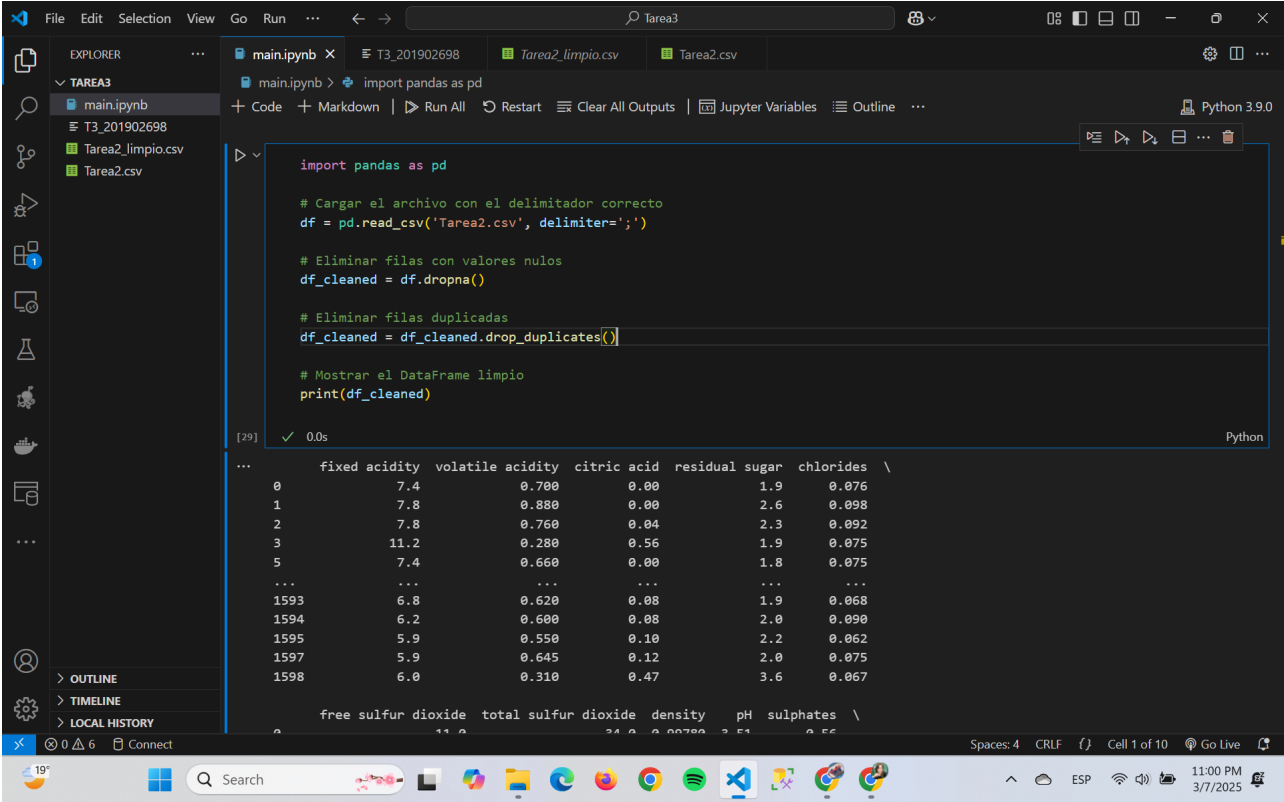


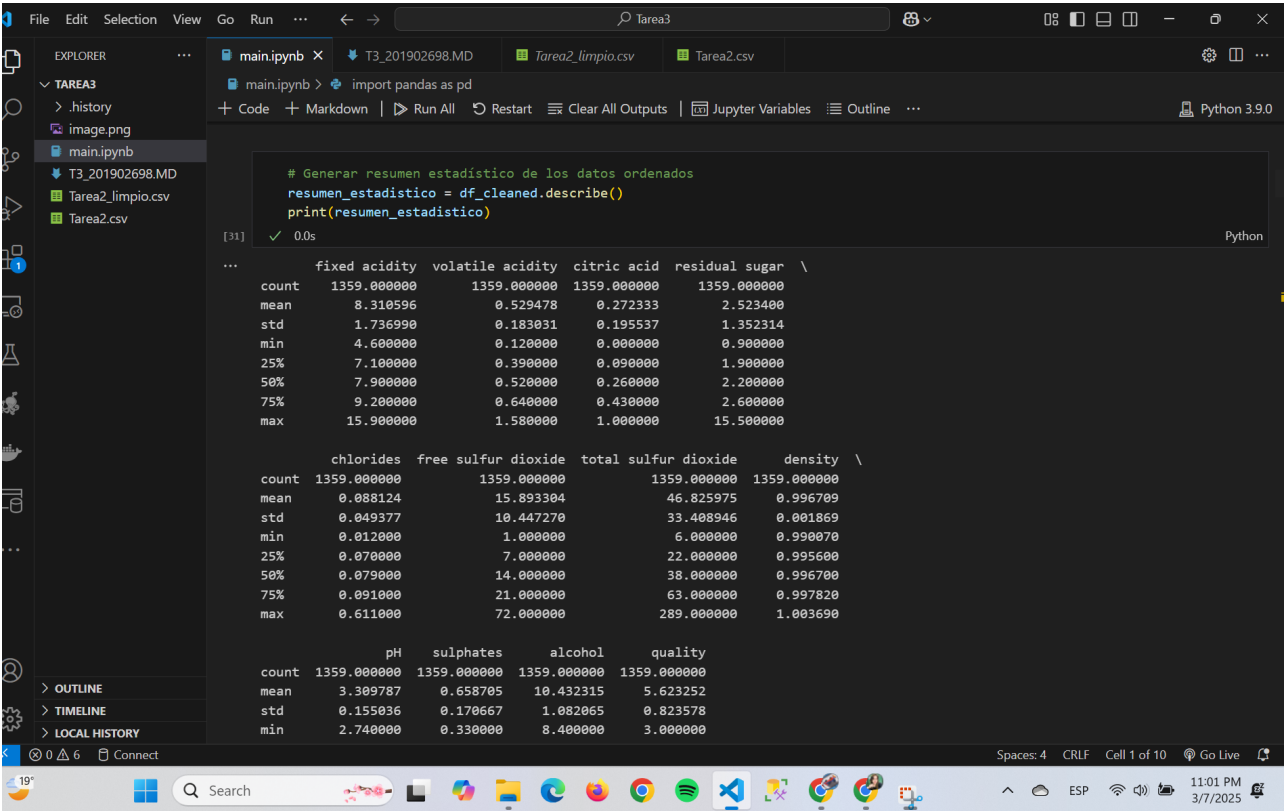
# Tarea 3

Pablo Javier Batz Contreras - 201902698

- Limpieza de datos



- Resumen de datos ordenados



• Matriz de correlacion

EXPLORER

TAREA3

image-1.png

image.png

main.ipynb

T3\_201902698.MD

Tarea2\_limpio.csv

Tarea2.csv

OUTLINE

TIMELINE

LOCAL HISTORY

main.ipynb

# Generar resumen estadístico de los datos ordenados

Code

Markdown

Run All

Restart

Clear All Outputs

Jupyter Variables

Outline

Python 3.9.0

```
# Generar matriz de correlación
matriz_correlacion = df_cleaned.corr()
print(matriz_correlacion)
```

[32] ✓ 0.0s Python

	fixed acidity	volatile acidity	citric acid	\
fixed acidity	1.000000	-0.255124	0.667437	
volatile acidity	-0.255124	1.000000	-0.551248	
citric acid	0.667437	-0.551248	1.000000	
residual sugar	0.111025	-0.002449	0.143892	
chlorides	0.085886	0.055154	0.210195	
free sulfur dioxide	-0.140580	-0.020945	-0.048004	
total sulfur dioxide	-0.103777	0.071701	0.047358	
density	0.670195	0.023943	0.357962	
pH	-0.686685	0.247111	-0.550310	
sulphates	0.190269	-0.256948	0.326062	
alcohol	-0.061596	-0.197812	0.105108	
quality	0.119024	-0.395214	0.228057	

