POSTGRADOS
ESPOL

JUNIO 2024

Clasificación y Análisis de la Calidad del Agua Superficial

Villamar Karla,
Filian Karla

Introducción

- **Recurso natural más vulnerables y explotado**
- **Fuentes de abastecimiento de agua**

Contaminadas

- **Fuentes naturales y antropogénicas**

Carga y Preprocesamiento de los Datos

Sample Point	Year	Month	pH	Hardness	TDS	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity

```
cd = pd.read_excel("water_potability.xlsx")
cd
```

	Punto	Year	Mes	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
0	P1	1996	Abril	NaN	204.890455	547.000000	7.300212	368.516441	564.308654	10.379783	86.990970	2.963135	0
1	P2	1996	Abril	3.716080	129.422921	593.000000	6.635246	NaN	592.885359	15.180013	56.329076	4.500656	0
2	P3	1996	Abril	8.099124	224.236259	502.000000	9.275884	NaN	418.606213	16.868637	66.420093	3.055934	0
3	P4	1996	Abril	8.316766	214.373394	533.000000	8.059332	356.886136	363.266516	18.436524	100.341674	4.628771	0
4	P5	1996	Abril	9.092223	181.101509	543.000000	6.546600	310.135738	398.410813	11.558279	3		
...	...	...	...	...	...	...	...	...	...	...			
3265	P6	2023	Junio	8.197353	203.105091	27701.794055	6.472914	328.886838	444.612724	14.250875	6		
3266	P7	2023	Junio	8.372910	169.087052	14622.745494	7.547984	NaN	464.525552	11.083027	3		
3267	P8	2023	Junio	8.989900	215.047358	15921.412018	6.297312	312.931022	390.410231	9.899115	5		
3268	P9	2023	Junio	6.702547	207.321086	17246.920347	7.708117	304.510230	329.266002	16.217303	2		
3269	P10	2023	Junio	11.491011	94.812545	37188.826020	9.263166	258.930600	439.893618	16.172755	4		

3270 rows × 13 columns

```
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Punto            3270 non-null   object
 1   Year             3270 non-null   int64
 2   Mes              3270 non-null   object
 3   ph               2779 non-null   float64
 4   Hardness         3270 non-null   float64
 5   Solids           3270 non-null   float64
 6   Chloramines      3270 non-null   float64
 7   Sulfate          2493 non-null   float64
 8   Conductivity     3270 non-null   float64
 9   Organic_carbon   3270 non-null   float64
 10  Trihalomethanes  3109 non-null   float64
 11  Turbidity        3270 non-null   float64
```

Análisis Exploratorio de Datos (EDA)

Resumen estadístico

```
cd.iloc[:,3:].describe(include="all").T
```

	count	mean	std	min	25%	50%	75%	max
pH	2779.0	7.081340	1.594089	0.000000	6.095820	7.036752	8.062436	14.000000
Hardness	3270.0	196.370369	32.901984	47.432000	176.844221	197.039446	216.671731	323.124000
TDS	3270.0	12052.082928	14037.159710	320.942611	900.250000	2914.000000	22717.107469	61227.196008
Chloramines	3270.0	7.121847	1.584330	0.352000	6.125756	7.128299	8.115665	13.127000
Sulfate	2493.0	333.760495	41.429448	129.000000	307.694522	333.073364	359.951766	481.030642
Conductivity	3270.0	426.223484	80.857317	181.483754	365.724999	421.926811	481.833045	753.342620
Organic_carbon	3270.0	14.285428	3.308409	2.200000	12.066018	14.219418	16.558601	28.300000
Trihalomethanes	3109.0	66.389487	16.184812	0.738000	55.816510	66.612984	77.330137	124.000000
Turbidity	3270.0	3.967575	0.780036	1.450000	3.440413	3.955122	4.500544	6.739000

