Informe Trabajo Práctico Especial

Fundamentos de la Ciencia de Datos

Integrantes:

- José María Goicoechea
- Ramiro Maestrojuán

Índice

Objetivo

El objetivo de este trabajo es poder comprender y analizar correctamente un dataset, aplicando las herramientas aprendidas durante la cursada para poder buscar relaciones entre sus variables, permitiendo así plantear diferentes hipótesis para posteriormente probarlas mediante testeos.

Metodología

Recibimos un dataset con mediciones del agua en diferentes zonas del Río de la Plata tomadas en el año 2022, habiéndose tomado una muestra por cada estación del año en cada una de las zonas.

La idea de este dataset es poder evaluar la calidad del agua de la muestra tomada en base a diferentes parámetros (como presencia de contaminantes, olores, diferentes elementos químicos en el agua, etc), para ello realizamos una serie de hipótesis sobre las muestras, basándonos en análisis que realizamos sobre los datos de las mismas. Debido a la meta del dataset, buscamos determinar hipótesis que tengan relación directa con la calidad del agua.

Para iniciar, al dataset lo encontramos con los datos “crudos” (valores repetidos, idénticos pero expresados de diferentes maneras, valores nulos expresados con palabras, y demás problemas del tipo) por lo que seguimos todo un proceso para poder dejar el dataset limpio y así poder analizarlo correctamente:

- Identificamos los datos incorrectos (nulos no representados cono NaN, repetidos, etc)
- Una vez identificados los datos “sucios” procedemos a limpiarlos
    - Reemplazamos las distintas representaciones de nulos con NaN.
    - Los valores iguales pero escritos de distinta manera los acomodamos de forma que sean todo lo mismo.
    - Las variables categóricas las pasamos a una representación booleana o binaria.
    - Las filas que tenían un 60% o más de datos nulos las eliminamos.
    - Reemplazamos los NaN dependiendo de la distribución de los datos de cada columna:
        - Si la columna tenía una distribución normal de los datos reemplazamos los NaN por el promedio de los valores de la columna.
        - Si no tenía distribución normal lo reemplazamos por el valor más cercano.
    - Las columnas que tenían valores como rangos (por ej: <0.1,<10 y así) lo reemplazamos por el valor más cercano en ese rango (<0.5=0.4).
    - Los datos se castearon a tipos de datos numéricos.

Contenido del dataset

Nos encontramos con una gran cantidad de muestras y de componentes en cada una. Se pudo ver que se realizaron 168 mediciones en diferentes sitios, y por lo general, una medición por estación del año. Además se vió que se cuentan con 30 variables en cada muestra: sitios: Localización específica donde se realizó el muestreo del agua:

- Código: Identificador único para cada muestra o estación de muestreo.
- Fecha: Fecha en la que se tomó la muestra de agua. Una por estación.
- Año: Año en que se realizó el muestreo.(2022)
- Campaña: Nombre o número de la campaña de monitoreo en la que se realizó el muestreo. Estación del año.
- Tem_agua: Temperatura del agua en grados Celsius.
- Tem_aire: Temperatura del aire en grados Celsius.
- Od: Oxígeno disuelto, medido en miligramos por litro (mg/L), esencial para la vida acuática.
- Ph: Medida de la acidez o alcalinidad del agua, en una escala de 0 a 14.
- Olores: Presencia de olores en el agua, que puede indicar contaminación.
- Color: Color del agua, que puede ser un indicador de la calidad del agua.
- Espumas: Presencia de espumas en la superficie del agua, que puede ser un signo de contaminación.
- Mat_susp: Materia suspendida, que se refiere a partículas sólidas que flotan en el agua.
- Colif_fecales_ufc_100ml: Unidades formadoras de colonias de coliformes fecales en 100 ml de agua, un indicador de contaminación fecal.
- Escher_coli_ufc_100ml: Unidades formadoras de colonias de Escherichia coli en 100 ml de agua, otro indicador de contaminación fecal.
- Enteroc_ufc_100ml: Unidades formadoras de colonias de enterococos en 100 ml de agua, que también indican contaminación fecal.
- Nitrato_mg_l: Concentración de nitratos en miligramos por litro (mg/L), que puede indicar contaminación por fertilizantes.
- Nh4_mg_l: Concentración de amonio en miligramos por litro (mg/L), que puede ser un indicador de contaminación orgánica.
- P_total_l_mg_l: Fósforo total en miligramos por litro (mg/L), que incluye todas las formas de fósforo en el agua.