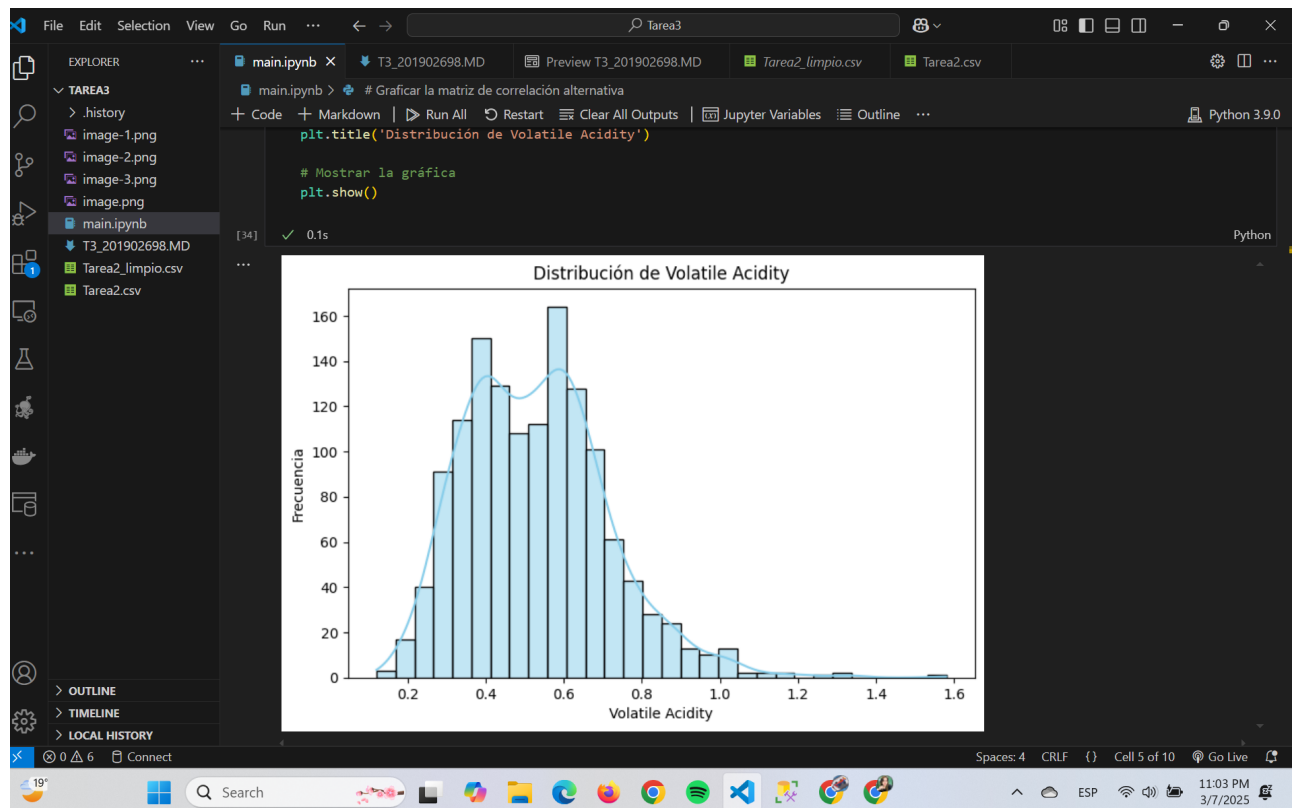
	residual sugar	chlorides	free sulfur dioxide	\
fixed acidity	0.111025	0.085886	-0.140580	
volatile acidity	-0.002449	0.055154	-0.020945	
citric acid	0.143892	0.210195	-0.048004	
residual sugar	1.000000	0.026656	0.160527	
chlorides	0.026656	1.000000	0.000749	
free sulfur dioxide	0.160527	0.000749	1.000000	
total sulfur dioxide	0.201038	0.045773	0.667246	
density	0.324522	0.193592	-0.018071	
pH	-0.083143	-0.270893	0.056631	