Variables numéricas

Variables Categóricas

```
cd.describe(include="object")
```

	Sample Point	Month
count	3270	3270
unique	10	12
top	P1	Abril
freq	327	280

```python
def clasificar_dureza(valor):
    if pd.isna(valor):
        return "Desconocido"
    if valor<60:
        return "Agua blanda"
    if valor<=120:
        return "Agua moderadamente dura"
    if  valor<=180:
        return "Agua dura"
    else:
        return "Agua muy dura"
cd_clasificado["Clasificacion_de_dureza"]=cd_clasificado["Hardness"].apply(clasificar_dureza)
cd
```

```python
def clasificar_uso_agua(cd):
    # Definir límites
    limites = {"Agua Potable": {"dureza": 180,"conductividad": 500,"TDS": 500}, "Agua Industrial": {"dureza": 120,"conductividad": 2000,"TDS": 1500},
        "Agua Agrícola": {"dureza": 180,"conductividad": 3000,"TDS": 2000},"Agua Recreacional": {"dureza": 120,"conductividad": 1000,"TDS": 1000}}
    clasificaciones = []

    for i, row in cd.iterrows():
        dureza = row["Hardness"]
        conductividad = row["Conductivity"]
        TDS = row["TDS"]

        clasificacion = []

        if dureza <= limites["Agua Potable"]["dureza"] and conductividad <= limites["Agua Potable"]["conductividad"] and TDS <= limites["Agua Potable"]["TDS"]:
            clasificacion.append("Agua Potable")
        elif dureza <= limites["Agua Industrial"]["dureza"] and conductividad <= limites["Agua Industrial"]["conductividad"] and TDS <= limites["Agua Industrial"]["TDS"]:
            clasificacion.append("Agua Industrial")
        elif dureza <= limites["Agua Agrícola"]["dureza"] and conductividad <= limites["Agua Agrícola"]["conductividad"] and TDS <= limites["Agua Agrícola"]["TDS"]:
            clasificacion.append("Agua Agrícola")
        elif dureza <= limites["Agua Recreacional"]["dureza"] and conductividad <= limites["Agua Recreacional"]["conductividad"] and TDS <= limites["Agua Recreacional"]["TDS"]:
            clasificacion.append("Agua Recreacional")

        clasificaciones.append(", ".join(clasificacion) if clasificacion else "No Clasificable")

    cd["Usodeagua"] = clasificaciones
    return cd

cd_clasificado = clasificar_uso_agua(cd)
cd_clasificado
```

											Trihalomethanes	Turbidity	Usodeagua	Clasificacion_de_dureza
										783	86.990970	2.963135	No Clasificable	Agua muy dura
										013	56.329076	4.500656	Agua Agrícola	Agua dura
										637	66.420093	3.055934	No Clasificable	Agua muy dura
										524	100.341674	4.628771	No Clasificable	Agua muy dura
4	P5	1996	Abril	9.092223	181.101509	543.000000	6.546600	310.135738	398.410813	11.558279	31.997993	4.075075	No Clasificable	Agua muy dura
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
										14.250875	62.906205	3.361833	No Clasificable	Agua muy dura
										11.083027	38.435151	4.906358	No Clasificable	Agua dura
										9.899115	55.069304	4.613843	No Clasificable	Agua muy dura
										16.217303	28.878601	3.442983	No Clasificable	Agua muy dura
										16.172755	41.558501	4.369264	No Clasificable	Agua moderadamente dura

