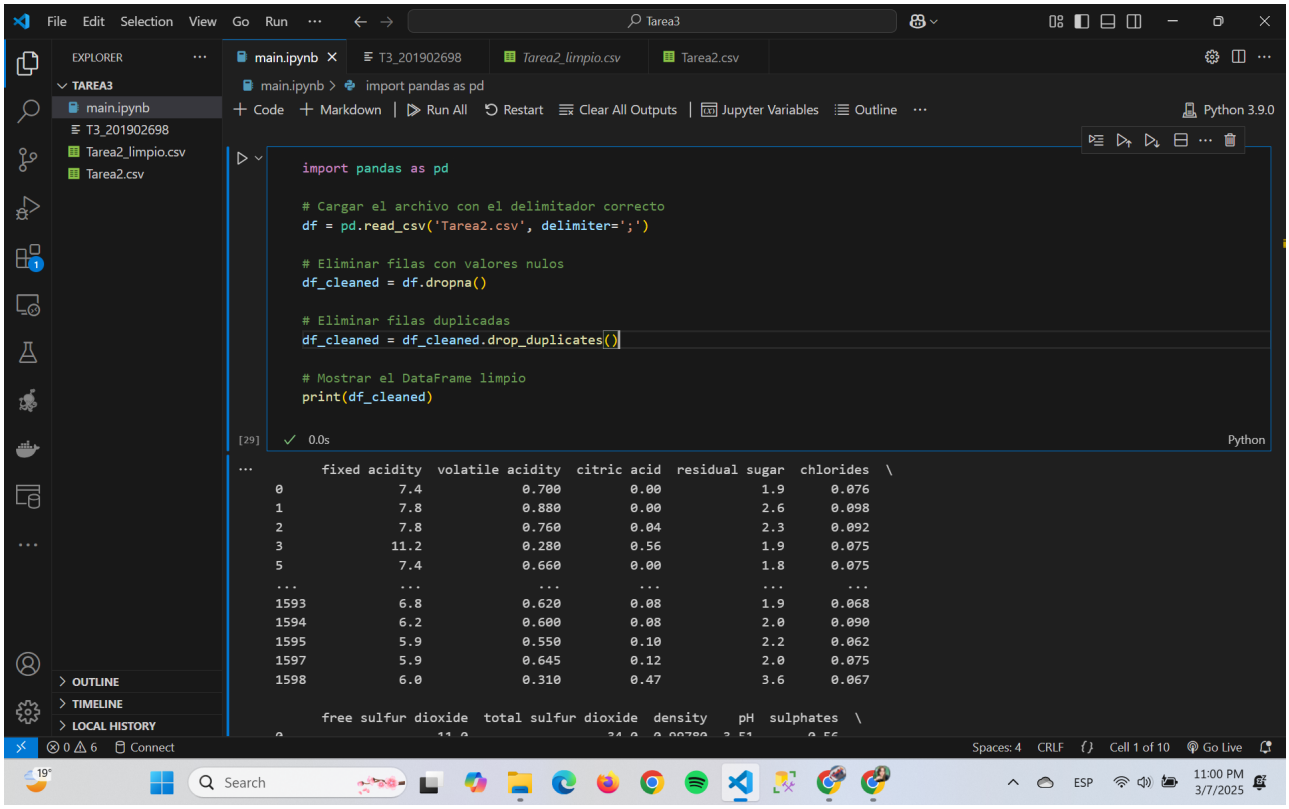
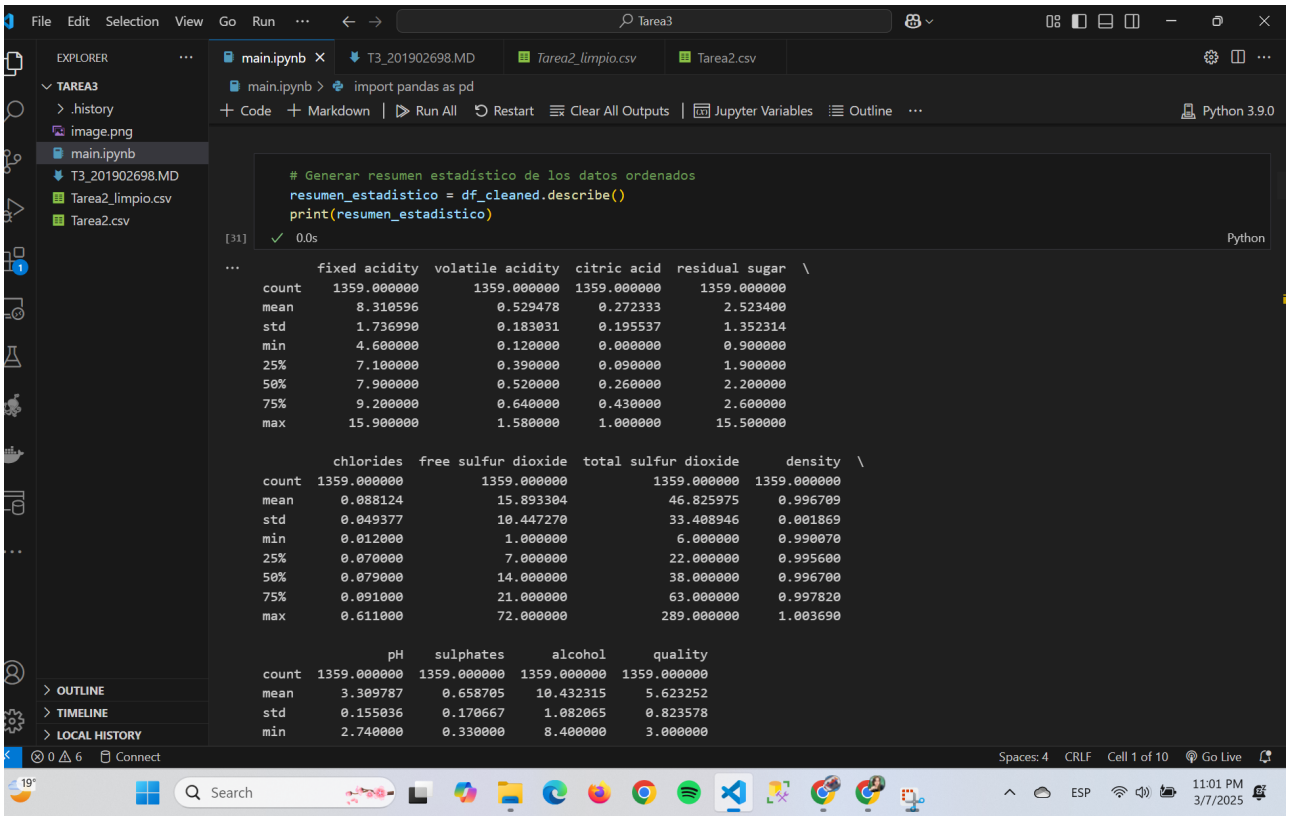
Tarea 3