Matriz de Correlación

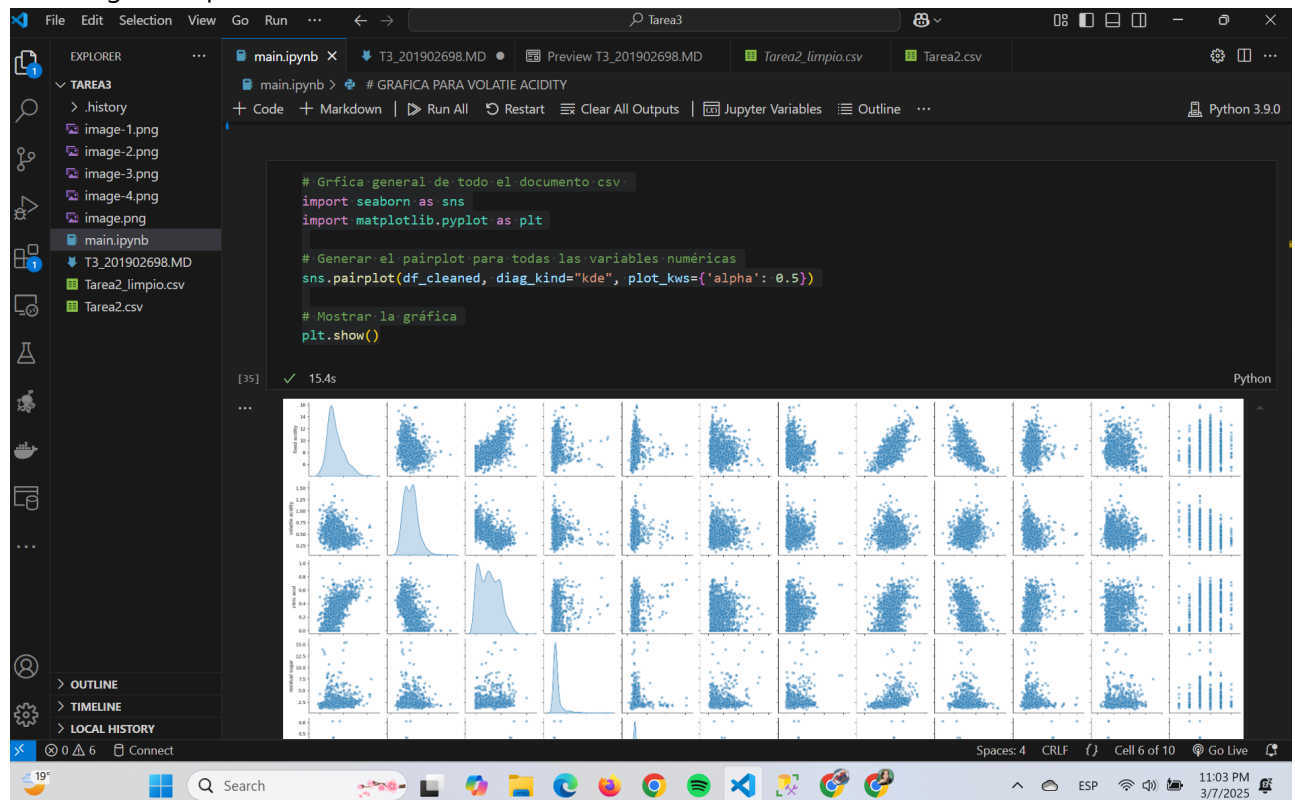
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
fixed acidity	1.000000	-0.255124	0.667437	0.111025	0.085886	-0.140580	-0.103777	0.670195	-0.686685	0.190269	-0.061596	0.119024
volatile acidity	-0.255124	1.000000	-0.551248	-0.002449	0.055154	-0.020945	0.071701	0.023943	0.247111	-0.256948	-0.197812	-0.395214
citric acid	0.667437	-0.551248	1.000000	0.143892	0.210195	-0.048004	0.047358	0.357962	-0.550310	0.326062	0.105108	0.228057
residual sugar	0.111025	-0.002449	0.143892	1.000000	0.026656	0.160527	0.201038	0.324522	-0.083143	0.000749	0.000000	0.000000
chlorides	0.085886	0.055154	0.210195	0.026656	1.000000	0.000749	0.045773	0.193592	-0.270893	0.000000	0.000000	0.000000
free sulfur dioxide	-0.140580	-0.020945	-0.048004	0.160527	0.000749	1.000000	0.667246	-0.018071	0.056631	0.000000	0.000000	0.000000
total sulfur dioxide	-0.103777	0.071701	0.047358	0.201038	0.045773	0.667246	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000
density	0.670195	0.023943	0.357962	0.324522	0.193592	-0.018071	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000
pH	-0.686685	0.247111	-0.550310	-0.083143	-0.270893	0.056631	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000
sulphates	0.190269	-0.256948	0.326062	0.000749	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000
alcohol	-0.061596	-0.197812	0.105108	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000
quality	0.119024	-0.395214	0.228057	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000

2 / 8

- Analisis Distribución de Volatile Acidity



- Grafico general para el archivo CSV.



