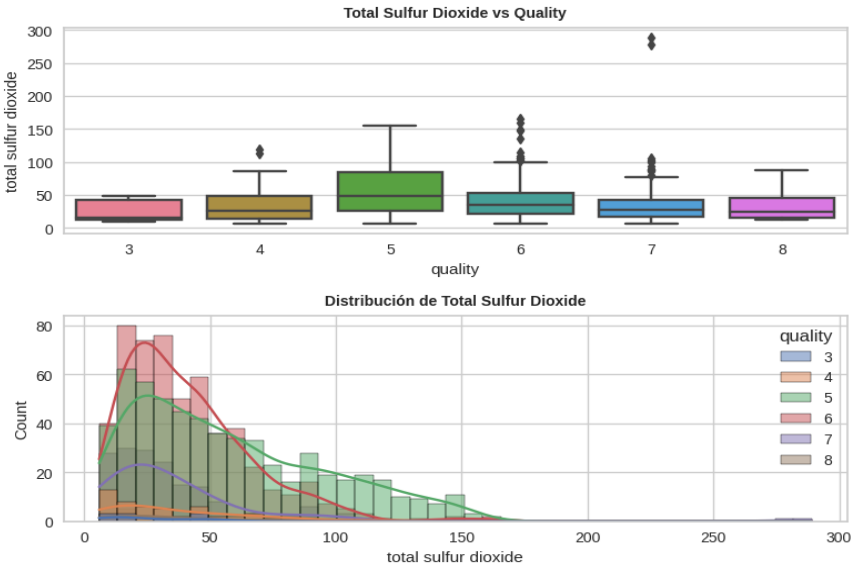
La variable Residual Sugar muestra una distribución altamente sesgada hacia la derecha debido a la presencia de numerosos valores atípicos, especialmente en las categorías 5 y 6.

Chlorides:



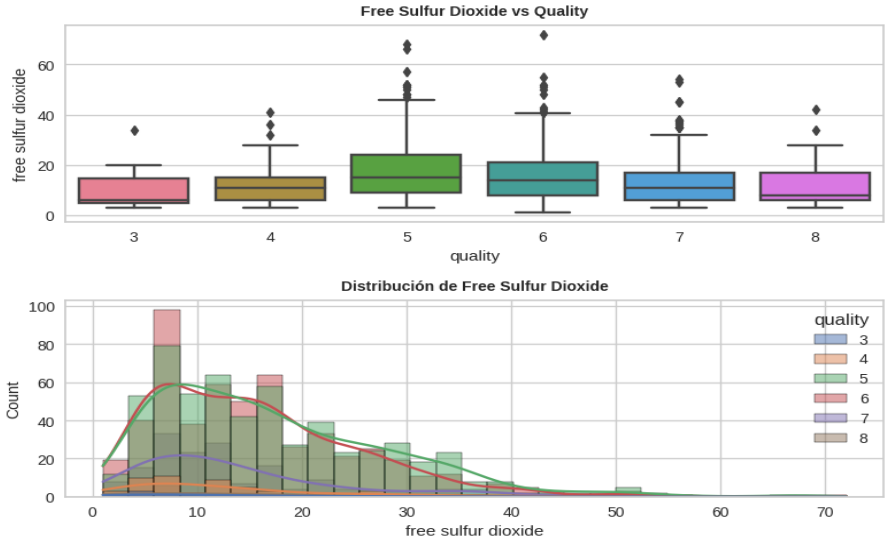
Similar a Residual Sugar, Chlorides también muestra una distribución asimétrica hacia la derecha debido a la presencia de valores atípicos, pero la distribución se mantiene relativamente simétrica.

Total Sulfur Dioxide:



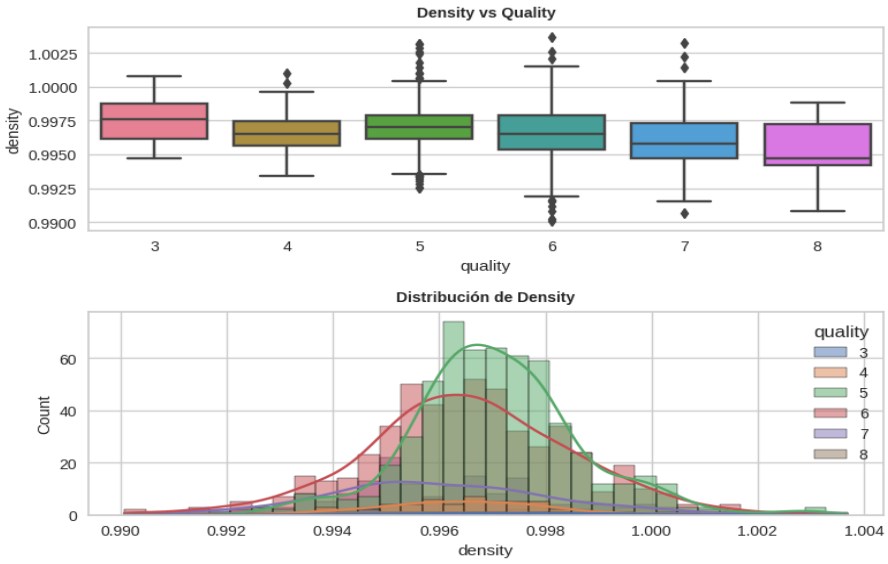
Al igual que el ácido cítrico, Total Sulfur Dioxide muestra una asimetría pronunciada hacia la derecha, acentuada por la existencia de valores atípicos, especialmente en las categorías 5 y 6.

Sulfuros de Dióxido Libres (Free sulfur dioxide):



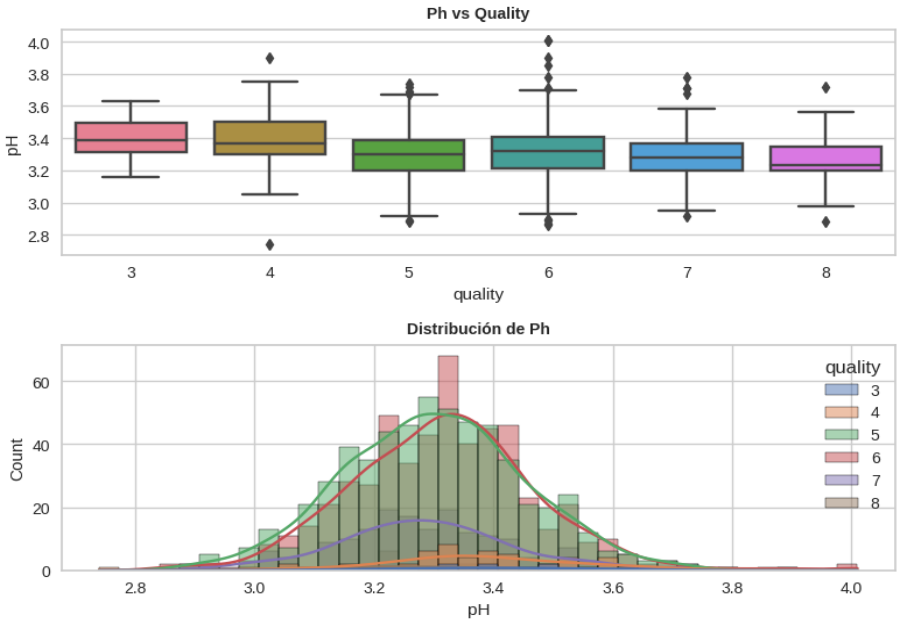
Esta variable parece carecer de valores atípicos significativos, pero su distribución muestra una marcada asimetría hacia la derecha.

Densidad:



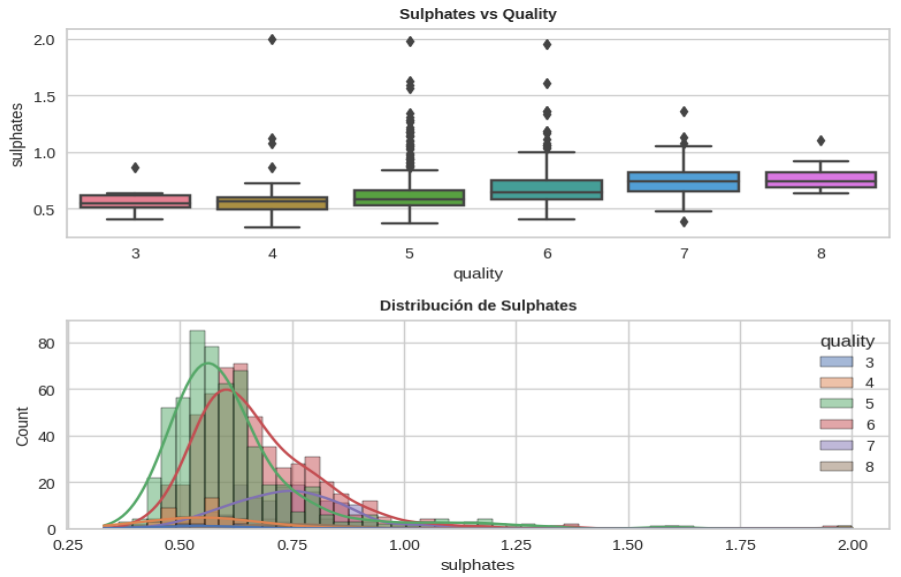
La Densidad es una de las variables que parece seguir una distribución normal, y la ausencia de valores atípicos contribuye a su simetría.