- ➢ Fosf_ortofos_mg_l: Concentración de ortofosfatos en miligramos por litro (mg/L), que es un nutriente importante.
- ➢ Dbo_mg_l: Demanda biológica de oxígeno en miligramos por litro (mg/L), que mide la cantidad de oxígeno requerido por microorganismos para descomponer materia orgánica.
- ➢ Dqo_mg_l: Demanda química de oxígeno en miligramos por litro (mg/L), que mide la cantidad total de oxígeno requerido para oxidar materia orgánica e inorgánica.
- ➢ Turbiedad_ntu: Turbidez del agua medida en unidades NTU (Nephelometric Turbidity Units), que indica la claridad del agua.
- ➢ Hidr_deriv_petr_ug_l: Hidrocarburos derivados del petróleo en microgramos por litro (µg/L), que indican contaminación por productos petroleros.
- ➢ Cr_total_mg_l: Concentración total de cromo en miligramos por litro (mg/L), un metal pesado que puede ser tóxico.
- ➢ Cd_total_mg_l: Concentración total de cadmio en miligramos por litro (mg/L), otro metal pesado que es tóxico en altas concentraciones.
- ➢ Clorofila_a_ug_l: Concentración de clorofila a en microgramos por litro (µg/L), que indica la cantidad de fitoplancton en el agua.
- ➢ Microcistina_ug_l: Concentración de microcistinas en microgramos por litro (µg/L), que son toxinas producidas por ciertas algas.
- ➢ Ica: Índice de calidad del agua, que puede ser un valor calculado para evaluar la calidad general del agua.
- ➢ Calidad_de_agua: Clasificación general de la calidad del agua basada en parámetros medidos.

Con el dataset ya limpio comenzamos con el análisis univariado, el cual consiste en analizar cada columna del dataset (cada variable) con diferentes herramientas como gráficos, frecuencia de los datos, valores, promedios e informándonos sobre cada una de esas variables para así poder detectar diferentes anomalías en cada una de las columnas. Esto nos ayudaría en el análisis bivariado y multivariado para la formulación de hipótesis.