- Grafico de valores repetidos

EXPLORER

TAREA3

.history

image-1.png

image-2.png

image-3.png

image-4.png

image-5.png

image.png

main.ipynb

T3\_201902698.MD

Tarea2\_limpio.csv

Tarea2.csv

OUTLINE

TIMELINE

LOCAL HISTORY

main.ipynb

T3\_201902698.MD

Preview T3\_201902698.MD

Tarea2\_limpio.csv

Tarea2.csv

main.ipynb

# Grfica general de todo el documento csv

+ Code + Markdown | Run All | Restart | Clear All Outputs | Jupyter Variables | Outline

Python 3.9.0

```
# Grafica para los valores mas repetidos

import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Contar las frecuencias de los valores más repetidos en una columna específica
columna = 'volatile acidity' # Cambia esto si quieres otra columna
valores_repetidos = df_cleaned[columna].value_counts().head(10) # Top 10 valores más repetidos

# Crear el gráfico de barras
plt.figure(figsize=(10, 5))
sns.barplot(x=valores_repetidos.index, y=valores_repetidos.values, palette="viridis")

# Etiquetas y título
plt.xlabel(columna)
plt.ylabel('Frecuencia')
plt.title(f'Top 10 valores más repetidos en {columna}')
plt.xticks(rotation=45)

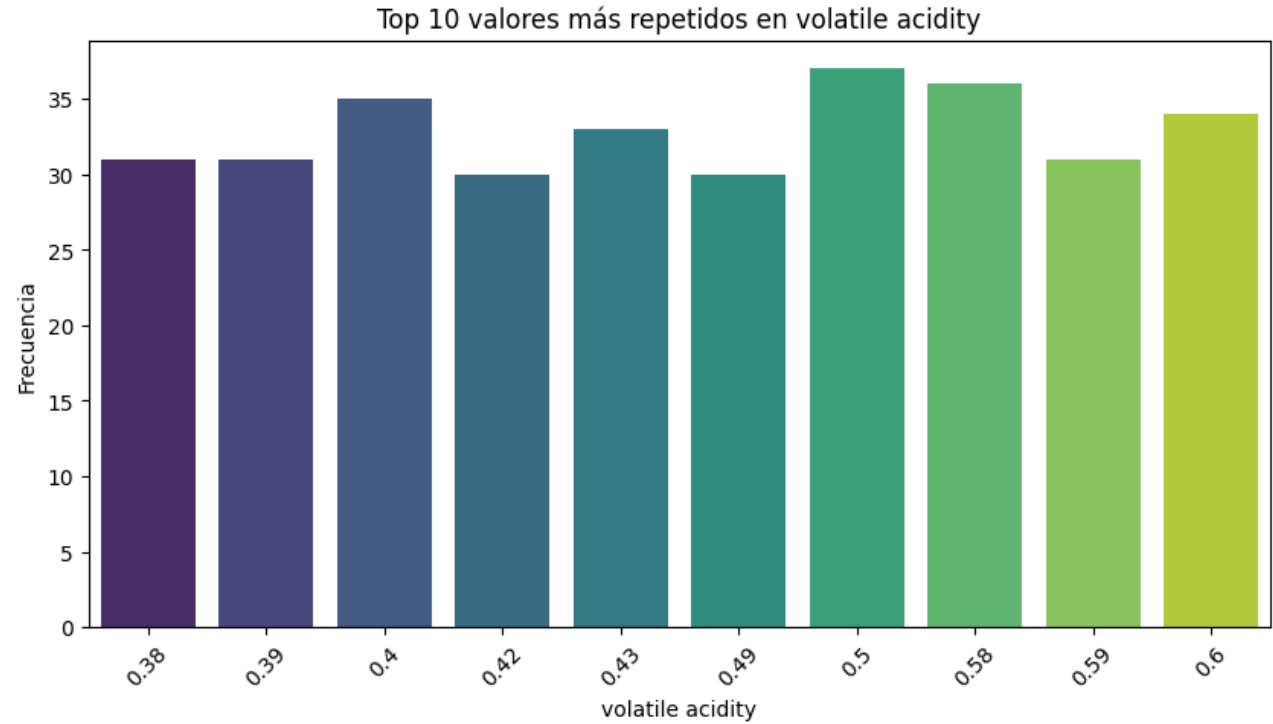
# Mostrar el gráfico
plt.show()
```

