Introducción

Enunciado

**Objetivo**

Este proyecto tiene como objetivo consolidar conocimientos sobre la manipulación de datos con Pandas, el manejo de bases de datos relacionales (SQLite) y la preparación de los datos para una posible inserción en bases de datos no relacionales (MongoDB). El tiempo estimado de trabajo es de 2-3 horas por alumno.

Desarrollo

Instrucciones

Sin límite de tiempo

Desarrollo

**Instrucciones detalladas**

**1. Descarga de datos**

- Descarga los datasets para vinos tintos y blancos desde la siguiente URL: [Wine Quality Dataset]([http://archive.ics.uci.edu/dataset/186/wine+quality](file:///C:\Users\MBITSC~1\AppData\Local\Temp\msohtmlclip1\01\clip_filelist.xml)). - Asegúrate de que ambos datasets se encuentren disponibles en tu entorno de trabajo como archivos CSV. - Idealmente, intenta llevarlo a cabo de manera programática para no tener que hacer operaciones manualmente.

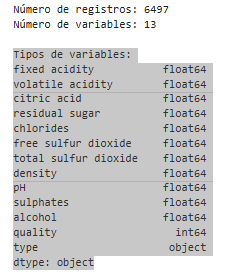
**2. Combinar los datos**

- Usa Pandas para cargar ambos datasets en dataframes en memoria. - Combina los datos en un único dataframe añadiendo una columna adicional que indique el tipo de vino (`red` o `white`). - ¿Cuántos registros tenemos? ¿Cuántas variables y de qué tipo?

Combined-wines tiene 6497 registros (filas)

Red 1599

White 4898

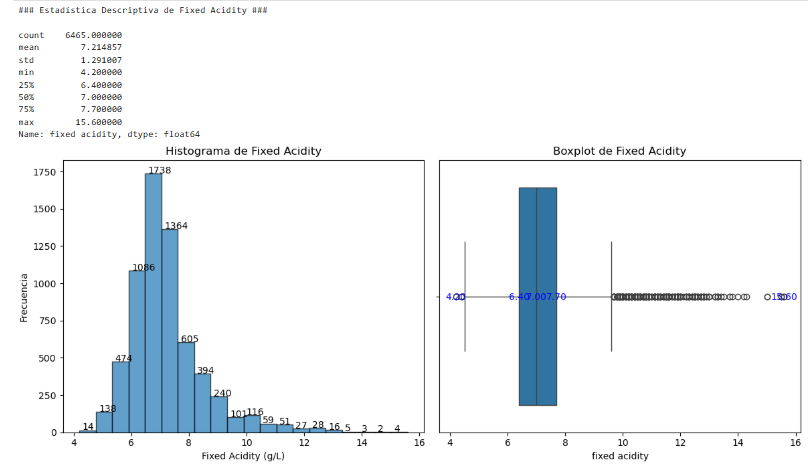


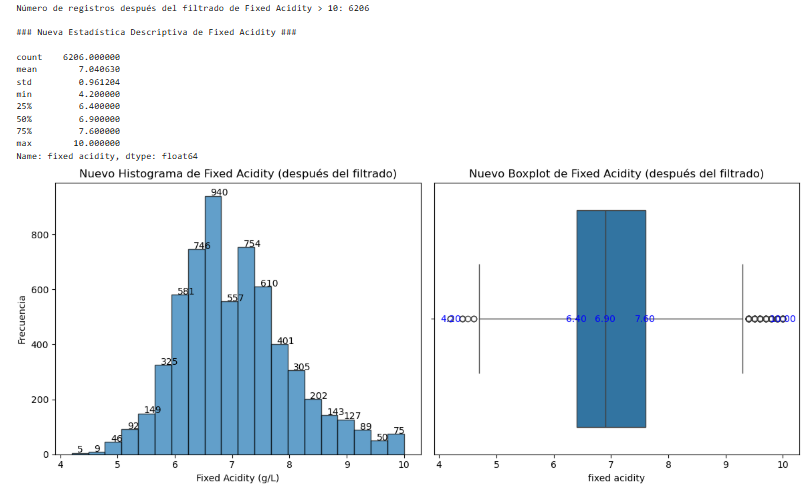
**3. Filtrar atípicos y manejar datos ausentes**

- Realiza un análisis estadístico o inspección visual de cada columna numérica para identificar valores atípicos. - Usa Pandas para filtrar y eliminar los datos atípicos y los valores ausentes. Explica en tu entrega qué criterios utilizaste para identificar los atípicos.

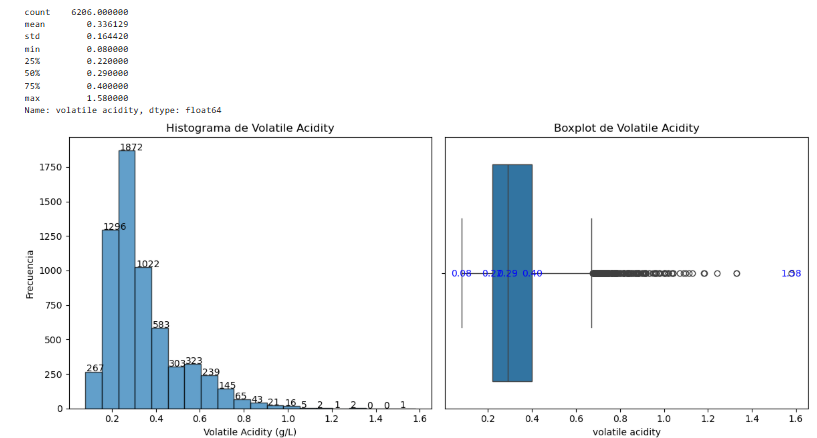
Primera revisión NO hay valores NULL

Usamos la función dataframe.isnull().sum() si nos da 0 es que no ha encontrado nada vacio.