Pablo Javier Batz Contreras - 201902698

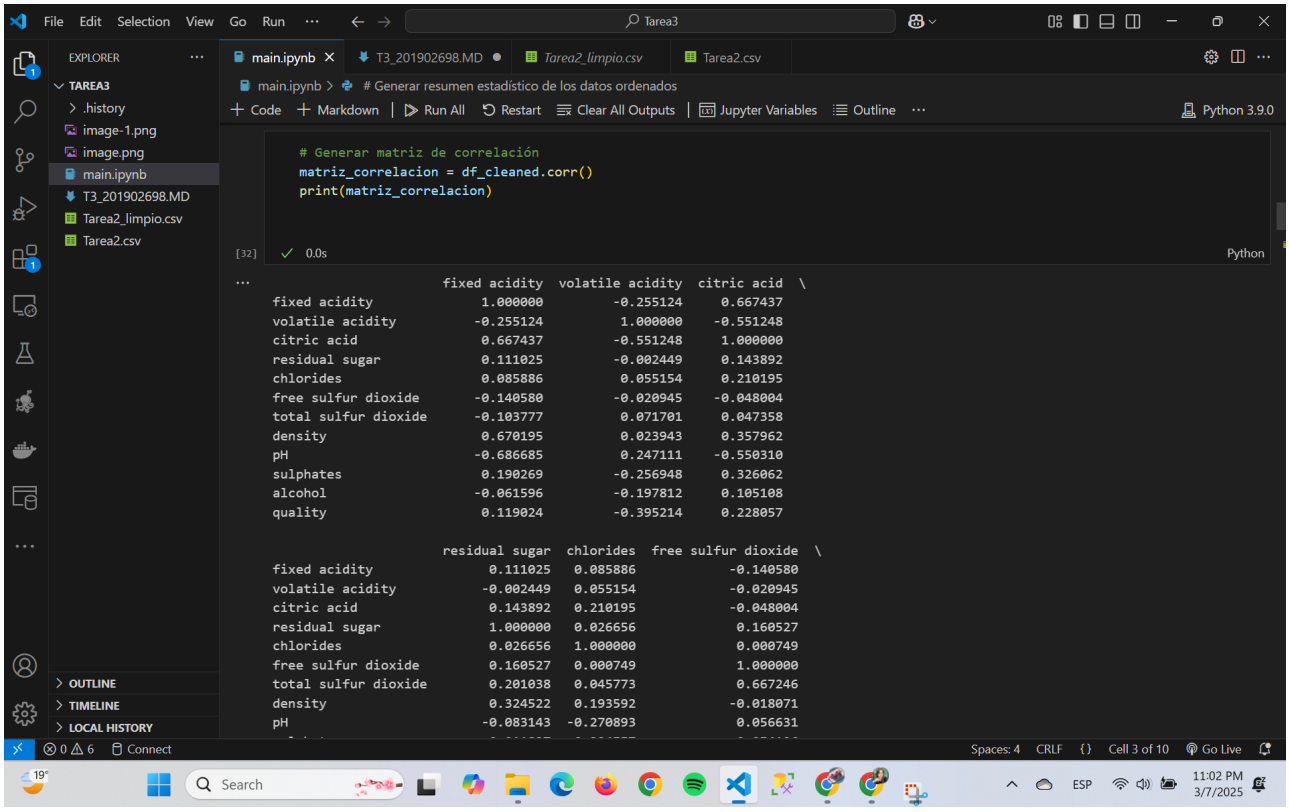
- Limpieza de datos

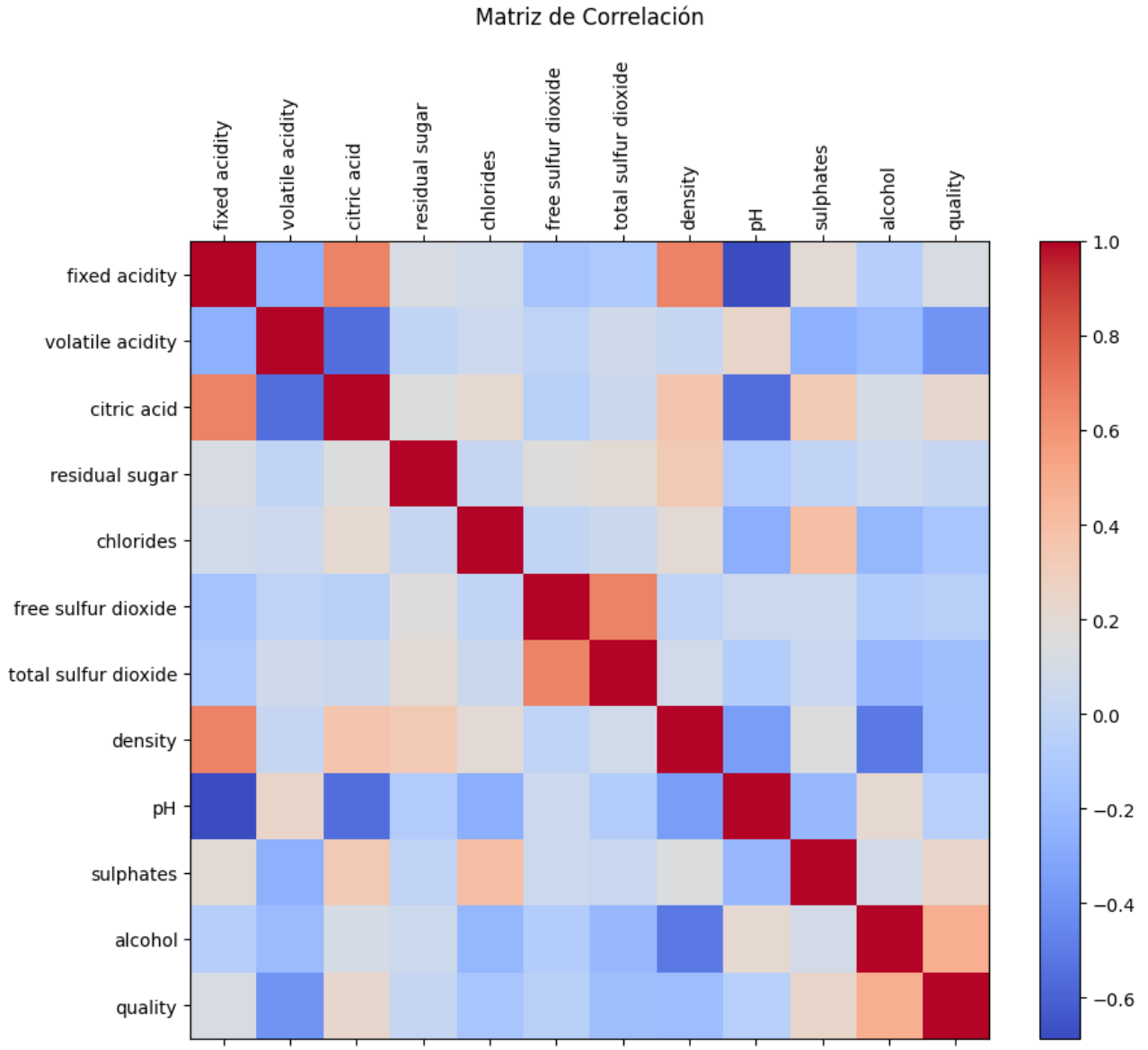


- Resumen de datos ordenados