[36] ✓ 0.1s

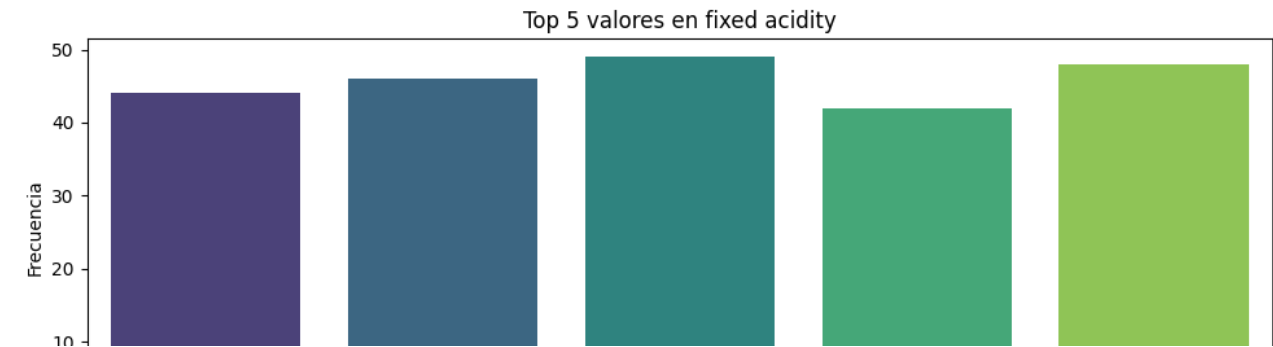
C:\Users\xavsc\AppData\Local\Temp\ipykernel\_15272\1726547837.py:13: FutureWarning: Passing 'palette' without assigning 'hue' is deprecated and will be removed in v0.14.0. Assign the 'x' variable to 'hue' and set