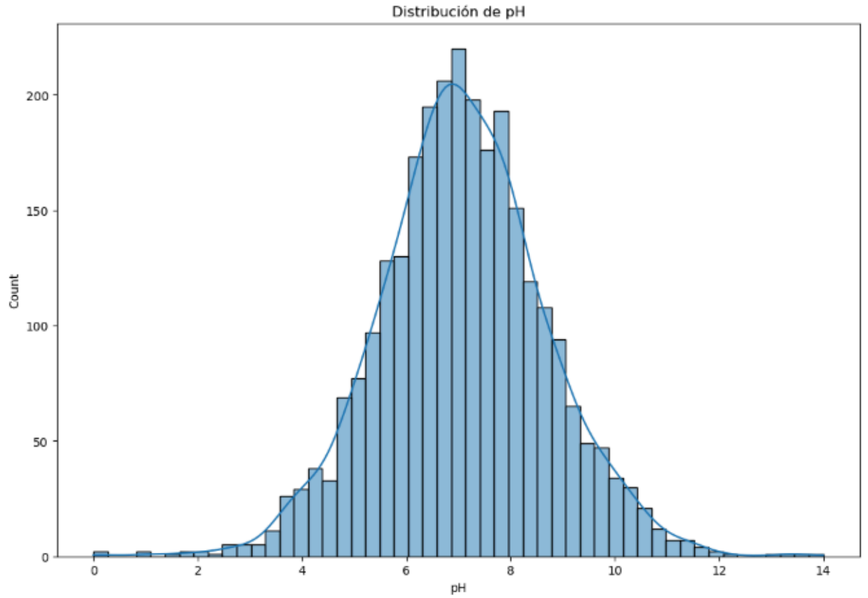
Visualización de Datos

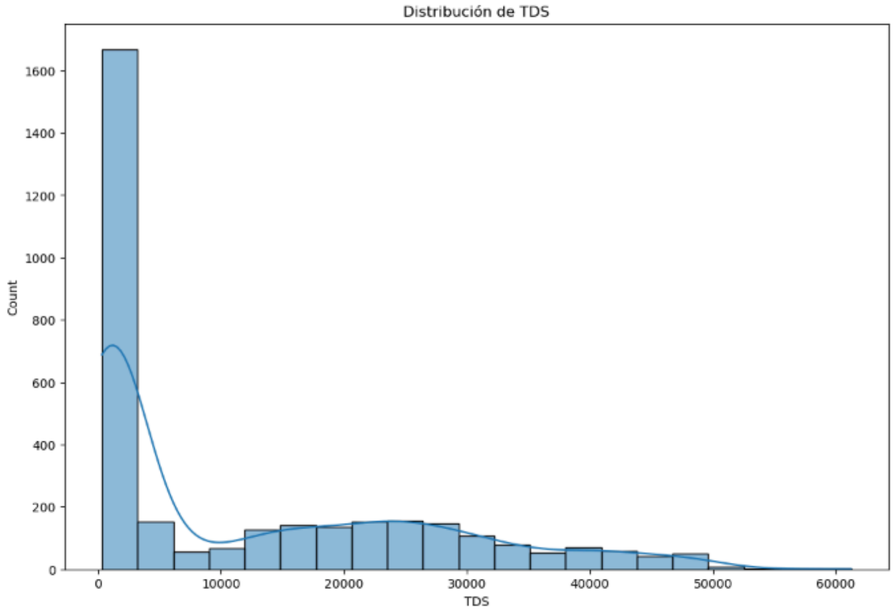
Distribución numérica de cada variable

```
def gen_distribucion(cd_clasificado):
    variables=["pH","Hardness","TDS","Conductivity",
               "Sulfate","Turbidity", "Chloramines"]
    for i in variables:
        fig, ax = plt.subplots(figsize=(12, 8))
        sns.histplot(data=cd_clasificado,
                     x=cd_clasificado[i], kde=True)
        ax.set_xlabel(i)
        ax.set_title(f'Distribución de {i}')

    return
```

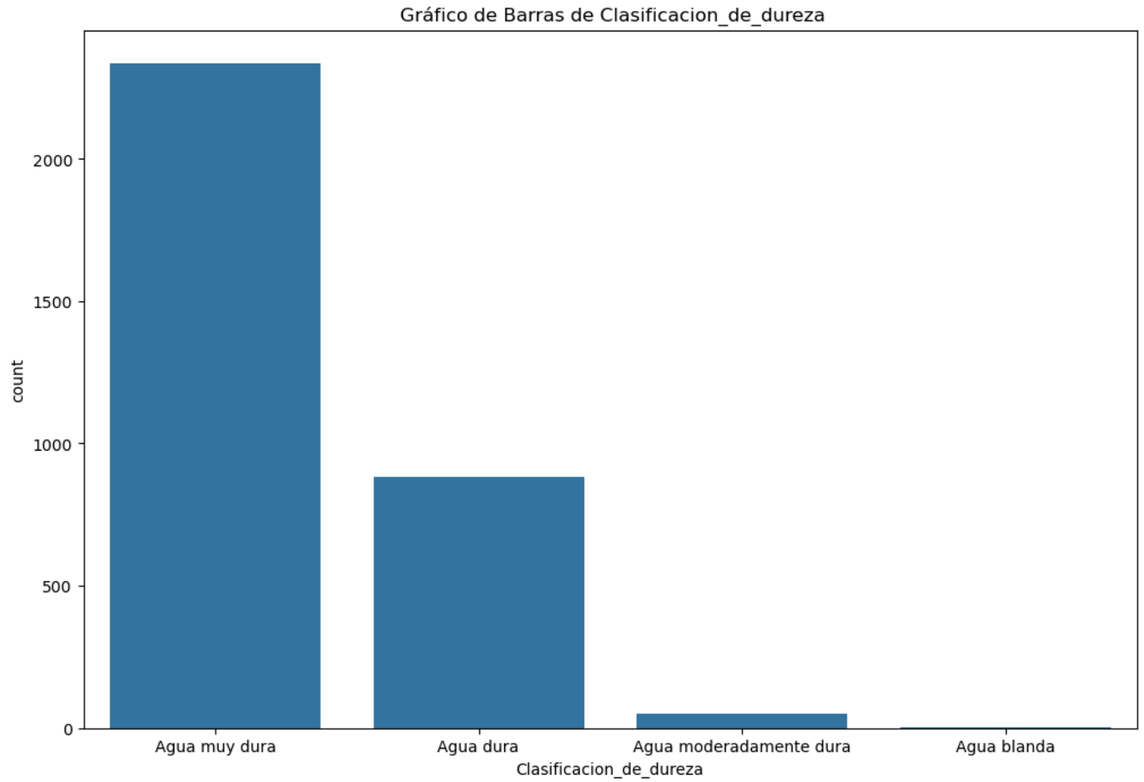
```
gen_distribucion(cd_clasificado)
```