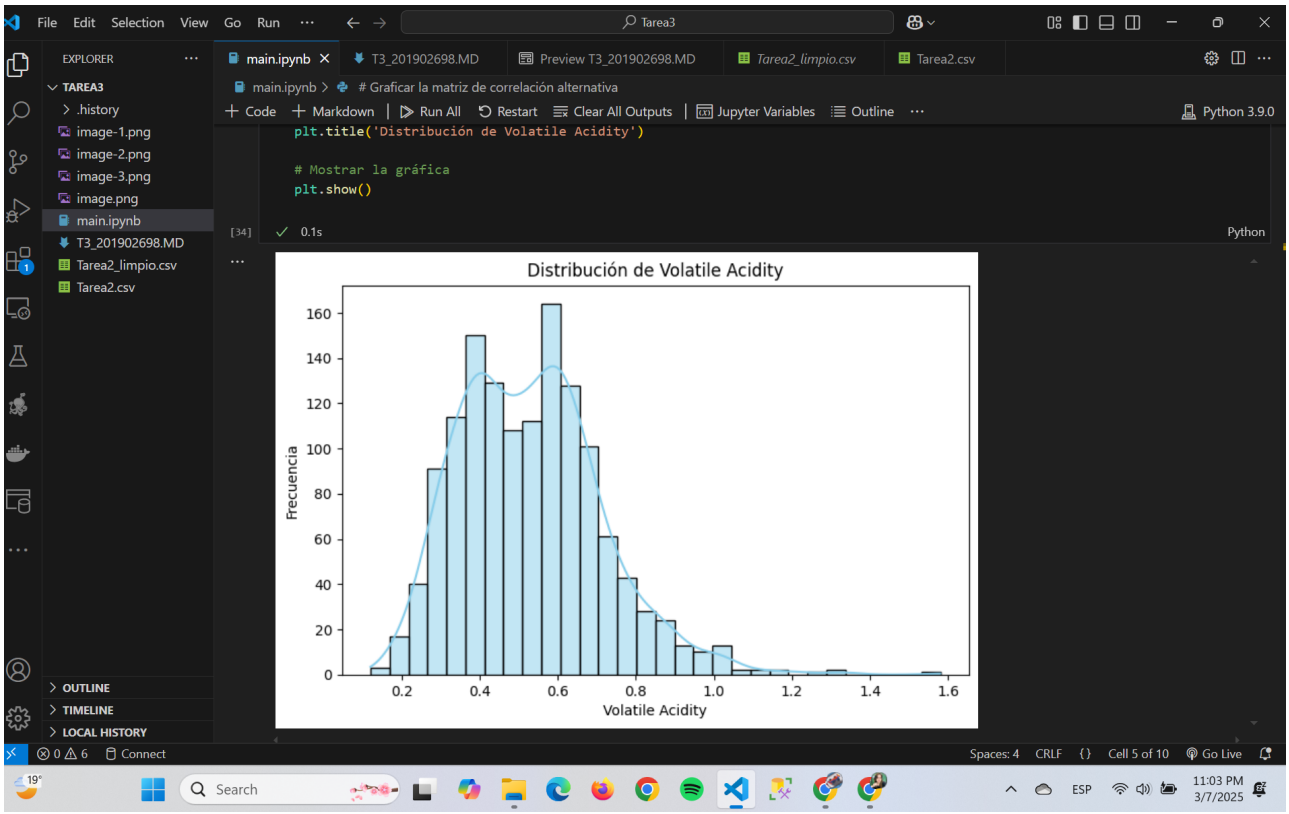


- Matriz de correlacion

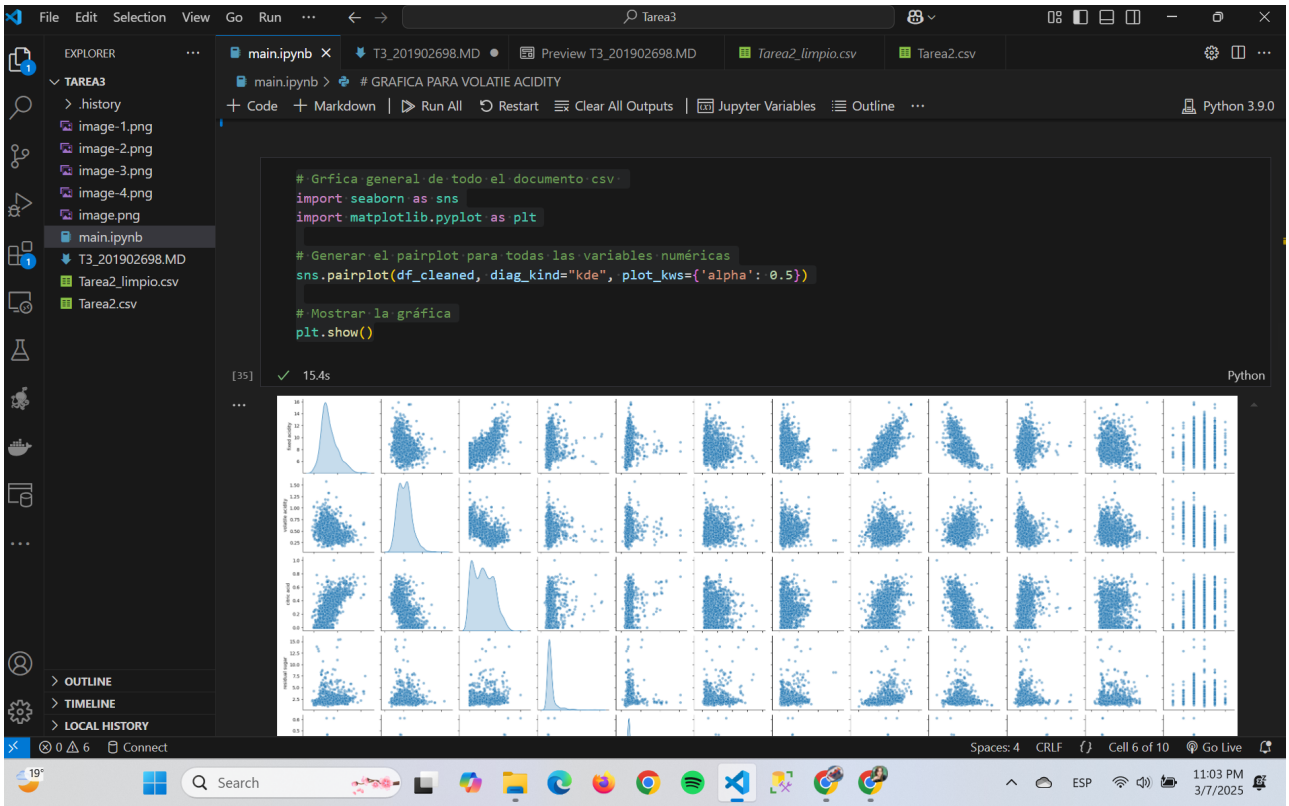




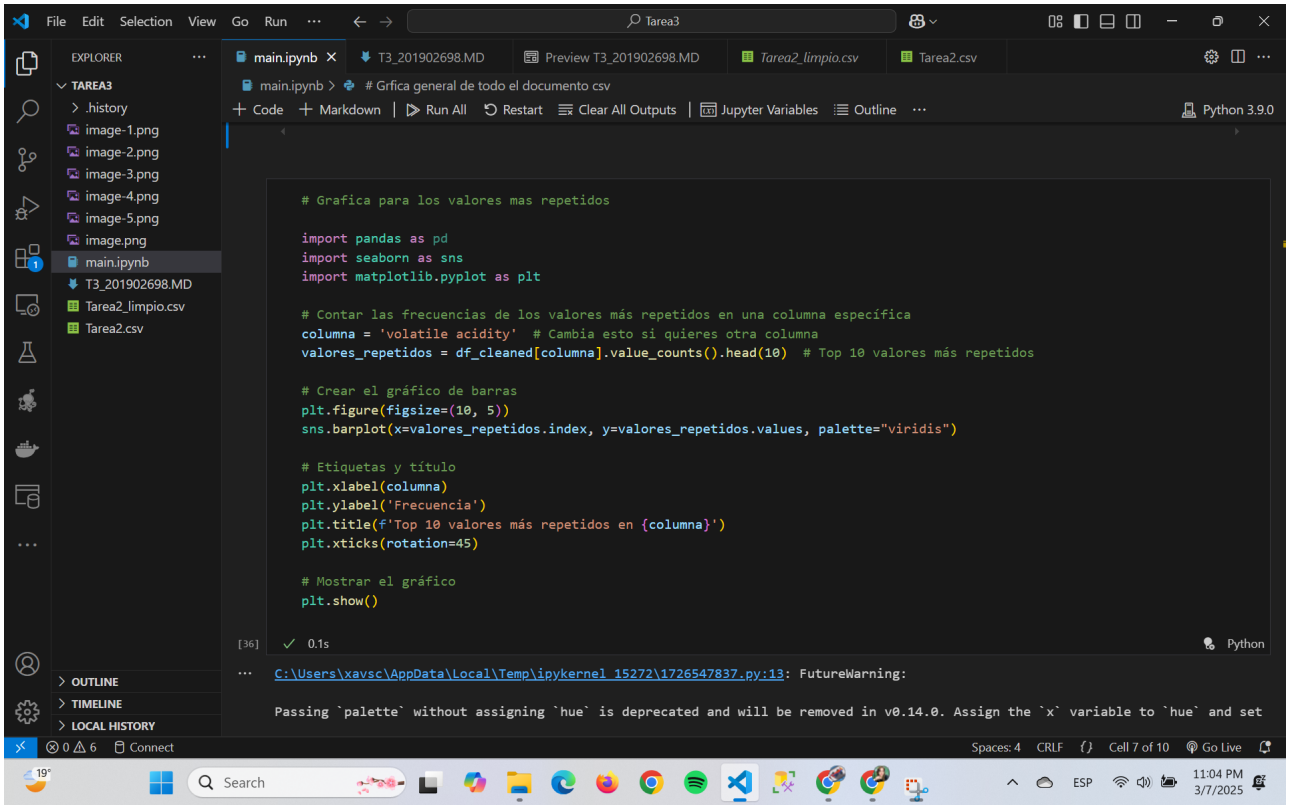
- Analisis Distribución de Volatile Acidity