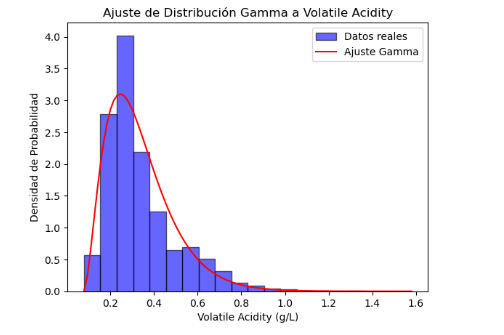




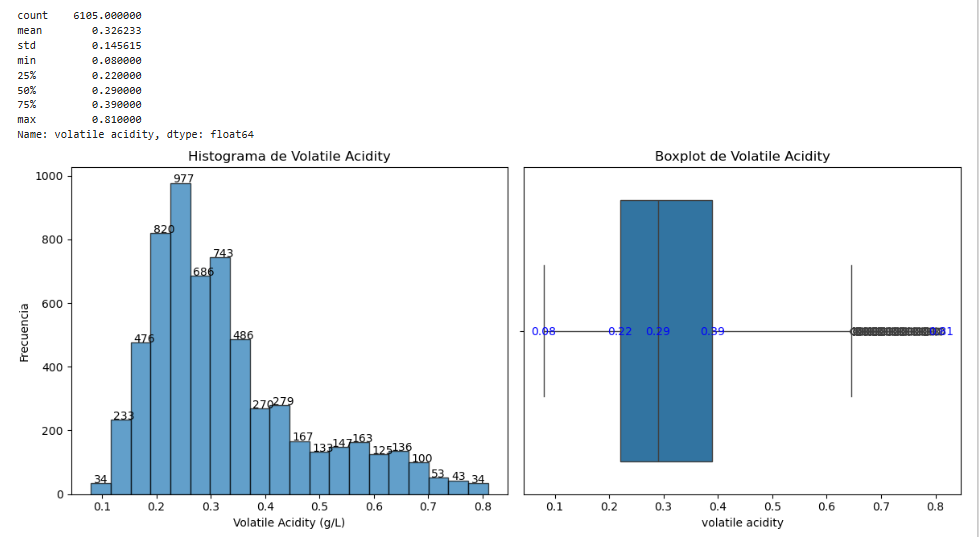
Sigueinte variable Volatile Acidity



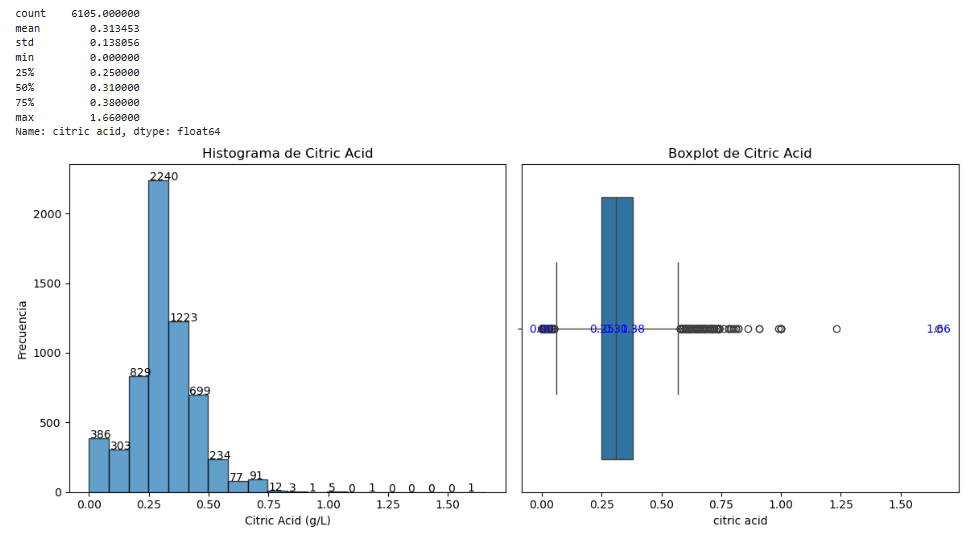
Revisamos y pensamos que podría ser más una Gamma



Se hace un Test de P valor y resulta que no es una Gamma, pero como ya habíamos filtrado hemos “perdido 105 registros” de la cola por la derecha.