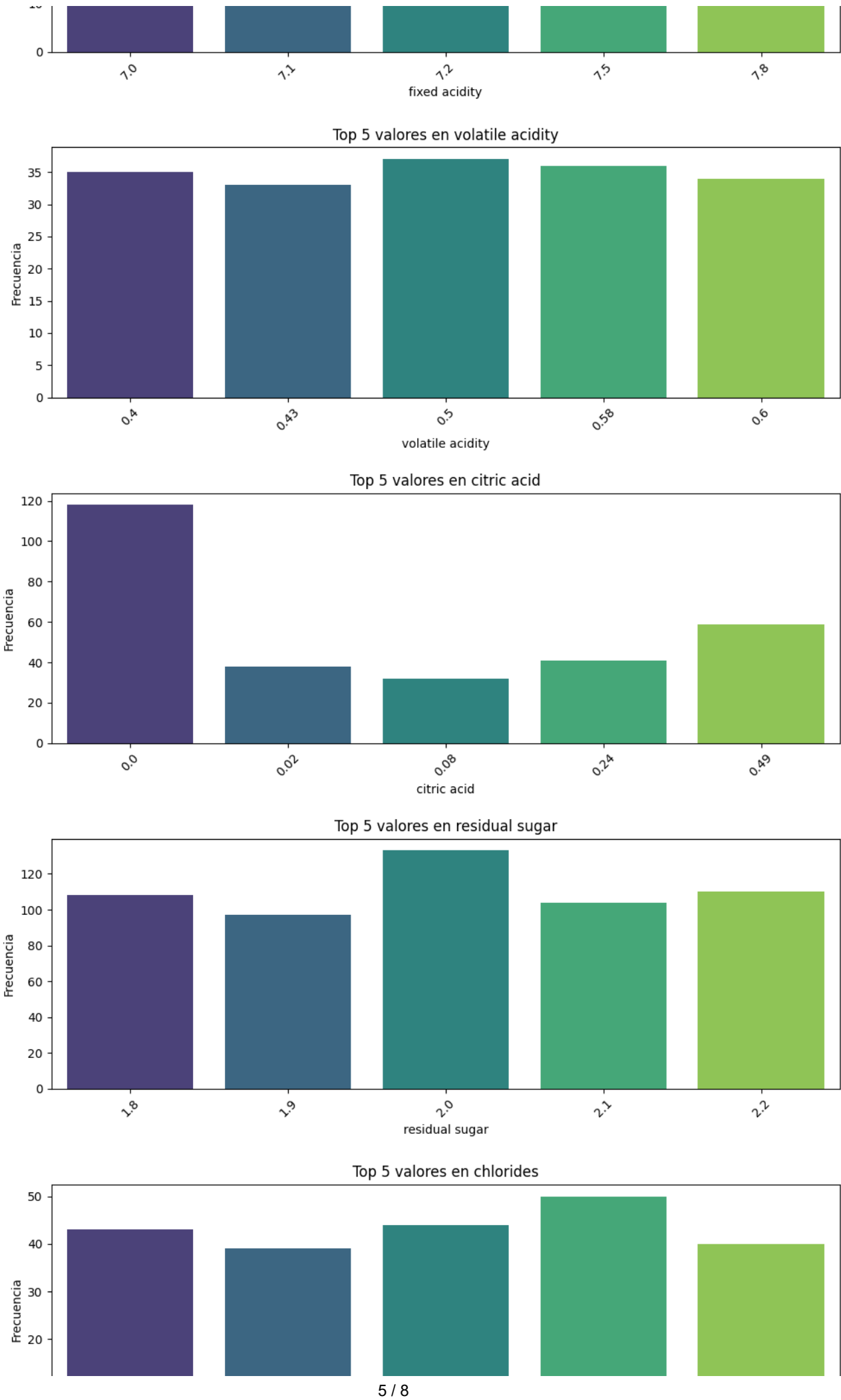
Spaces: 4 CRLF {} Cell 7 of 10 Go Live

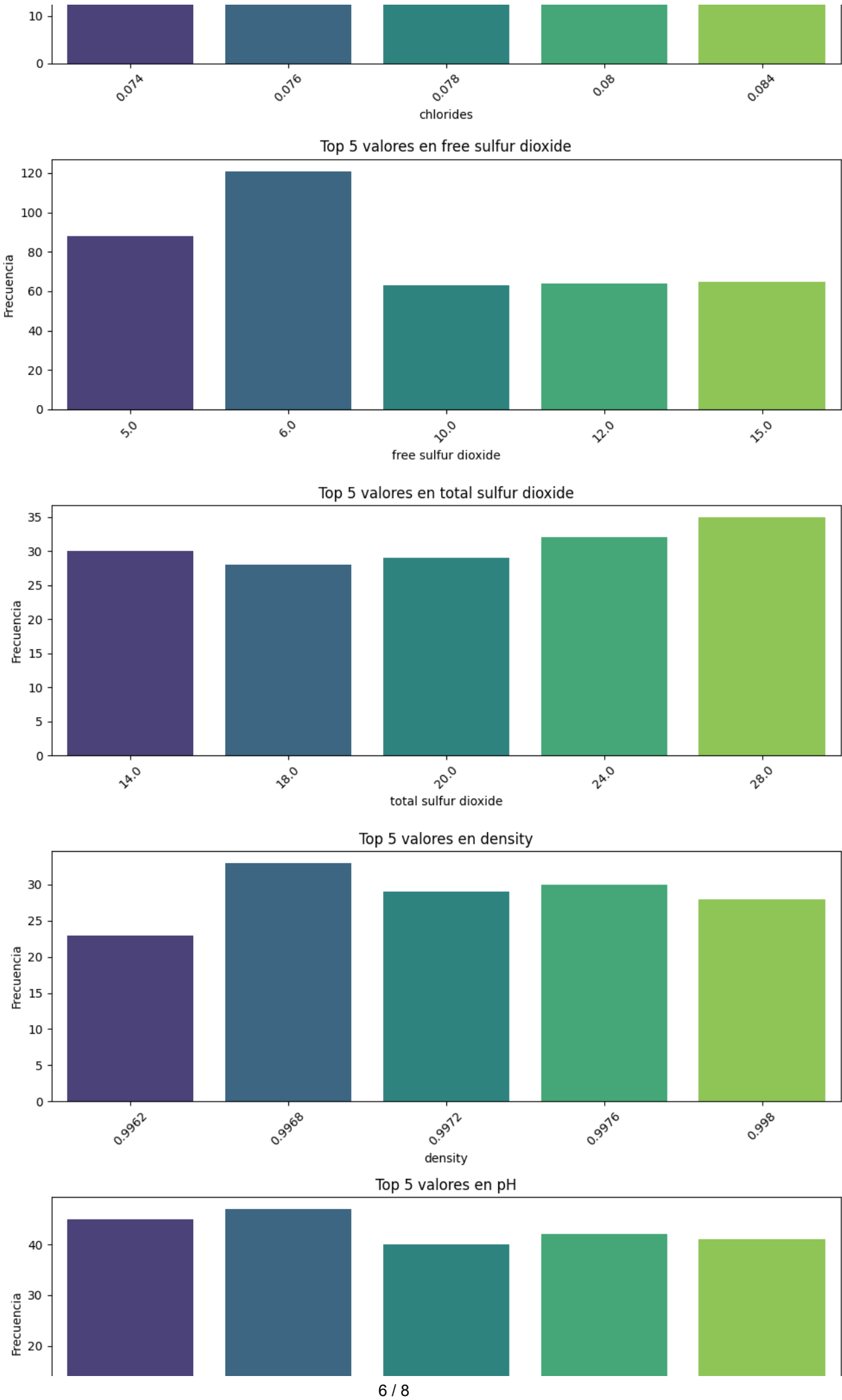
19° Search ESP 11:04 PM 3/7/2025