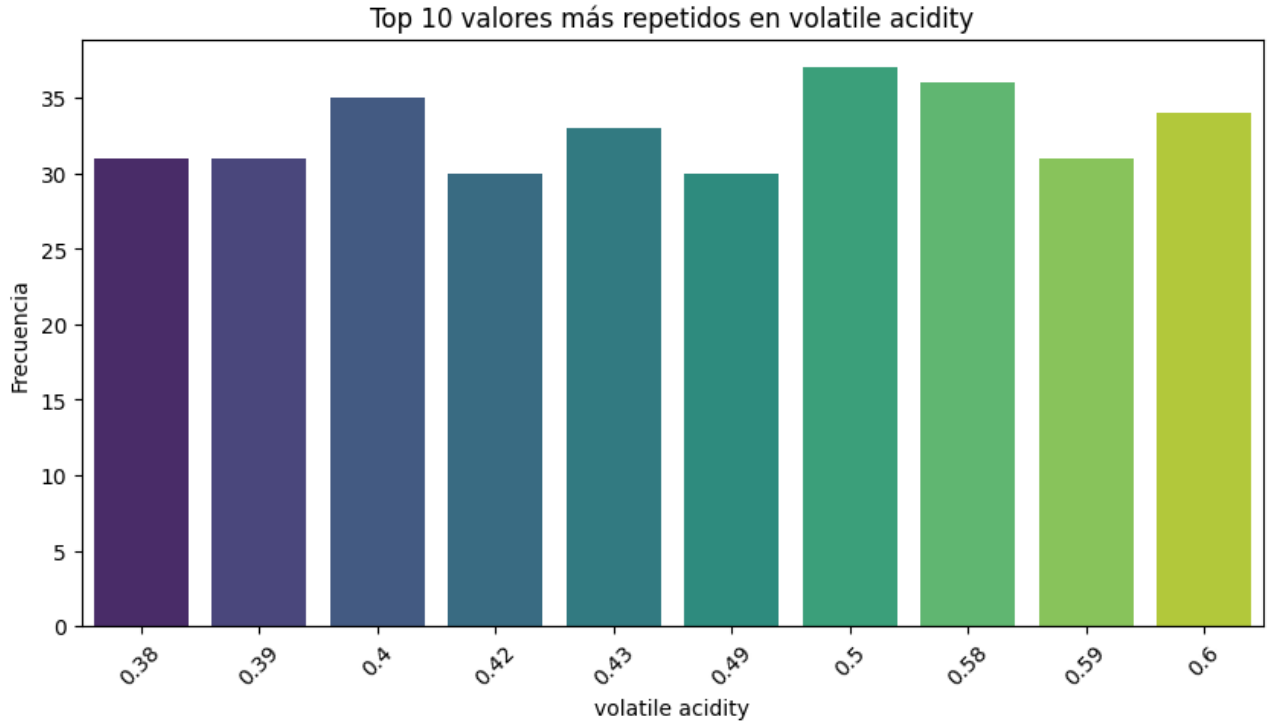


- Grafico general para el archivo CSV.