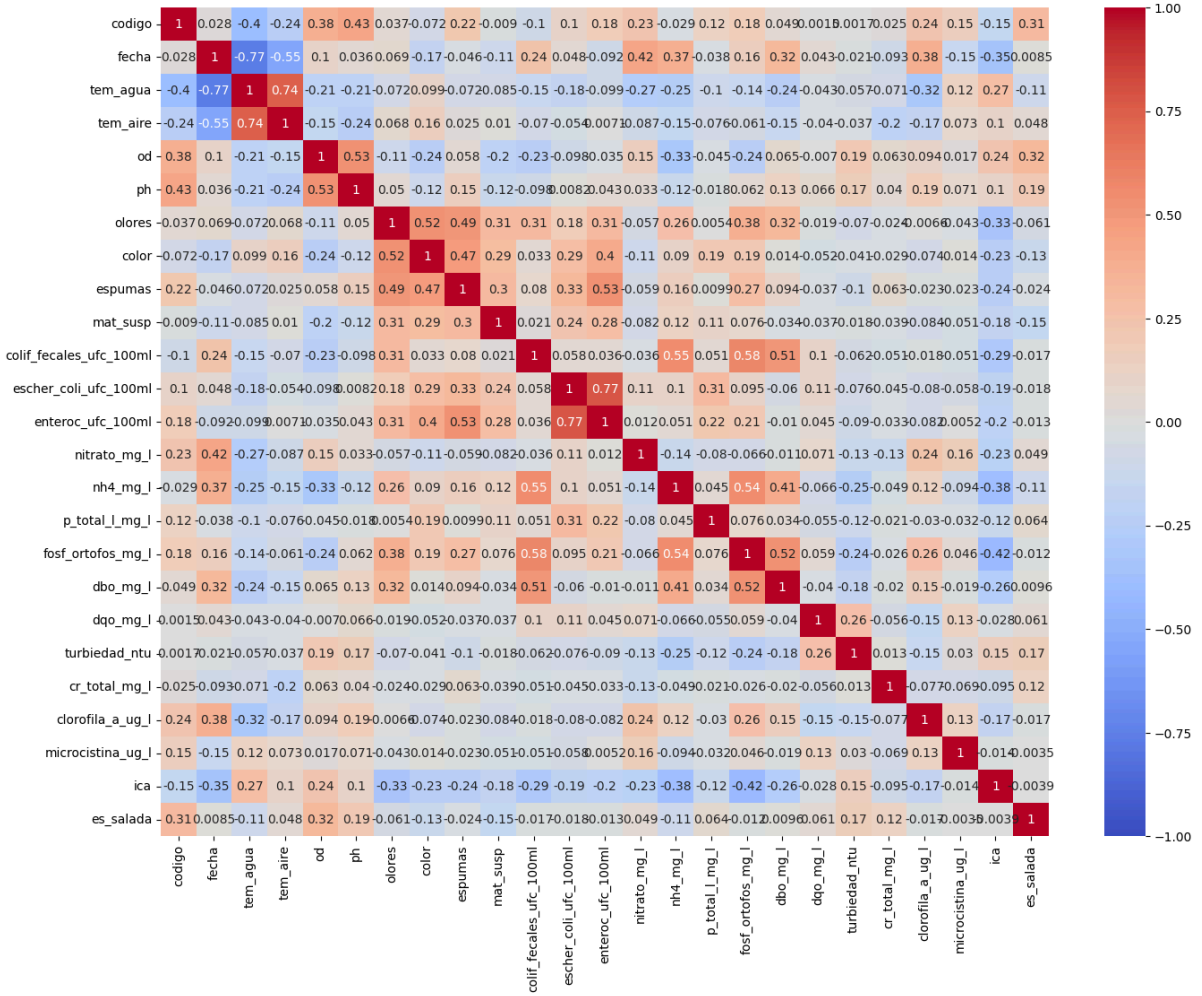
Nos encontramos ciertas variables que contenían en alguna muestra concreta valores extraordinarios, con valores fuera de lo normal en comparación a otras. Debido a que estos, en todos sus casos, se trataron de situaciones con números coherentes en cuánto a la naturaleza de ellos, se decidió proceder sin eliminarlos.

Terminado el análisis univariado comenzamos con el análisis bivariado y multivariado.
En este tipo de análisis busca identificar y describir relaciones entre dos o más variables, por lo que aplicando diferentes herramientas de análisis (gráficos, matriz de correlaciones, etc) pudimos observar relaciones entre variables que podrían afectar o no la calidad del