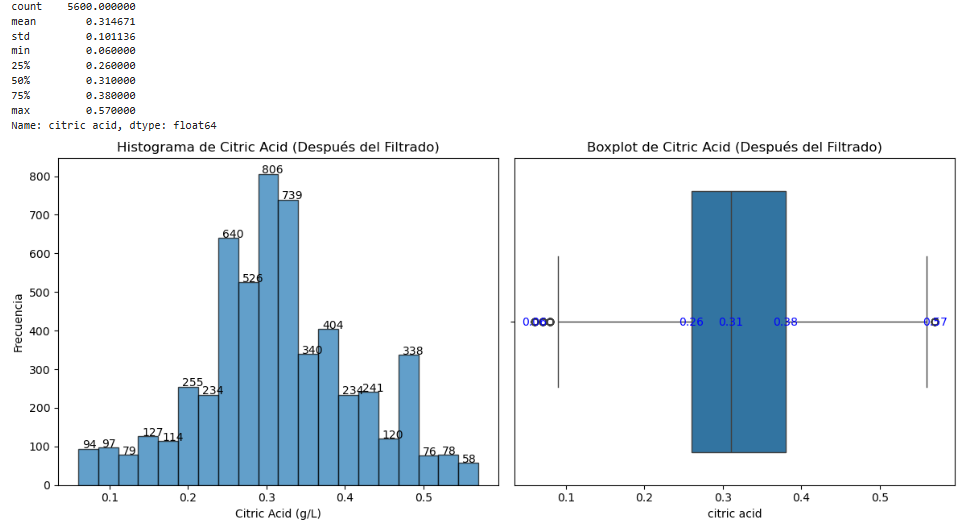


Ahora continuamos con la siguiente variable

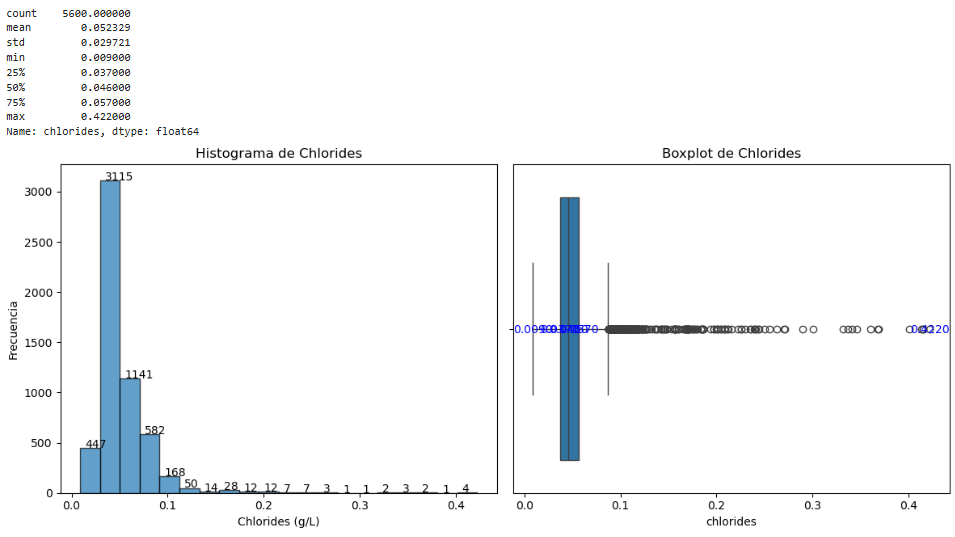


En este caso hemos comprobado antes si era un Normal, y no lo es.

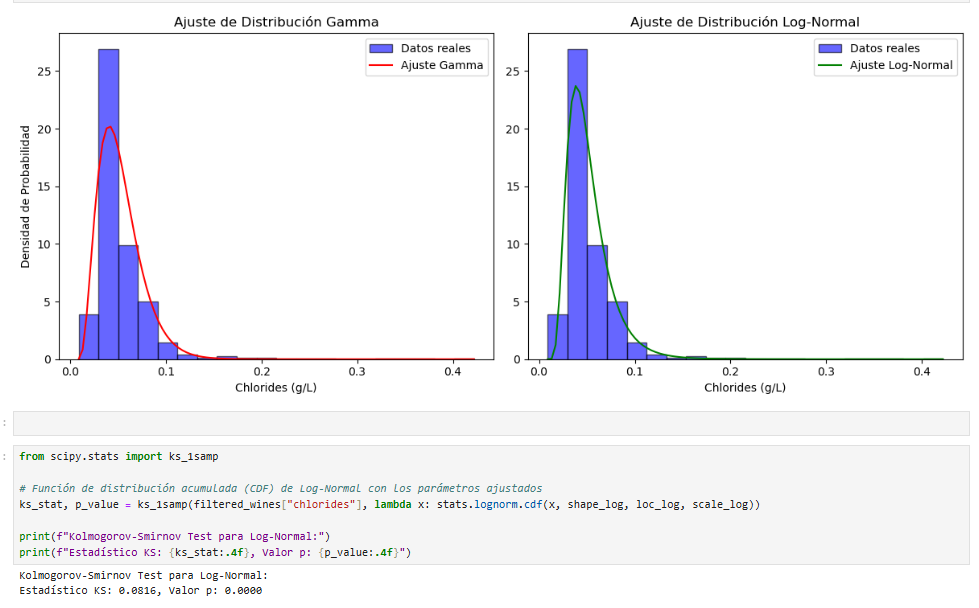
Por ello decidimos usar el método IQR para filtrar datos, después de filtrar nos quedamos con 5,600 (292 por abajo y 213 por arriba se han eliminado)



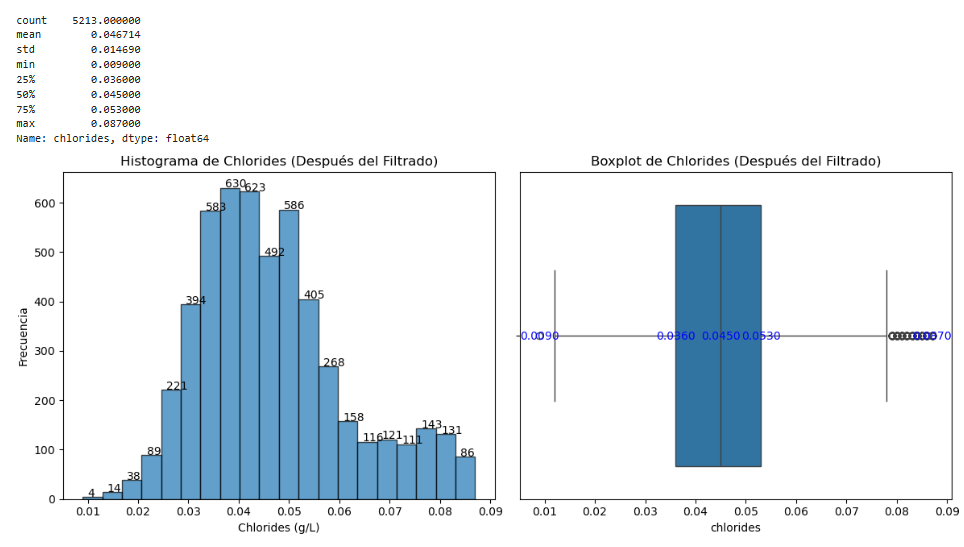
Nos vamos a la siguiente variable Residual Sugar