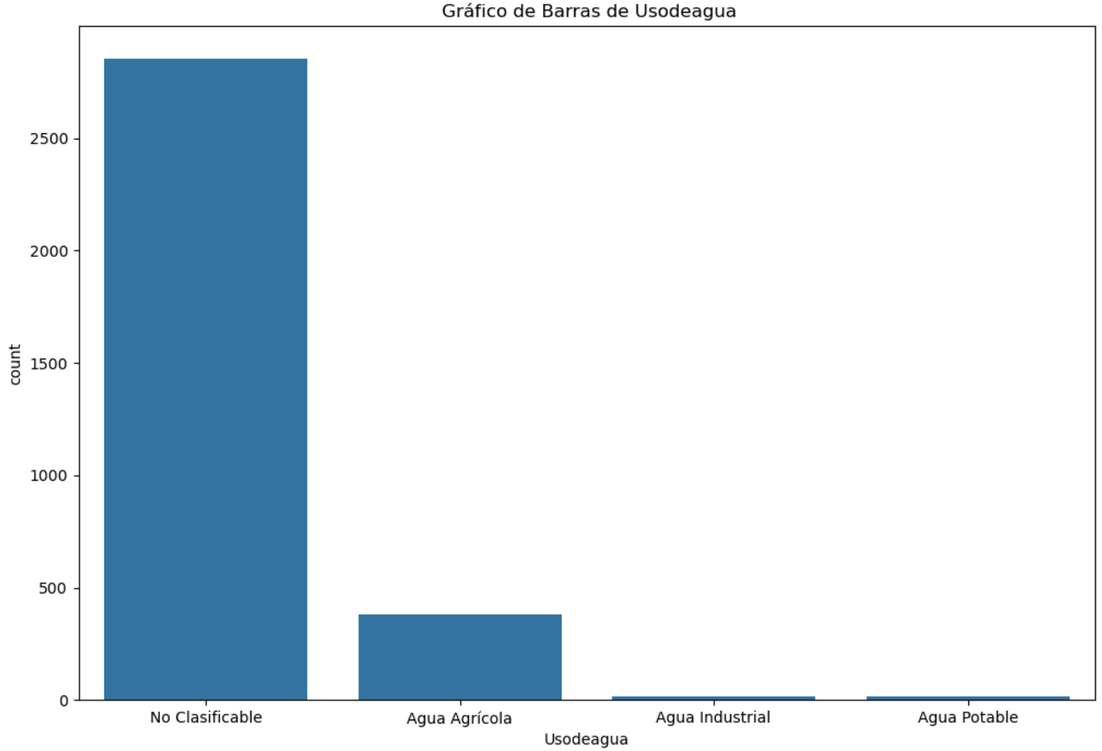




Comportamiento de variables categóricas

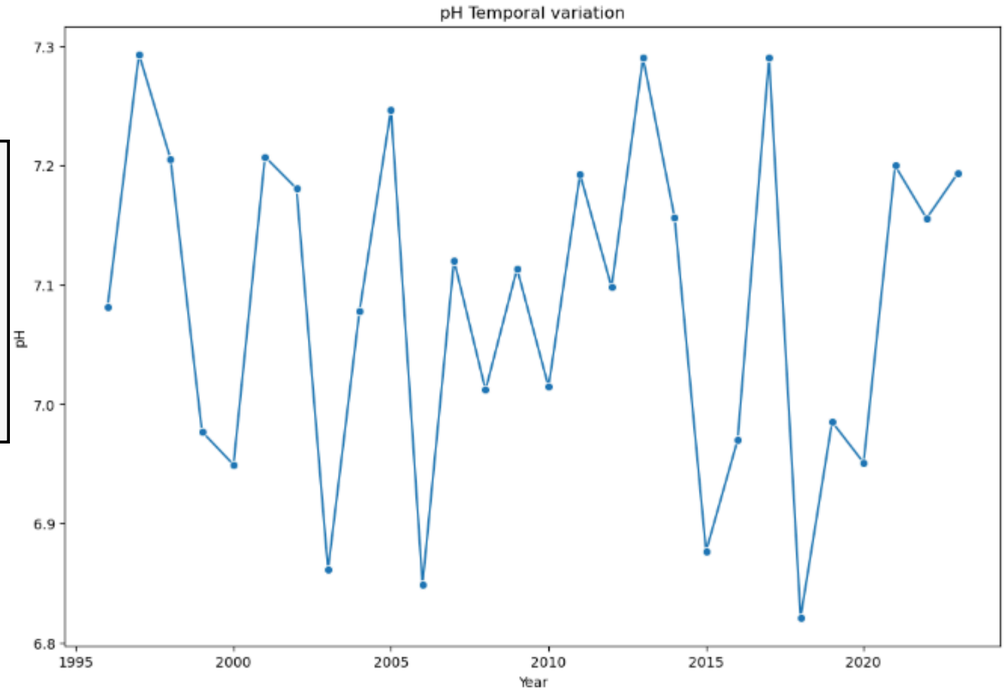
```python
def gen_variables_categorica(cd_clasificado):
    variables=["Usodeagua","Clasificacion_de_dureza"]
    for i in variables:
        fig, ax = plt.subplots(figsize=(12, 8))
        sns.countplot(data=cd, x=i)