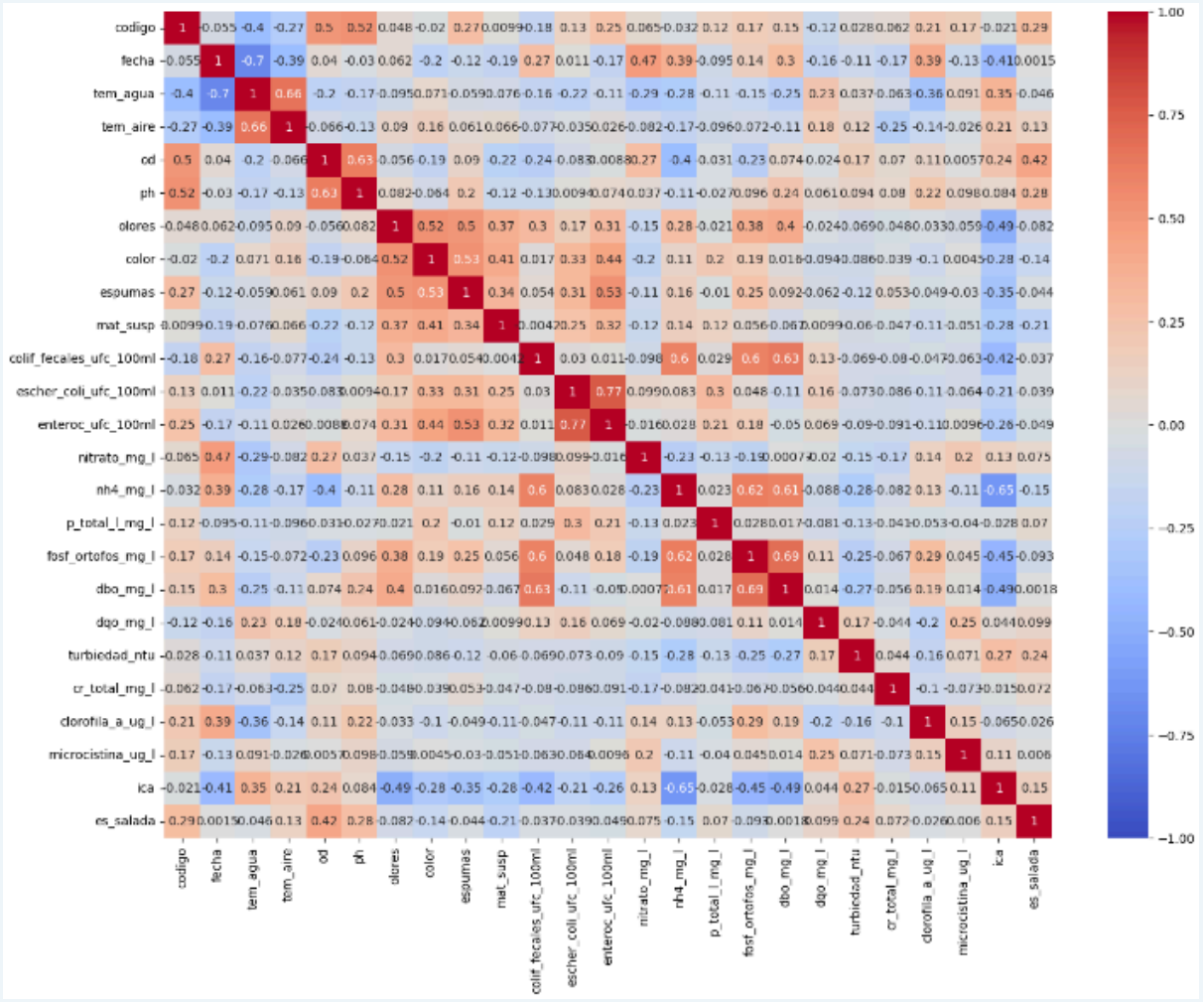
agua, así que a partir de este análisis pudimos plantear diferentes hipótesis. Luego de realizar el análisis bivariado nos encontramos con una disyuntiva, ya que nos encontramos con suficiente material con el que trabajar, así que nos sumergimos en este, y dejamos el análisis multivariado para cuándo lo terminemos de exprimir.