Pensamos que podría ser otra distribución una Gamma o una Log Normal

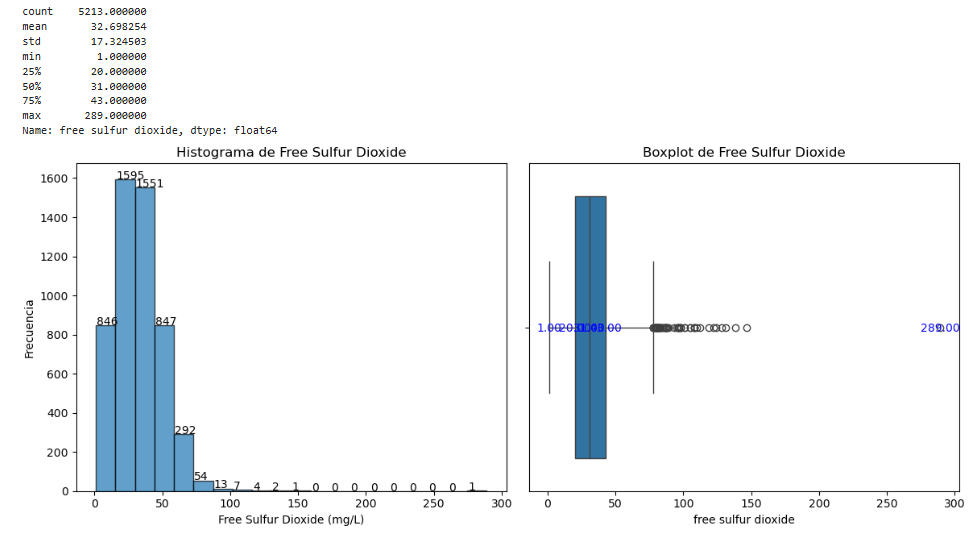


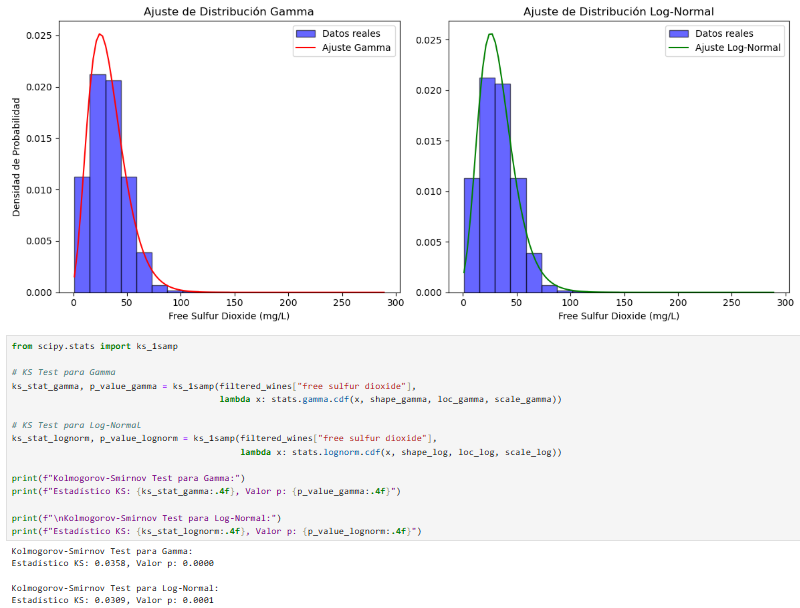
Pero como vemos en el test no podemos aceptarlo, por lo que usaremos el método IQ para filtrar.



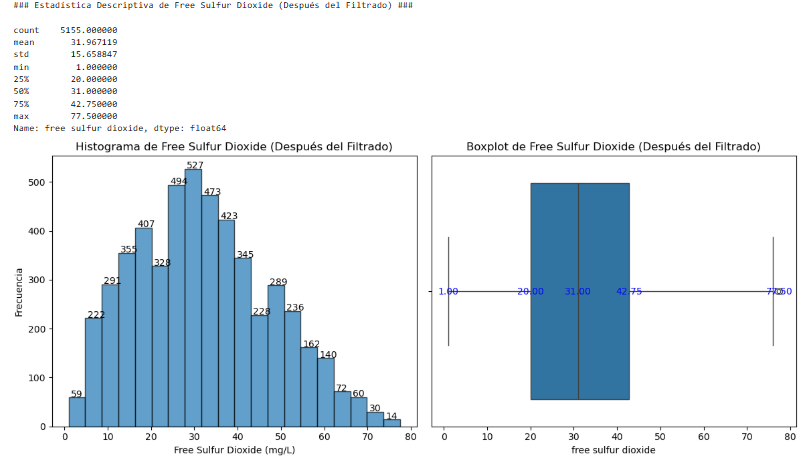
Buen ajuste con la nueva curva, hemos quitado unos 400 datos