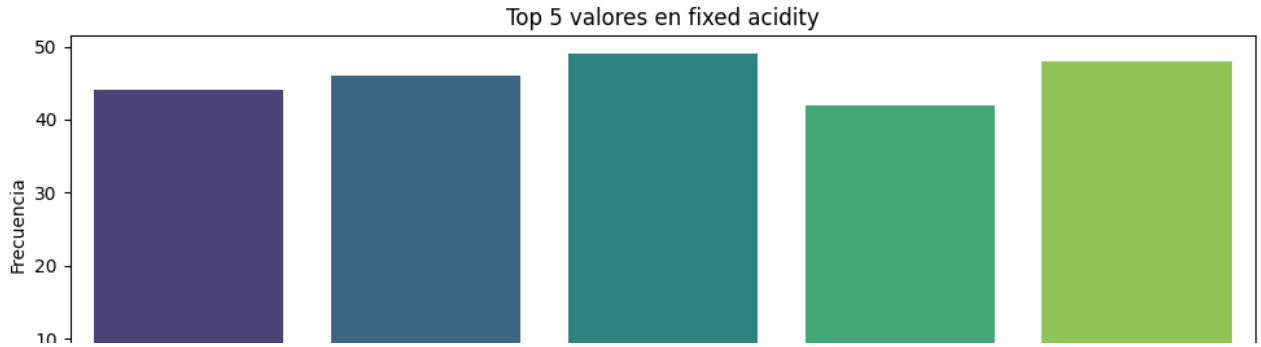


- Grafico de valores repetidos

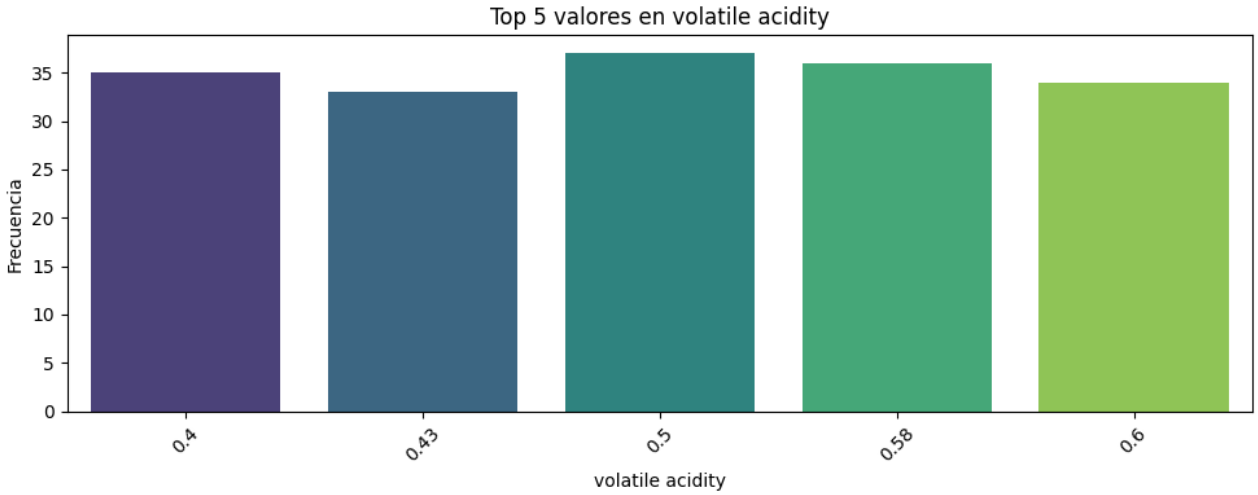


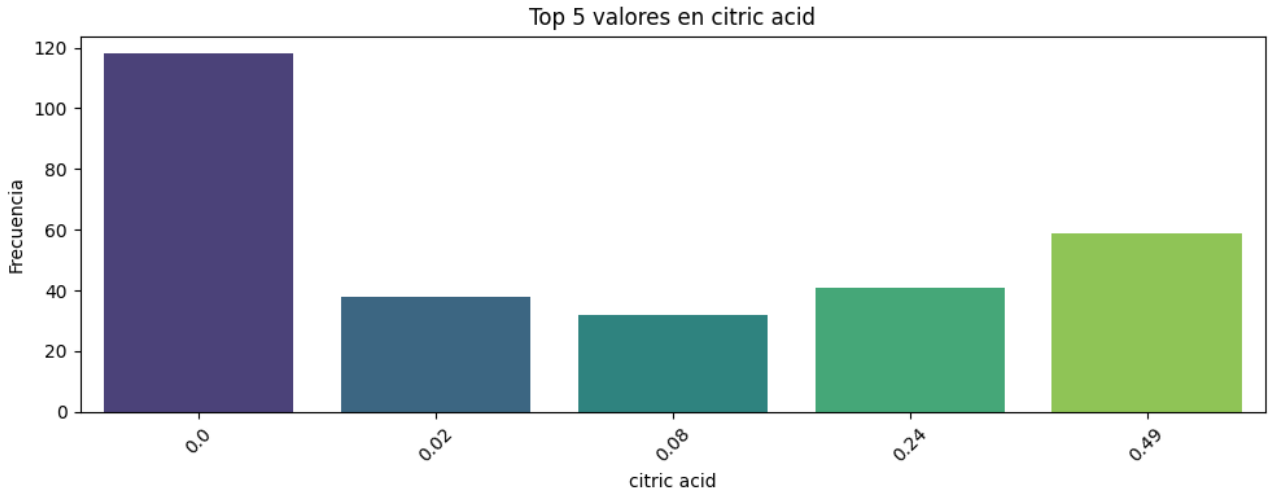


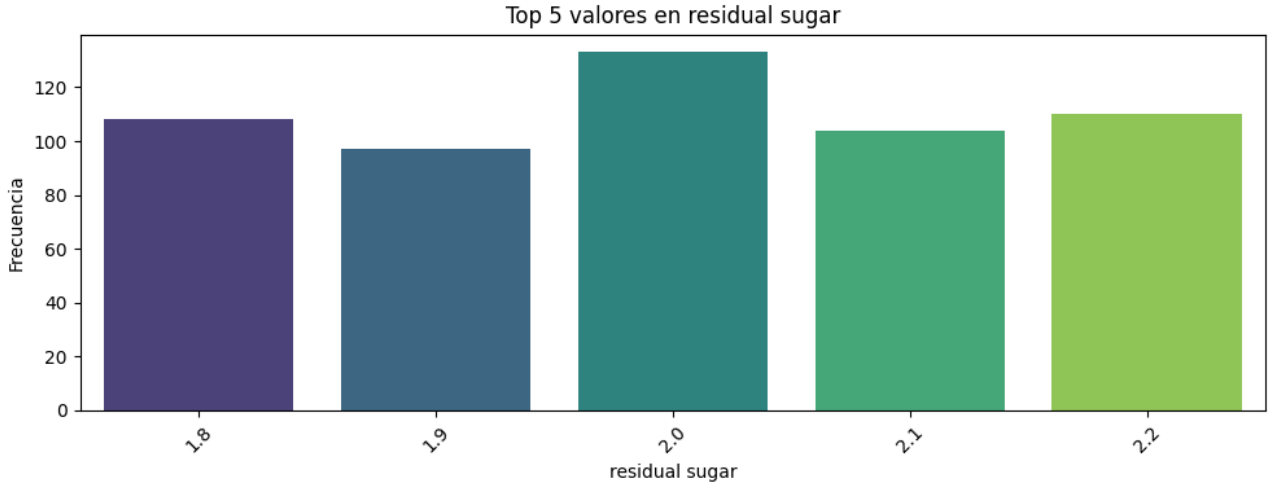
- Top 5 de valores mas repetidos