        ax.set_title(f' Gráfico de Barras de {i}')
        plt.show()
    return
```

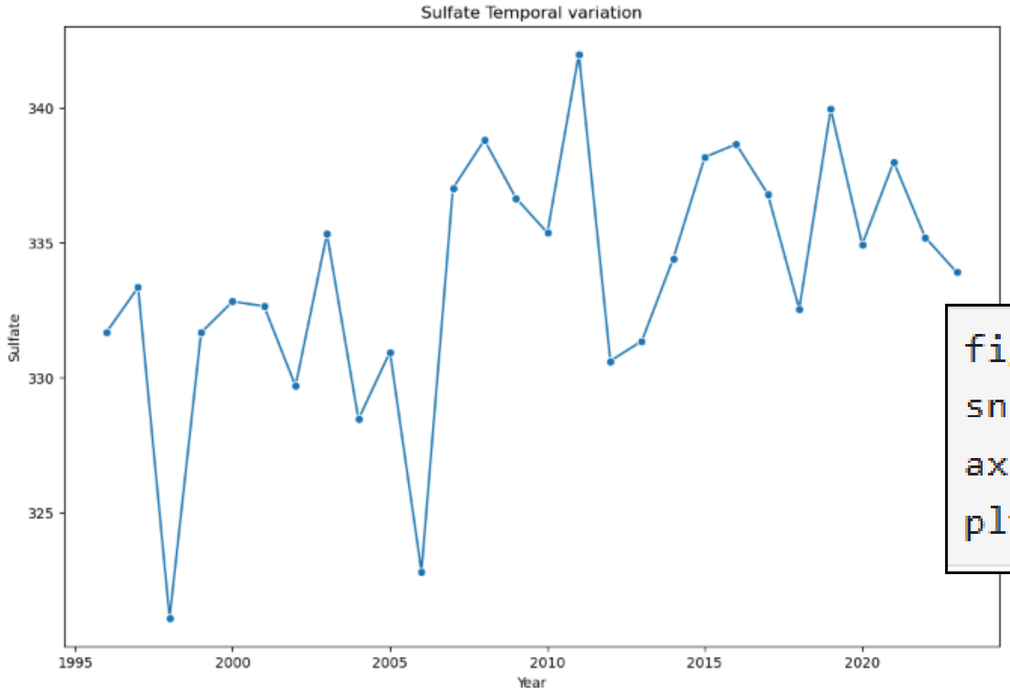




Tendencias de variable a lo largo del tiempo

```
fig, ax = plt.subplots(figsize=(12, 8))
sns.lineplot(data=cd, x="Year", y="pH", marker='o', ci=None)
ax.set_ylabel("pH")
ax.set_title("pH Temporal variation")
plt.show()
```