Para la siguiente Variable Sulfur Dioxinde



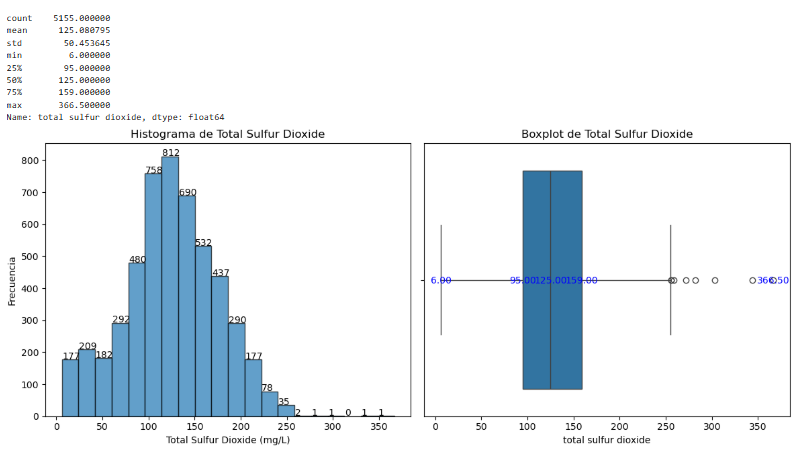


Como no se ajustan estas distribuciones, toca aplicar el IQR

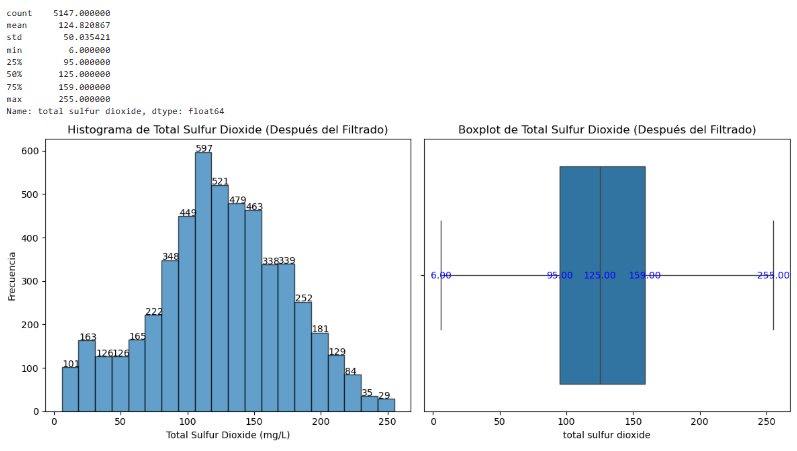


Parece ahora más un distribución normal.

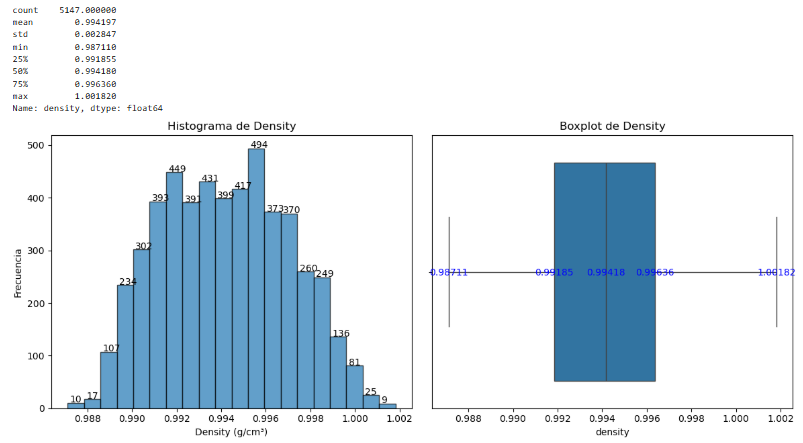
En la siguiente variable Total Sulfur Dioxide



Aplicamos directamente un IQR

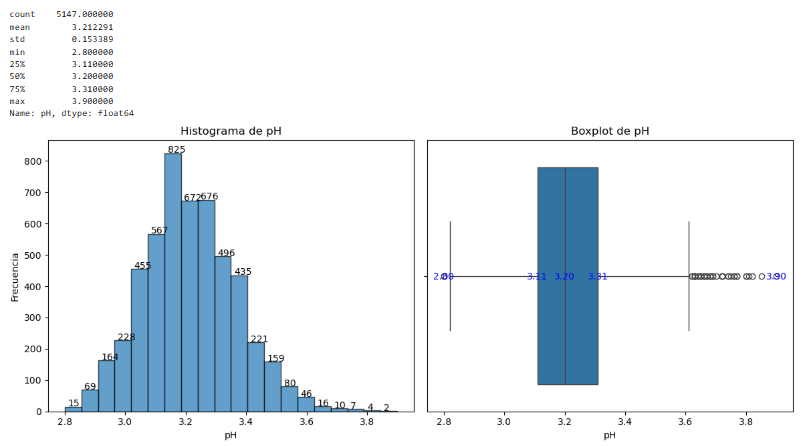


La siguiente variable es Density

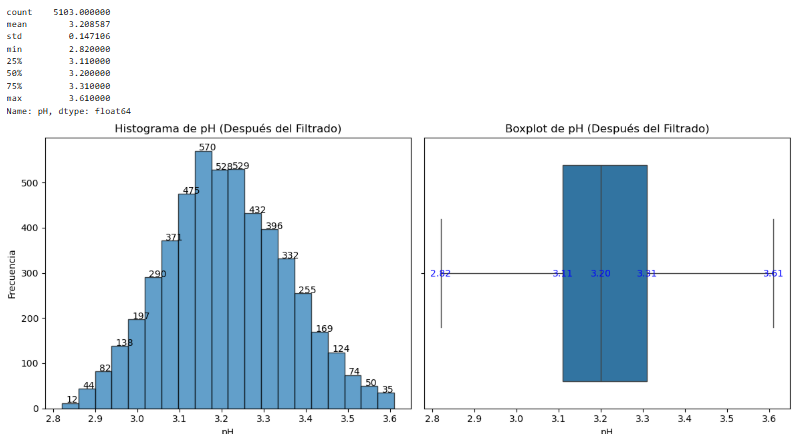


Y en mi opinión no tiene valores que filtrar.