Con el análisis terminado surgen diferentes hipótesis, las cuales a partir de los gráficos anteriormente mencionados se plantean las mismas, una vez planteadas las mismas lo que se busca es rechazar la hipótesis nula (es decir que no se cumpla lo contrario a lo que plantea nuestra hipótesis) mediante el test correspondiente en función de la distribución y relación de los datos:test-t, Mann-Whitney U o Kruskall-Wallis) de manera que se valide nuestra hipótesis.

A continuación, podemos ver un gráfico donde se detallan las correlaciones de los atributos con los que contamos.
Dicho sea de paso, se decidió agregar una columna para diferenciar las muestras de las zonas de agua salada con las muestras de las zonas de agua dulce debido a que el Río de La Plata en su desembocadura contiene agua salada, mientras que en el resto de su cuerpo cuenta con agua dulce. Teniendo esto en cuenta, en el dataset hay ciertas variables que varían según el tipo de agua donde se encuentra, tal como el ph, por lo que buscamos diferenciar las muestras para ver si podríamos identificar el comportamiento de los atributos según la zona geográfica. También se contó con una división de la columna ‘calidad_de_agua’ para así poder visualizar más notoriamente ciertos comportamientos según los grupos.
Esta es una de 4 matrices que se contaron para el análisis. Las demás, son con los mismos datos, pero estos se encuentran divididos según la calidad de las muestras.



matriz de agua extremadamente deterioradas(usada como fundamento en ciertas hipótesis):



Hallazgos

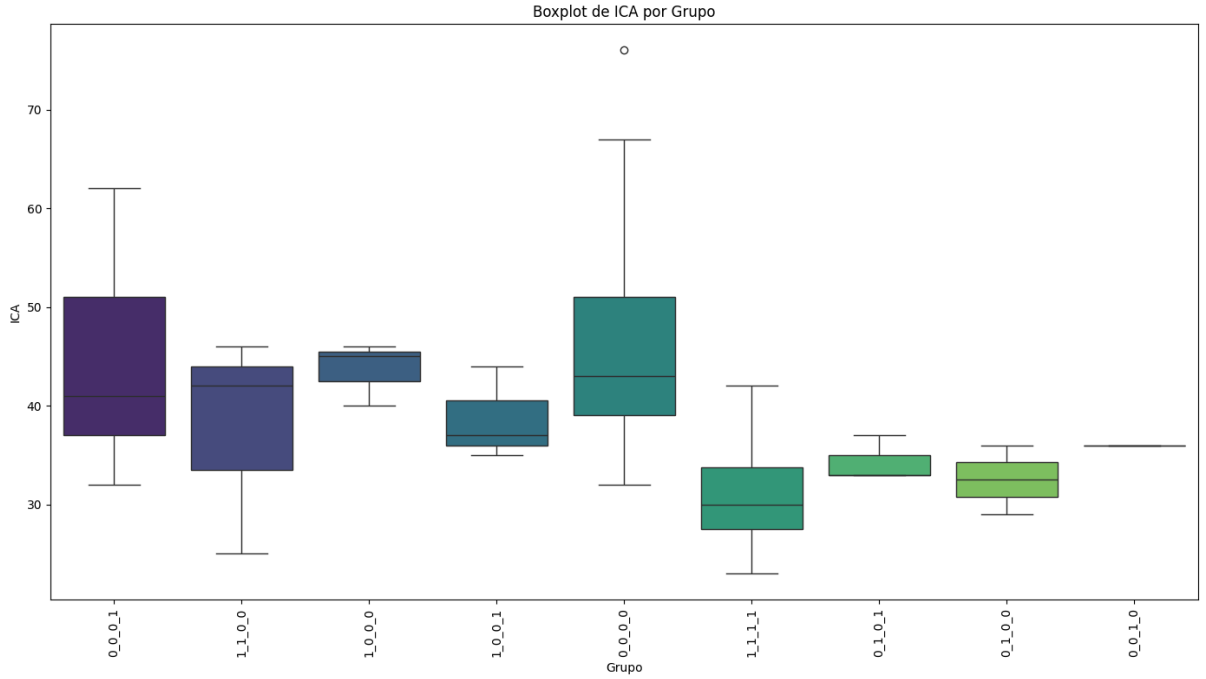
Con el análisis mencionado anteriormente pudimos hacer hallazgos interesantes sobre la calidad del agua en relación sobre que variables en conjunto pueden o no afectarla, pero el tipo de agua no nos arrojó en detalle grandes correlaciones. Estos hallazgos los realizamos mediante el planteo de las siguientes hipótesis:
Notar que en cada una marcamos los gráficos que nos hicieron iniciar nuestros planteos. Los cálculos y justificaciones son debidamente mostrados en el Jupyter notebook adjuntado. Además de las siguientes cuatro hipótesis se realizaron dos que no fueron analizadas profundamente. Se pudo llegar a cinco de ellas en base a un análisis bivariado, y una mediante análisis multivariado, por lo que se decidió no incluir posteriormente metodología de disminución de dimensionalidad y demás que no haya sido utilizada en su anterioridad más allá de la regresión lineal múltiple que se realiza en la segunda hipótesis no analizada.