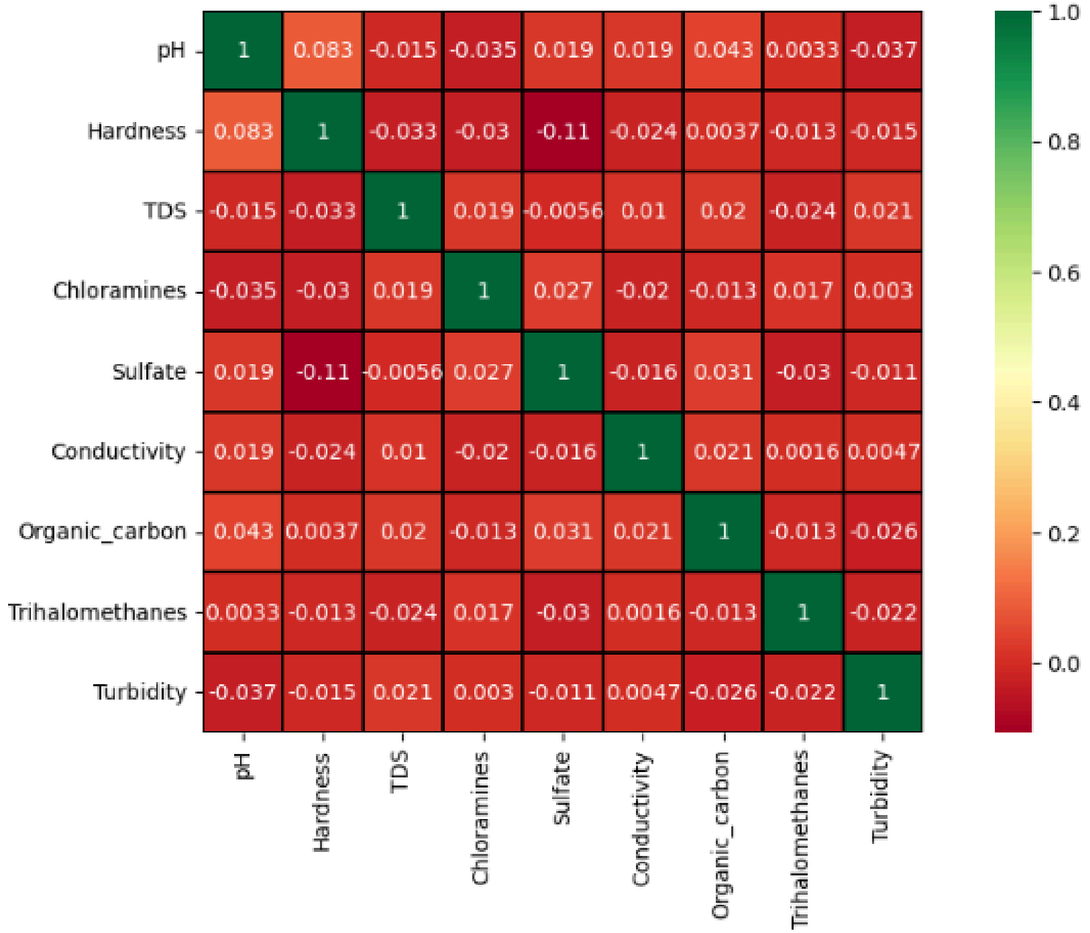



```
fig, ax = plt.subplots(figsize=(12, 8))
sns.lineplot(data=cd, x="Year", y="Sulfate", marker='o', ci=None)
ax.set_title("Sulfate Temporal variation")
plt.show()
```

Mapa de calor

```
fig, ax = plt.subplots(figsize=(12, 6))
# Pearson correlation
cor = a.corr(numeric_only=True)

sns.heatmap(cor, cmap='RdYlGn', annot=True,
            linewidths=0.01, linecolor='black',
            square=True, ax=ax)
plt.show()
```