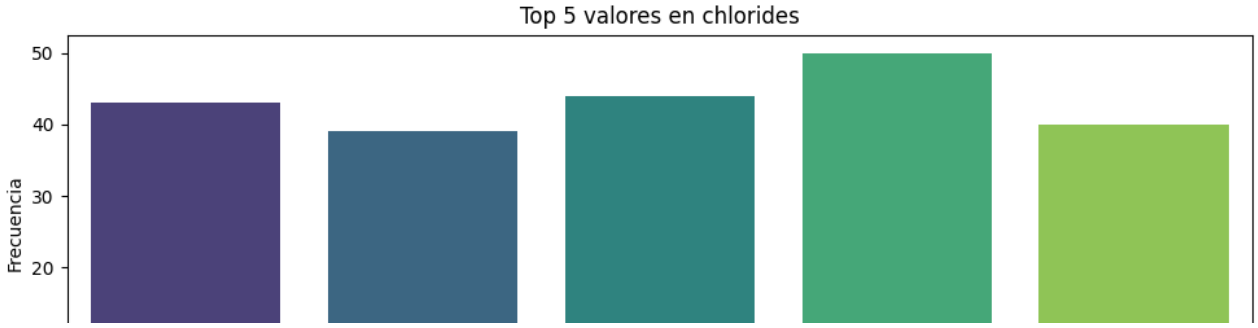


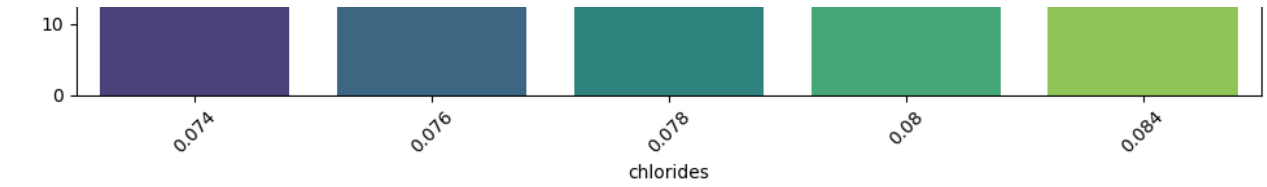

0
7.0
7.1
7.2
7.5
7.8
fixed acidity

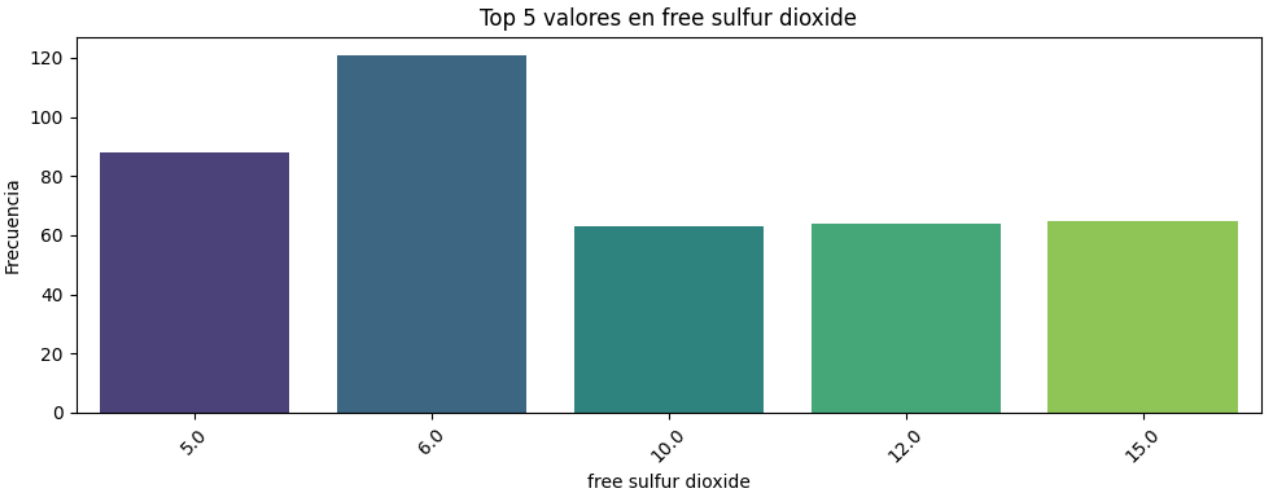

Top 5 valores en volatile acidity
Frecuencia
0
5
10
15
20
25
30
35
0.4
0.43
0.5
0.58
0.6
volatile acidity