Hipótesis 1: La ausencia de olores, color, materia suspendida y de espumas indica un aumento significativo en el índice de calidad de agua en comparación a las muestras que cuentan con algunos de ellos.

Al analizar los datos detectamos que la presencia de estas 4 variables en el agua podría afectar o tendría relación con el deterioro de la calidad de la misma, mientras más presencia de esta variables en conjunto más afectado se ve el ICA por lo que planteamos la hipótesis.

Se dividieron las muestras en dos grupos: las aguas que no contaban con presencia de ninguna de las cuatro variables, o la muestra contaba con al menos una en su análisis. Así analizamos el comportamiento del ICA en cada grupo.

Los datos no tenían una distribución normal, pero si eran homocedásticos por lo que pudimos validar la hipótesis por el test de Mann-Whitney U. Al aplicar el test. nos dió un p-valor menor a 0.05 lo cual indica que se rechaza la hipótesis nula, esto quiere decir que nuestra hipótesis se valida.



Teniendo en cuenta que se agrupó por las distintas combinaciones de dichos elementos y que un 1 denota que sí están presentes, y un 0 denota que no lo están se puede detallar aquí que el coeficiente aumenta en base a que disminuya la cantidad de los atributos mencionados.

```python
# Test de Mann-Whitney U para comparar el índice ICA en el grupo que no tiene ninguno de los elementos categóricos contaminantes con las
stat, p = stats.mannwhitneyu(ausencia_de_elem_sensibles, presencia_de_elem_sensibles)
print(f"Test de Mann-Whitney U para ICA: Estadístico={stat:.3f}, p-valor={p:.3f}")

# Interpretación de los resultados
alpha = 0.05  # Nivel de significancia
if p > alpha:
    print("No hay suficiente evidencia para rechazar la hipótesis nula.")
    print("No hay una diferencia significativa en el índice de Calidad del Agua entre aguas que no presentan elementos categóricos conta
else:
    print("Se rechaza la hipótesis nula.")
    print("Existe una diferencia significativa en el índice de Calidad del Agua entre aguas que no presentan elementos categóricos conta
```

✓ 0.0s Python

```
Test de Mann-Whitney U para ICA: Estadístico=2856.000, p-valor=0.002
Se rechaza la hipótesis nula.
Existe una diferencia significativa en el índice de Calidad del Agua entre aguas que no presentan elementos categóricos contaminantes y agu
```

HIPÓTESIS 2: La calidad del agua es notablemente diferente si la muestra se toma en invierno.

Luego de analizar distintos gráficos notamos que en las estaciones frías (invierno, otoño) existe un deterioro en el índice ICA, es decir, en las muestras tomadas en las estaciones frías se observó una peor calidad del agua.