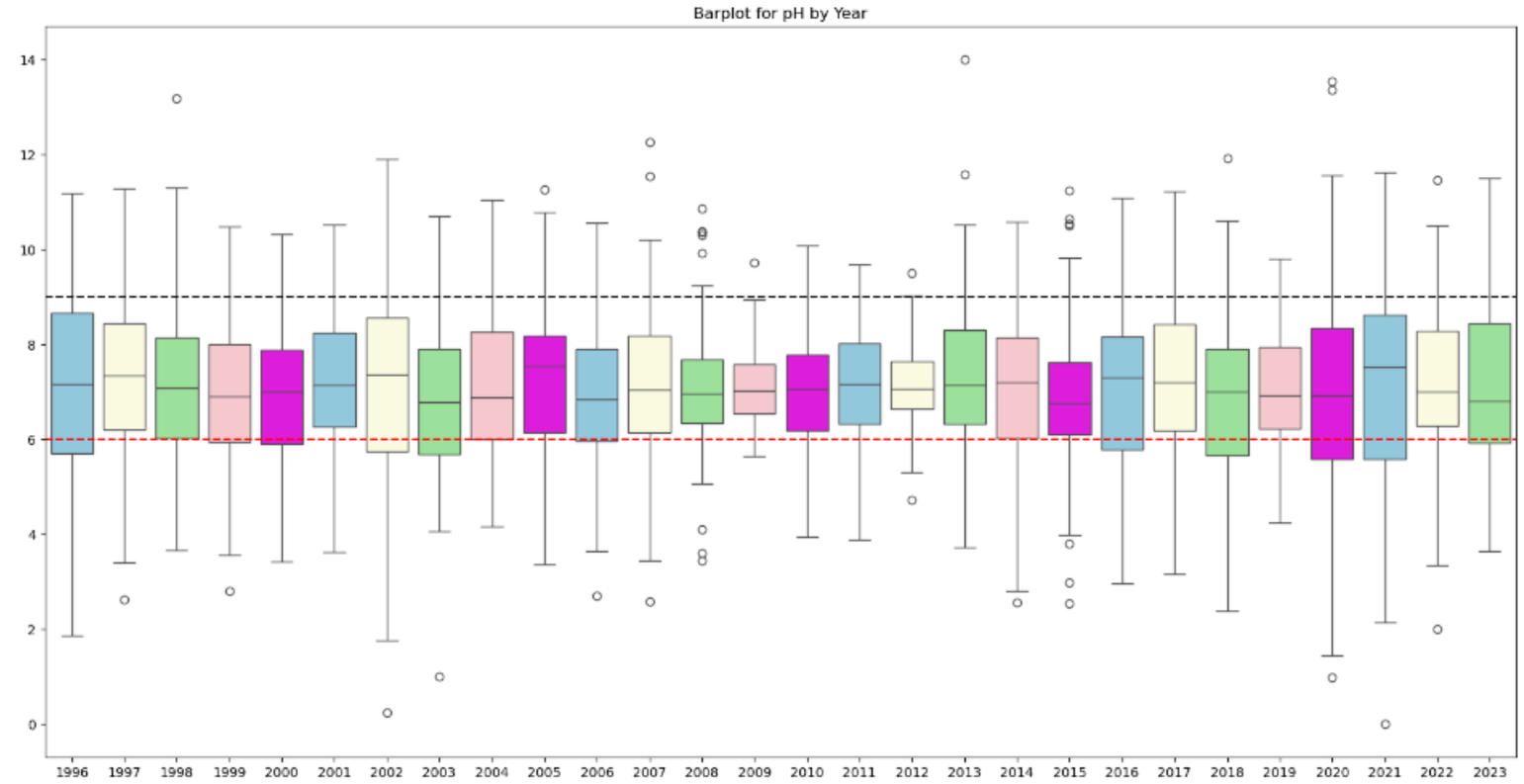
Boxplots

```
fig, ax = plt.subplots(figsize=(20,10))

#Create plot with seaborn
sns.boxplot(data=cd_clasificado, x="Year", y="pH", palette=["skyblue","lightyellow","lightgreen","pink","magenta"])

#Set labels and title
ax.set_title("Barplot for pH by Year")
ymin=ax.axhline(y=6, c="red", ls="--")
ymax=ax.axhline(y=9, c="black", ls="--")

plt.show()
```