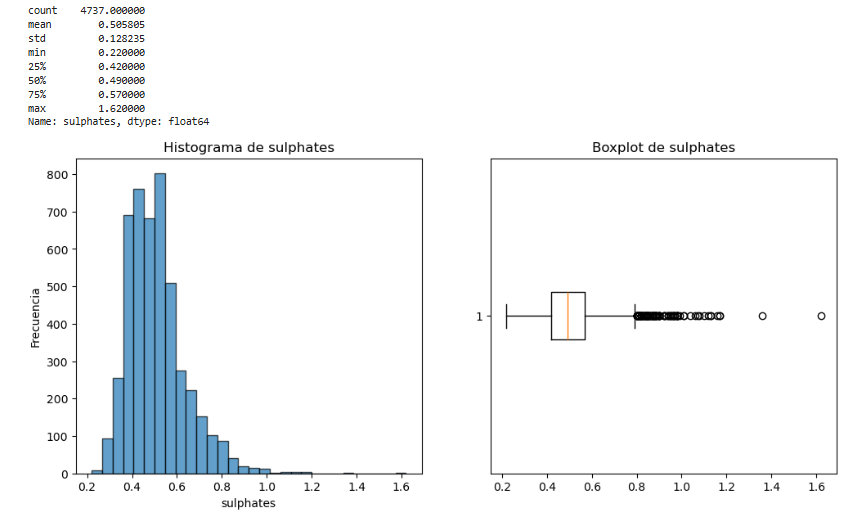
Seguimos con la otra variable PH

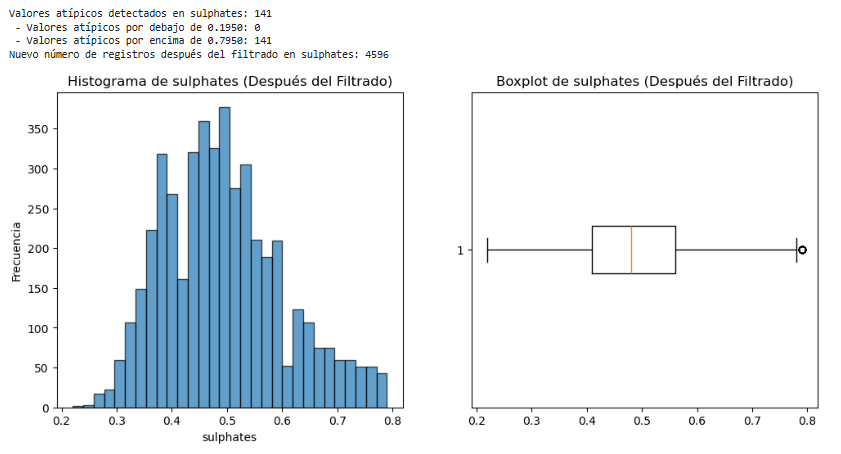


No es una distribución normal, así que aplicamos IQR

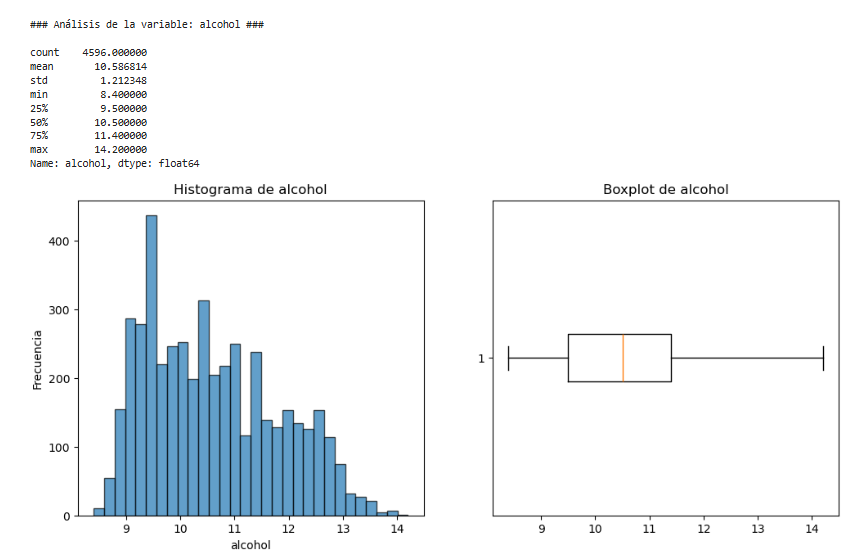


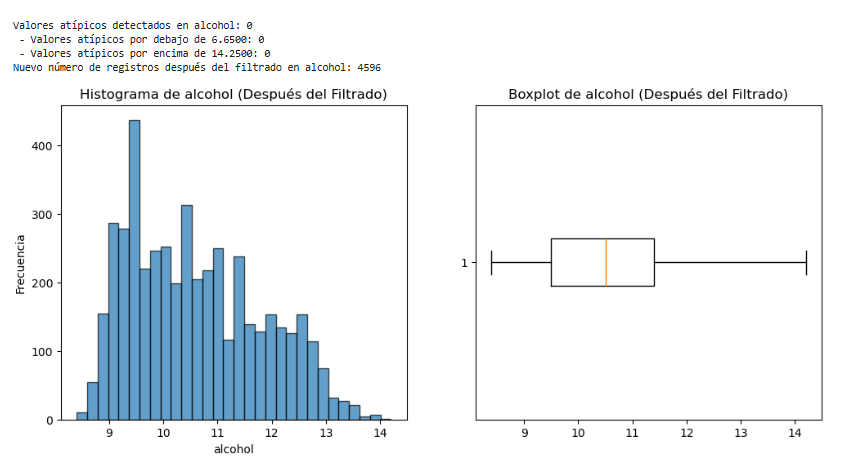
La siguiente variable sería Sulphates





Alchol





De un inicio teníamos 6497 datos = 1599 red + 4898 white

Al final hemos filtrado hasta 4,596 datos..