pH:



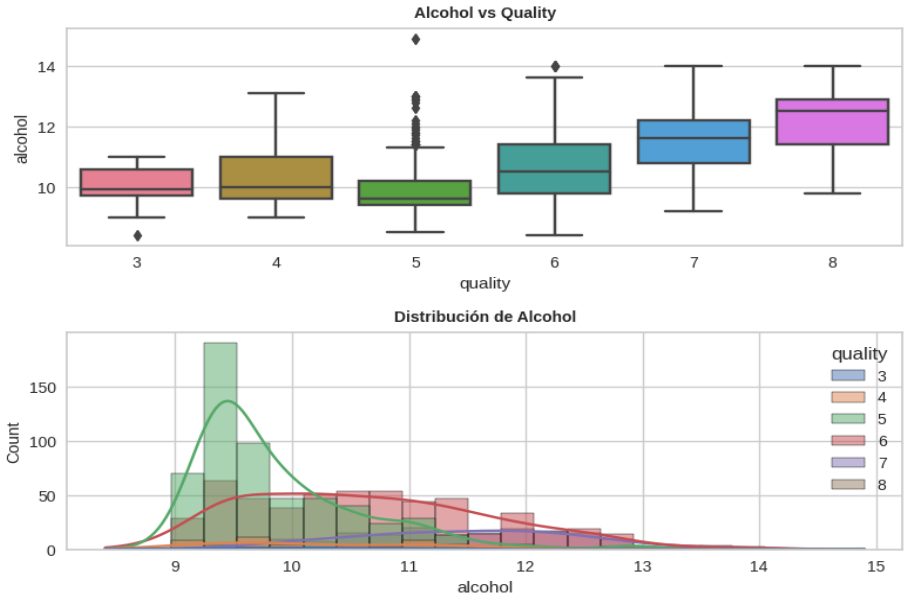
El pH es similar a la Densidad en términos de distribución, aunque muestra una pequeña cantidad de valores atípicos, lo que resulta en una leve asimetría hacia la derecha.

Sulfitos:



A pesar de su distribución aparentemente normal, Sulfitos también presenta una asimetría hacia la derecha debido a la alta cantidad de valores atípicos.

Alcohol:



Alcohol es la primera variable que muestra una distribución marcadamente asimétrica y parece tener una gran cantidad de valores atípicos.

Estos hallazgos son esenciales para comprender la distribución de las variables en nuestros datos y nos proporcionan información valiosa para futuros análisis y modelos. Las asimetrías y la presencia de valores atípicos deben considerarse al seleccionar las técnicas de análisis adecuadas.