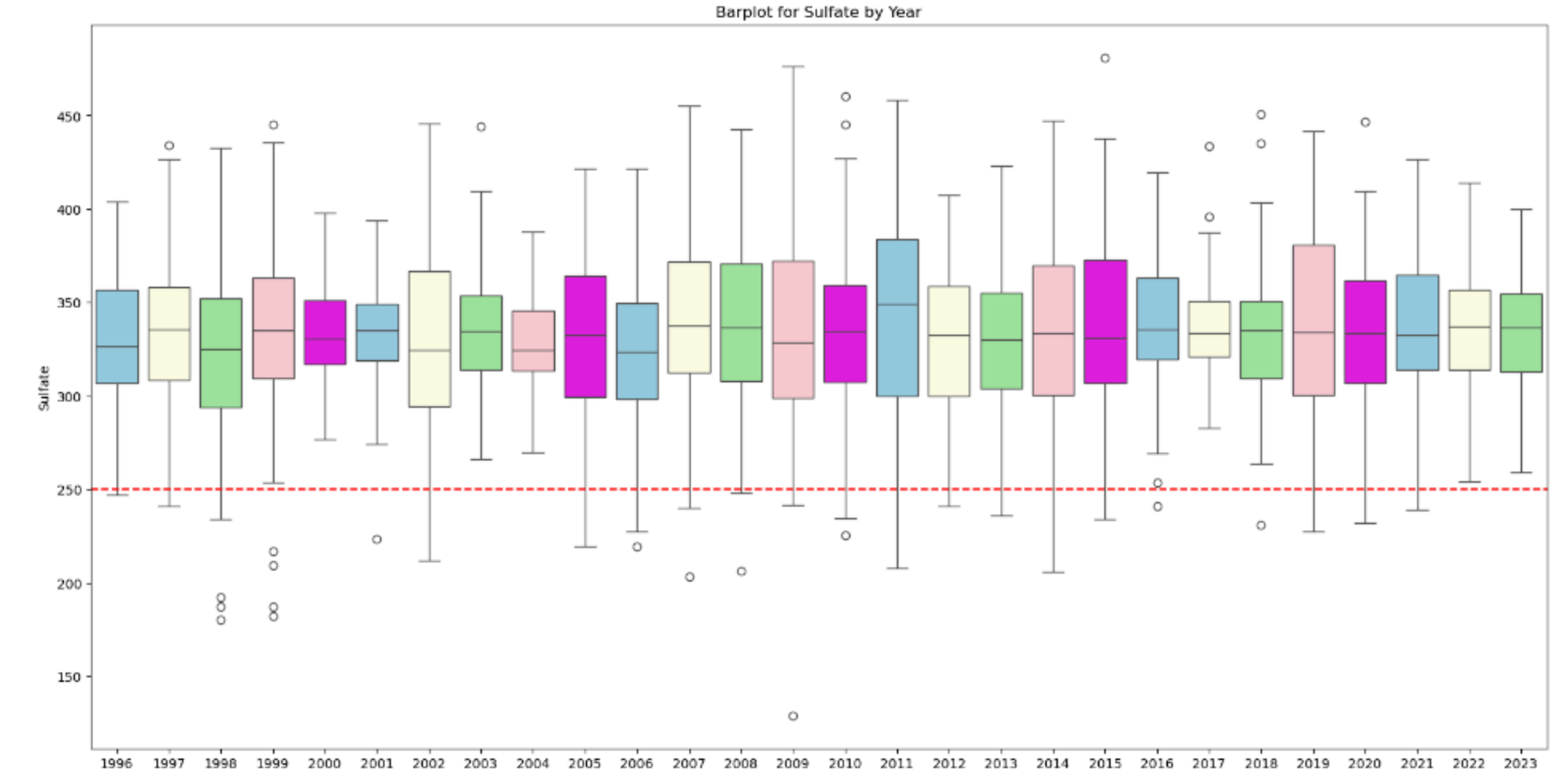
Boxplots

```
fig, ax = plt.subplots(figsize=(20,10))

#Create plot with seaborn
sns.boxplot(data=cd_clasificado, x="Year", y="Sulfate", palette=["skyblue","lightyellow","lightgreen","pink","magenta"])

#Set labels and title
ax.set_title("Barplot for Sulfate by Year")
ax.axhline(y=250, c="red", ls="--")

plt.show()
```



Resultados

Uso del Agua	
No clasificable	2856
Agrícola	381
Industrial	17
Agua Potable	16

Dureza del agua	
Muy dura	2336
Dura	882
Moderadamente Dura	51
Blanda	1

- Valores de pH se encuentran entre los valores permisibles para uso de agua para uso agrícola.
- No existe una correlación entre variables.
- Heterogeniedad en las concentraciones de pH y sulfatos.
- Se observa valores aberrantes en boxplots de pH y sulfatos.

Conclusiones y Recomendaciones

- El 87% de los registros de agua no cumplen con los criterios establecidos.
- El 98% de los puntos de muestreo corresponden a aguas muy duras y duras.
- Asimetría en histograma de TDS.
- El pH no es un factor limitante para el uso del agua en riego agrícola.
- No existe una correlación lineal entre las variables.

Se recomienda realizar nuevas campañas de muestreo donde se analicen diversos parámetros de calidad de agua.

THANK

YOU