Esto es eliminamos un 29% (posiblemente sea muchos datos en mi opinión)

**4. Almacenar los datos limpios en SQLite**

- Usa SQLite para almacenar el dataframe limpio en una base de datos persistente. - Sigue la documentación oficial de SQLite: [SQLite Python Documentation]([https://docs.python.org/3/library/sqlite3.html](file:///C:\Users\MBITSC~1\AppData\Local\Temp\msohtmlclip1\01\clip_filelist.xml)).

**5. Realizar 3 consultas en SQLite**

Basándote en los datos y las columnas del dataset, realiza las siguientes consultas:

1. **Consulta 1**: ¿Cuál es el promedio de calidad (`quality`) por tipo de vino (`type`)?

✅ Promedio de calidad por tipo de vino:

red: 5.73

white: 5.93

2. **Consulta 2**: ¿Cuántos vinos tienen un nivel de alcohol superior a 10.5, agrupados por tipo?

✅ Cantidad de vinos con alcohol > 10.5 por tipo:

red: 149 vinos

white: 1995 vinos

3. **Consulta 3**: Obtén el conteo de vinos por nivel de acidez (`fixed acidity`) agrupados en rangos (por ejemplo, de 0-5, 5-10, 10-15).

✅ Conteo de vinos por nivel de acidez:

Rango 0-5: 32 vinos

Rango 5-10: 4564 vinos

**6. Exportar datos a JSONLines**

De cara a una potencial insercion en una base de datos noSQL como `mongoDB`, podemos servirnos de pandas para preparar los datos.

- ¿Qué estructura de datos de python es la más similar a un documento noSQL?

Un DataFrame es lo más sencillo para usar por Python y lo podemos sacar fácilmente de la SQL lite

- Usa Pandas para transformar los datos de una de las consultas en un archivo JSONLines.

Hecho, usamos la consulta de Calidad Promedio de cada vino, por ser pequeña y la guardamos en carpeta Outputs

- Usa la librería `jsonlines` para guardar el archivo.

exported\_wine\_data.jsonl

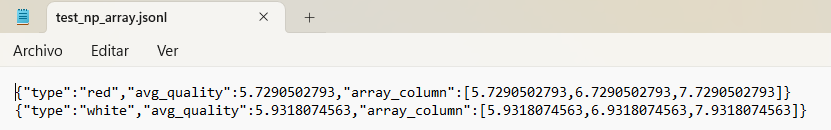
- ¿Qué problemas podrían surgir al transformar un dataframe en jsonlines?

1 problemas con las fechas, que el Json no entendería dierecatamente como fechas

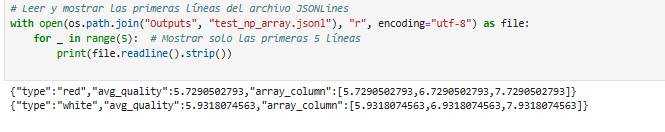
2 problemas con tipos de datos no serializables (Objetos o clases personalizadas), json se puede perder con clases o funciones

3 problema con arrays (np.array) se convierten en listas o cadenas. Pandas lo guarda como listas pero en un JSONlines.

- Añade una columna que sea originalmente un `np.array`,¿qué sucede al transformarlo en jsonlines?

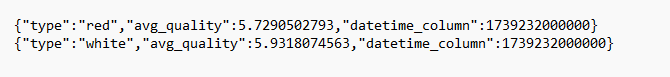


Se ven en la misma línea o fila, pero las series van entre corchetes



- Añade una columna que sea originalmente un `pd.datetime`,¿qué sucede al transformarlo en jsonlines?

Cuando guardamos una fecha por ejemplo 2025-02-01 como el json no tiene especificado ese tipo de formato, lo guarda como un “timestamp” que se representa con un número muy grande. Al intentar leerlo o hacer consultas podríamos tener problemas para leerlo.

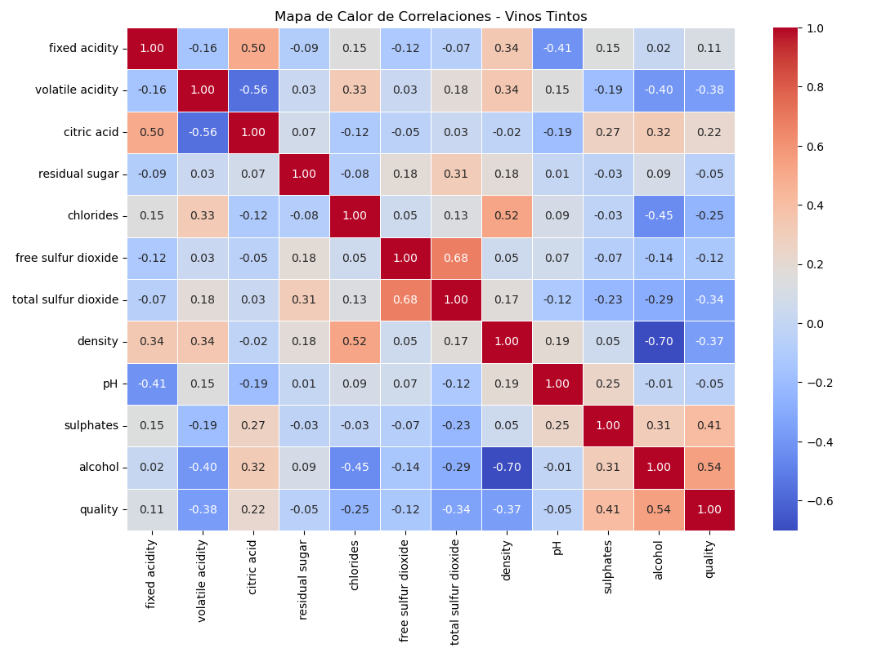


**7. Análisis de calidad de los vinos**

- Inspecciona qué caracteriza a los vinos tintos y blancos con mayor calidad (`quality`).