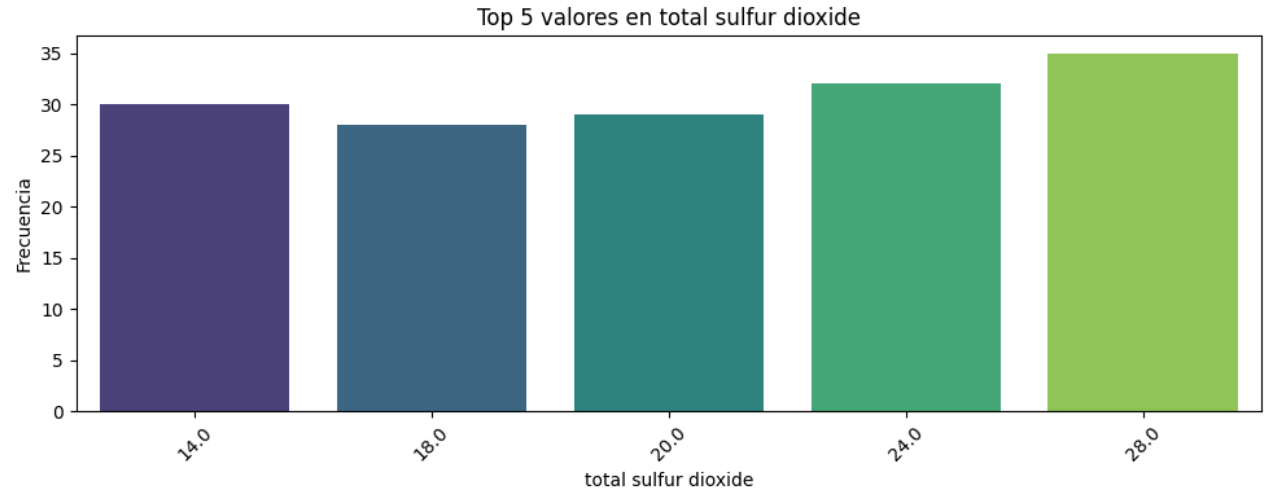

Top 5 valores en citric acid
Frecuencia
0
20
40
60
80
100
120
0.0
0.02
0.08
0.24
0.49
citric acid

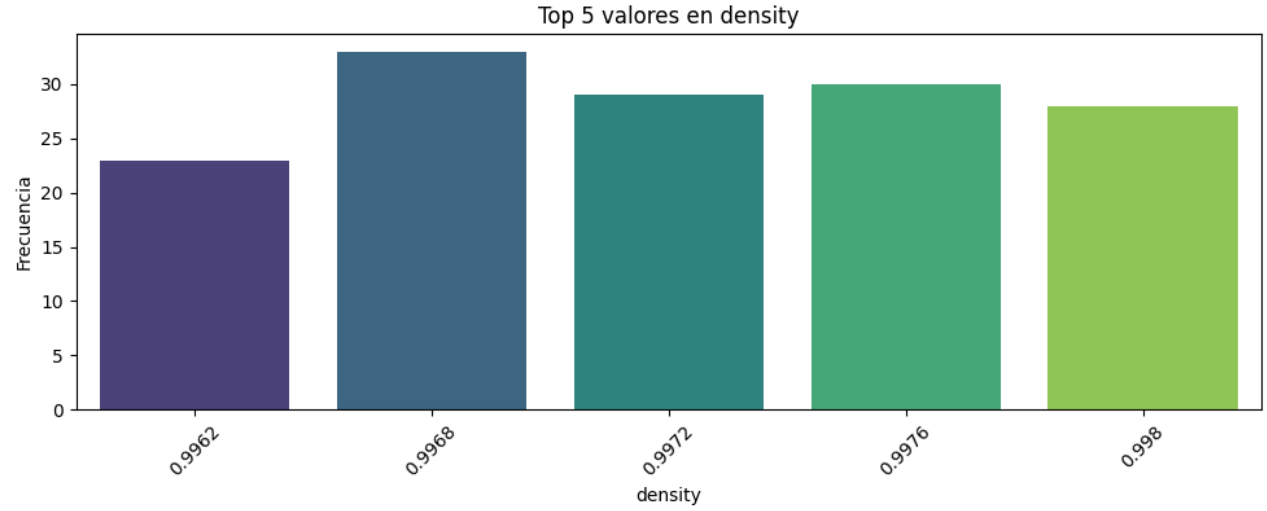

Top 5 valores en residual sugar
Frecuencia
0
20
40
60
80
100
120
1.8
1.9
2.0
2.1
2.2
residual sugar