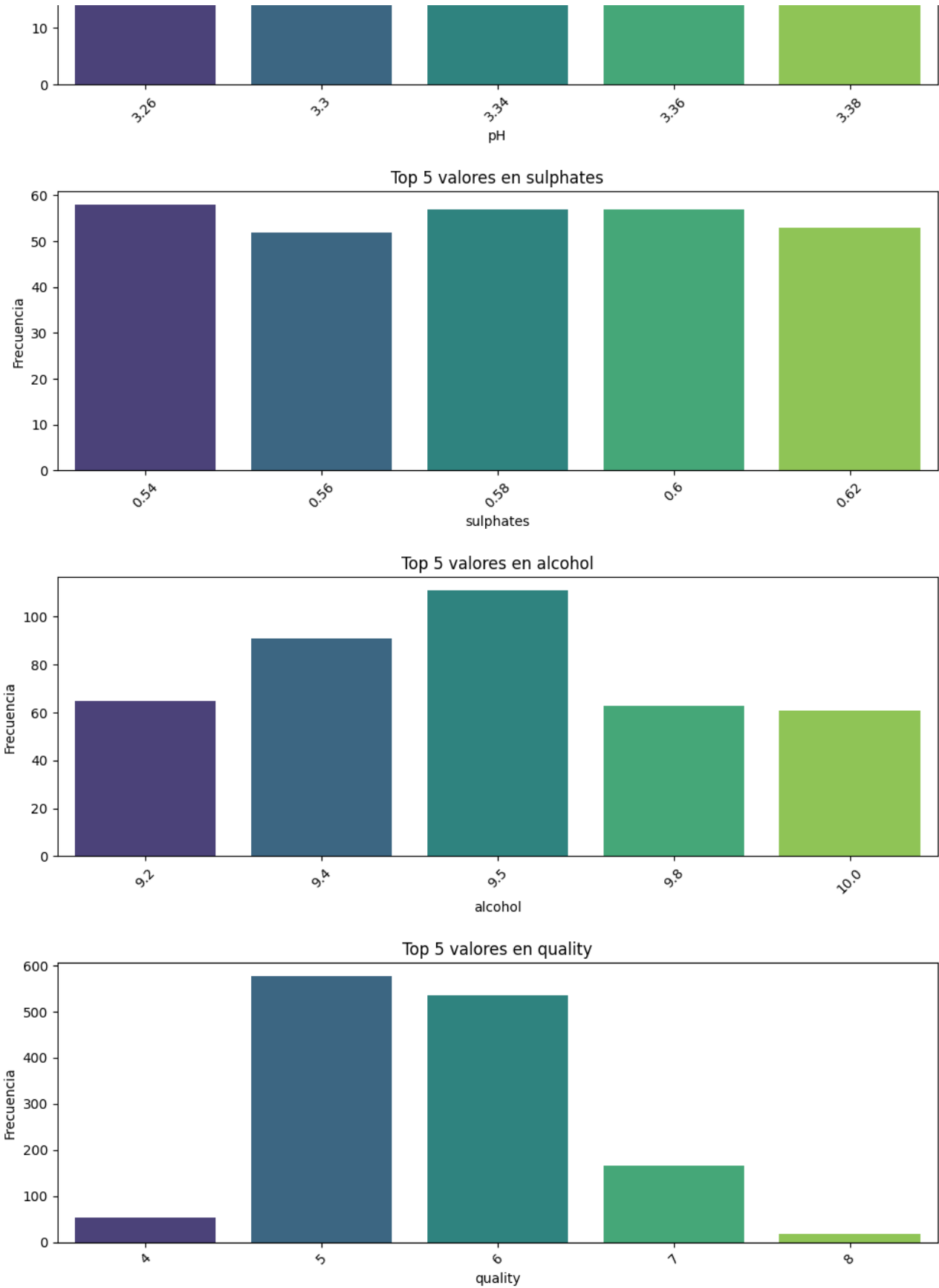


- Top 5 de valores mas repetidos









# 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.