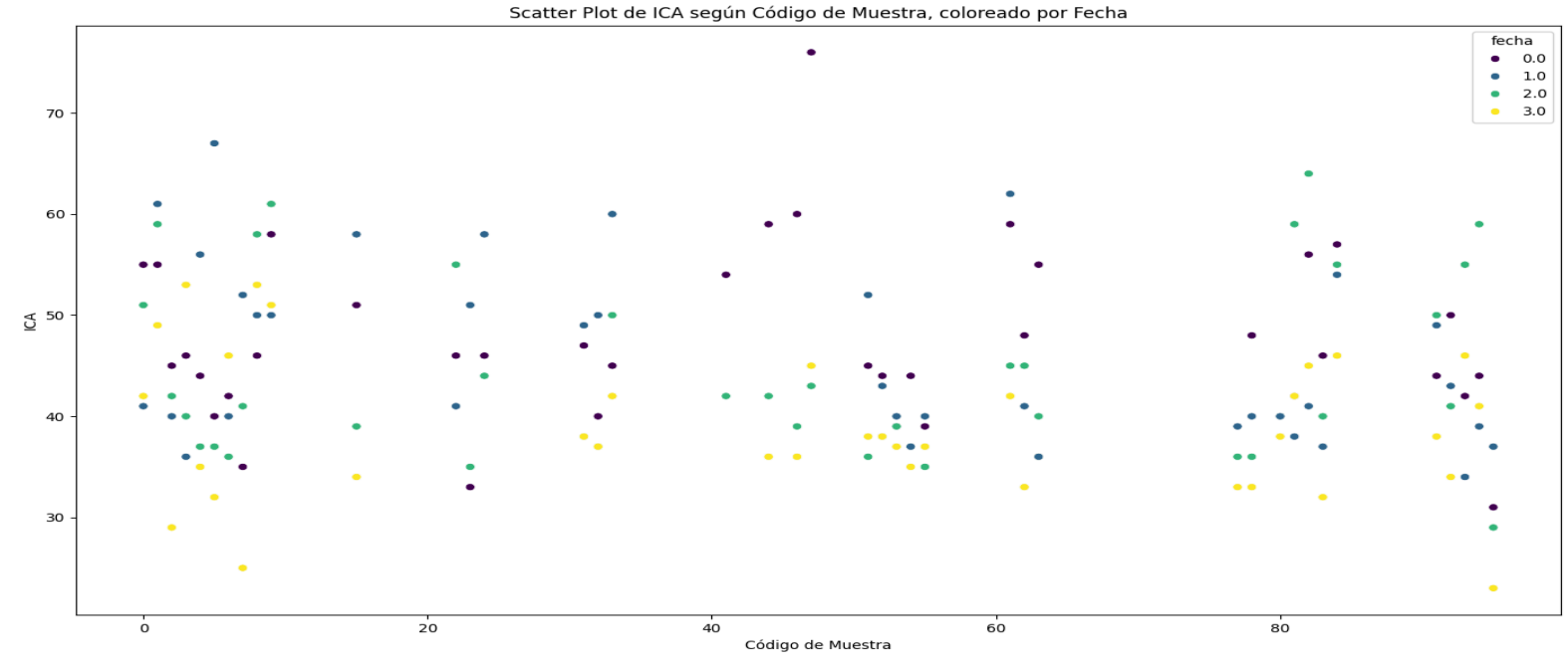
Decidimos entonces plantearnos si realmente la calidad del agua empeora notablemente en invierno.