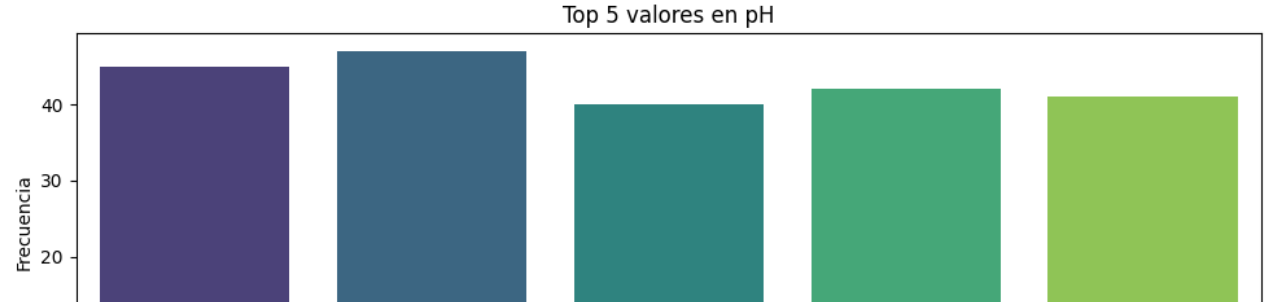

Top 5 valores en chlorides
Frecuencia
20
30
40
50

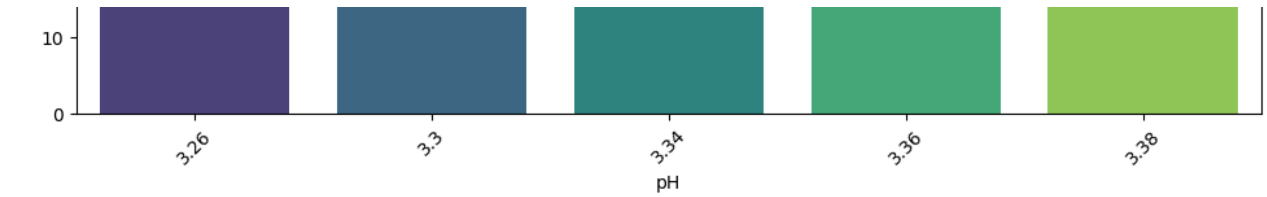

10
0
0.074
0.076
0.078
0.08
0.084
chlorides

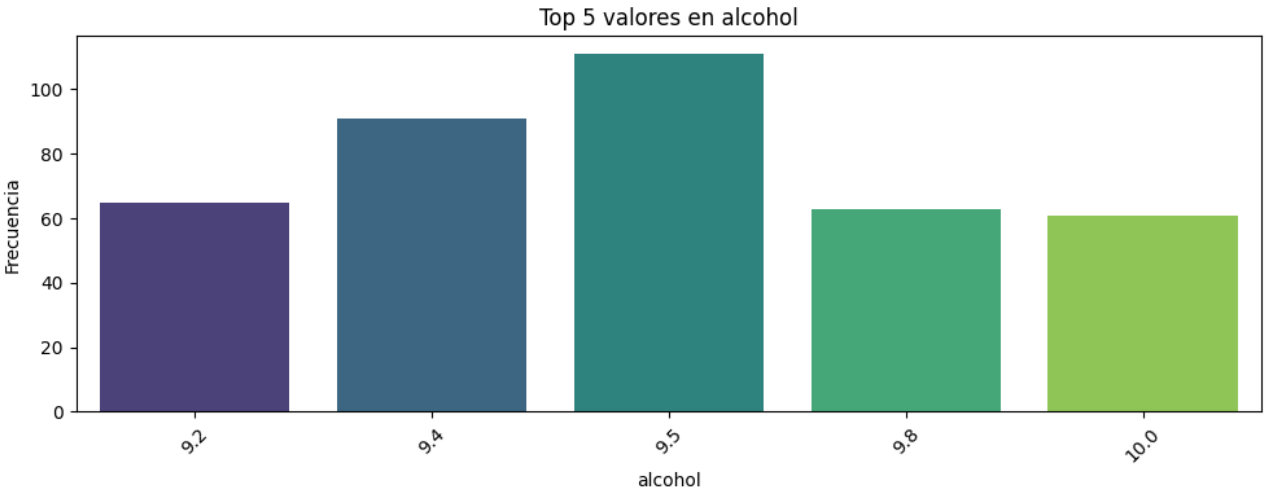

Top 5 valores en free sulfur dioxide
120
100
80
60
40
20
0
Frecuencia
5.0
6.0
10.0
12.0
15.0
free sulfur dioxide

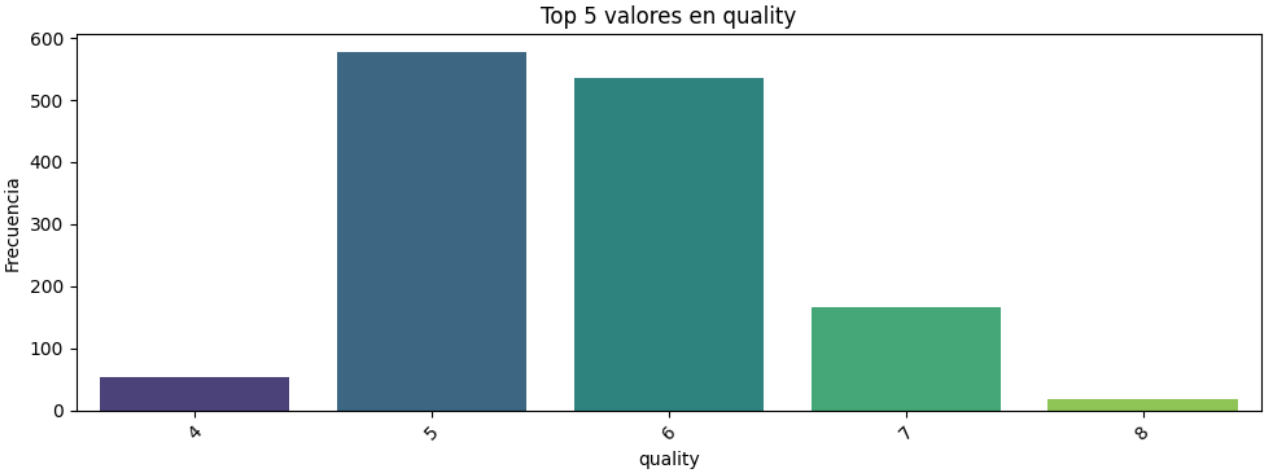

Top 5 valores en total sulfur dioxide
35
30
25
20
15
10
5
0
Frecuencia
14.0
18.0
20.0
24.0
28.0
total sulfur dioxide


Top 5 valores en density
30
25
20
15
10
5
0
Frecuencia
0.9962
0.9968
0.9972
0.9976
0.998
density


Top 5 valores en pH
40
30
20
Frecuencia







Conclusiones.

1. Calidad de los Datos

Después de eliminar valores nulos y duplicados, los datos están más limpios y confiables para el análisis. Sin embargo, si se eliminaron muchas filas, podría indicar problemas de calidad en la recolección de datos.

2. Relación entre Variables

El heatmap de correlación mostró qué variables tienen mayor relación entre sí. Por ejemplo:

- Si "volatile acidity" tiene una correlación negativa con la calidad, significa que **valores altos de acidez volátil pueden afectar negativamente la calidad del producto.
- Otras variables con correlaciones fuertes pueden indicar factores clave que influyen en el resultado.

3. Tendencias y Valores Más Frecuentes

El análisis de los valores más repetidos en cada columna nos ayuda a identificar tendencias. Por ejemplo:

- Si ciertos niveles de acidez, pH o alcohol son los más comunes, esto puede indicar un estándar de producción o una tendencia en los datos.
- Si hay valores que dominan demasiado una columna, puede ser necesario investigar si el dataset está desbalanceado.