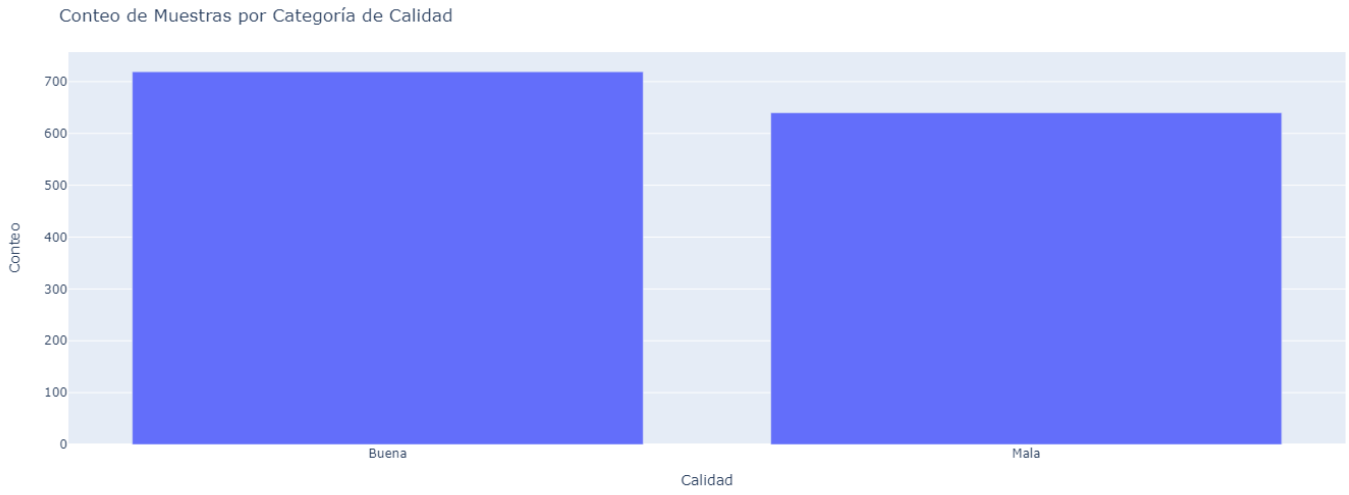
-Enfoque en 2 Categorías de Calidad-

En este punto, hemos tomado una decisión estratégica para simplificar y mejorar nuestro análisis. Originalmente, estábamos trabajando con 5 categorías de calidad de vino, pero notamos un desequilibrio en la cantidad de muestras, especialmente en la categoría de calidad 8. Esto podría sesgar nuestros resultados.

Por lo tanto, hemos optado por reagrupar las categorías de calidad en dos grupos más amplios: "Mala" y "Buena". Las categorías ahora se definen de la siguiente manera: "Mala" abarca las calidades 3, 4 y 5, mientras que "Buena" incluye las calidades 6, 7 y 8.

Esta decisión se basa en el entendimiento de que, según el autor y propietario de la base de datos, los vinos con una calificación de alrededor de 6.5 se consideran "buenos". La consolidación de estas categorías de calidad aborda el desequilibrio en el tamaño de las categorías y permite un análisis más equitativo. Además, facilitará la efectividad de nuestros modelos de aprendizaje automático al clasificar vinos en estas dos categorías distintas. Esta simplificación no solo mejora la precisión de nuestros análisis, sino que también facilita la toma de decisiones basadas en la calidad del vino.

Esto nos devolvió esencialmente dos categorías, la Mala, con 744 muestras, y la Buena, con 855 muestras, algo mucho más balanceado.



-Limpieza de Outliers-

En nuestra exploración de datos, abordamos el desafío de los valores atípicos mediante la eliminación de los mismos. Inicialmente, seguimos el enfoque tradicional, que consiste en eliminar los valores que se encuentran fuera del rango de ±2 veces el rango intercuartil (IQR). Sin embargo, observamos que esto resultaba en la eliminación de un número significativamente mayor de valores de lo esperado.

Para evitar una pérdida excesiva de datos y garantizar una muestra representativa, decidimos ajustar nuestros criterios y utilizar un rango más amplio, de ±2 veces el IQR, para definir los valores atípicos. Este enfoque nos brindó un margen adicional y nos permitió conservar más datos, especialmente aquellos relacionados con vinos de mejor calidad, que a menudo presentan niveles más altos de sulfitos. La decisión de realizar esta limpieza rigurosa se tomó para evitar sesgar nuestro análisis hacia las categorías de menor calidad y mantener una muestra más equilibrada de vinos de alta calidad.

Este proceso de limpieza nos permitió mejorar la calidad de nuestros datos al eliminar valores que podrían haber distorsionado nuestro análisis. Sin embargo, siempre debemos recordar que la eliminación de valores atípicos debe basarse en una comprensión sólida del dominio y en los objetivos del análisis. En este caso, nuestro objetivo principal es comprender la relación entre las variables y la calidad del vino y garantizar que nuestros modelos de aprendizaje automático funcionen de manera efectiva.

Finalmente llegamos a construir dos dataFrames, uno de cada categoría, sin Outliers, lo cual convirtió nuestras muestras en 507 para Mala y 614 para la Buena.