VINOS RED

Hacemos un Heatmap para ver correlación que existe entre las variables y el resultado quality



Sacamos las siguientes conclusiones

1.- Variable de Alcohol tiene la mayor correlación de 0.54 Vinos con mayor contenido alcohólico suelen ser percibidos como más equilibrados y estructurados, lo que influye positivamente en su calidad sensorial.

2.- Sulpahtes tiene un 0.41, posiblemente los vinos con más sulfitos pueden mantener mejor sus características organolépticas, afectando la percepción de calidad

3.- Density -0.37, una mayor densidad está asociada con vinos de menor calidad.

4.- Volatile acidity -0.38 posiblemente niveles elevados pueden generar aromas desagradables (como vinagre o acetona) lo que reduce la percepción de calidad del vino

5.- Total Sulfur -0.34 un alto contenido en dióxido de azufre suele estar asociado con menor calidad. Aunque se usa como conservante, niveles elevados pueden indicar sobreprotección del vino o mala gestión del proceso de vinificación, afectando el perfil sensorial.

Además existen correlaciones ENTRE variables interesantes.

Free Sulfur Dioxide y Total Sulfur Dioxide (+0.68)

Chlorides y Density (+0.52)

Fixed Acidity y Citric Acid (+0.50)

Fixed Acidity y Citric Acid (+0.50)

Density y Alcohol (-0.70)

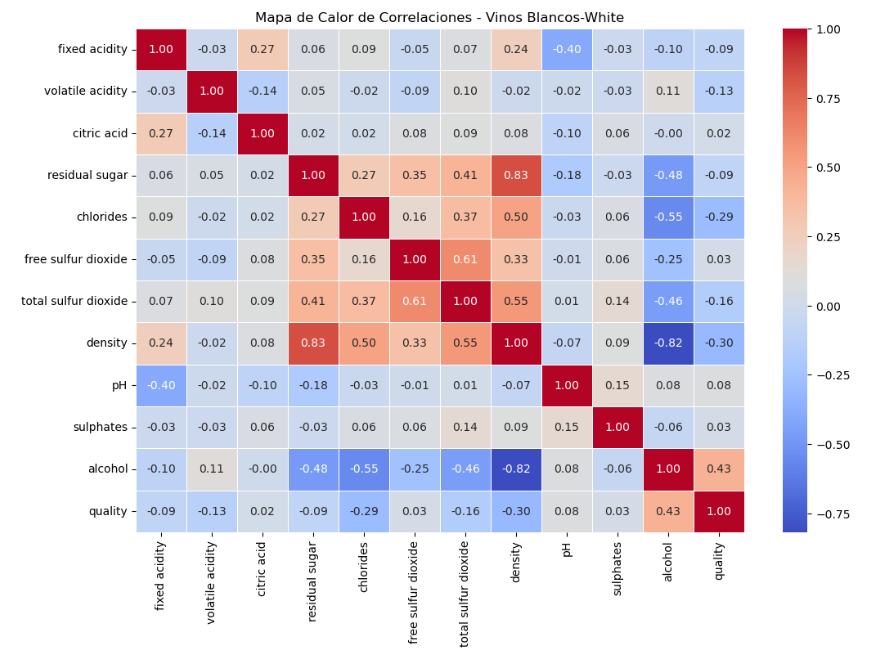
Conclusión

✔️ Algunas correlaciones como Alcohol vs Density o Free vs Total Sulfur Dioxide eran esperadas y confirmadas por la química del vino.

✔️ Otras, como la relación negativa entre Volatile Acidity y Citric Acid, nos pueden dar pistas sobre procesos de fermentación y su impacto en el vino.

✔️ Si quisiéramos hacer más análisis avanzados, podríamos aplicar técnicas de regresión o clustering para entender aún mejor estas relaciones.

Vamos a realizar ahora el mismo análisis para el vino White



Alcohol sigue siendo la variable más correlacionada con la calidad (+0.43), pero menos que en los tintos (+0.54)

Density y Residual Sugar están fuertemente correlacionados (+0.83)

Chlorides tiene una correlación más negativa con la calidad en blancos (-0.29) que en tintos (-0.25)

- Usa análisis estadístico, gráficos o cualquier técnica que consideres relevante para identificar patrones.

**Notas adicionales para la entrega**

- Asegúrate de documentar cada paso en tu entrega y justificar tus decisiones (por ejemplo, cómo identificaste los atípicos).

- El entregable principal será un Jupyter notebook con todos los pasos que habéis seguido y las explicaciones que consideréis pertinentes.

- Incluye en un fichero requirements.txt todas las librerías que estés utilizando.

- La entrega se hará tanto a través de Leemons en formato .zip como en el repositorio de Git asignado para el proyecto.

- Cada grupo debe trabajar en su propia rama del repositorio de GIT [https://github.com/francisco-mbit/PC1](file:///C:\Users\MBITSC~1\AppData\Local\Temp\msohtmlclip1\01\clip_filelist.xml)

- Loggead el esfuerzo en diferentes commits, asegurándoos de realizar al menos un commit por persona.

- Realizad un commit final con el mensaje: **ENTREGA FINAL**.

¡Buena suerte con el proyecto! 😊