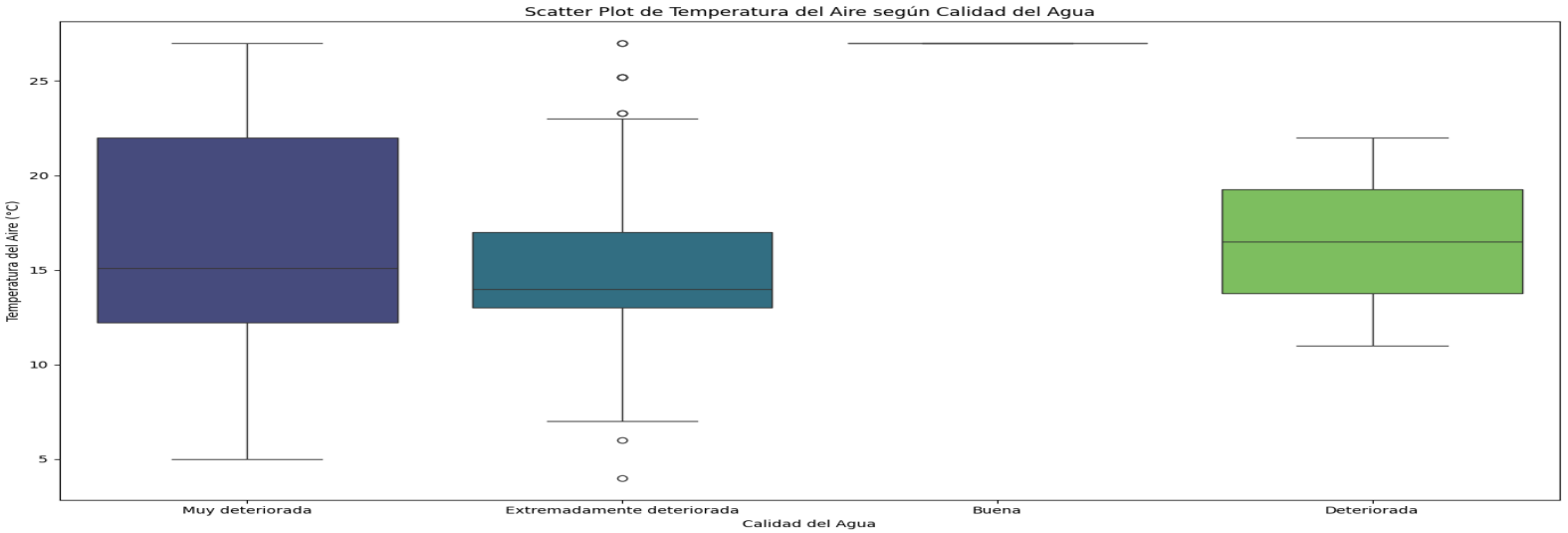
Como se puede ver en los gráficos a continuación, no se notaron comportamientos necesariamente diferentes entre las demás estaciones, solo en invierno, razón por la cuál se resuelve esta hipótesis.

Para validar la hipótesis por el test-t buscamos la normalidad de los datos dividiendo en dos grupos a las muestras: si son o no son de Invierno. Finalmente, no se validó. Luego, buscamos la homocedasticidad de los datos lo cual pudimos validar, entonces planteamos el test de Mann-Whitney U.

La hipótesis nula del test de Mann-Whitney U sugiere que no hay una diferencia significativa en términos de tendencia central entre los dos grupos que se están comparando. En nuestra hipótesis esto se traduce a que la distribución de la calidad del agua de las muestras tomadas en invierno y en otras estaciones son idénticas.

Al probarlo nos arroja un p-valor menor a 0.05, por lo que se rechazó la hipótesis nula y confirmamos que existe una diferencia significativa entre las muestras tomadas en invierno y las demás estaciones. Nota: En los dos primeros gráficos a continuación un 0 denota Verano, un 1 primavera, 2 Otoño y 3 Invierno.






```python
# Test de Mann-Whitney U para comparar ICA entre muestras de invierno y las de las otras estaciones
stat, p = stats.mannwhitneyu(muestras_invernales, muestras_no_invernales)
print(f"Test de Mann-Whitney U para ICA: Estadístico={stat:.3f}, p-valor={p:.3f}")

# Interpretación de los resultados
alpha = 0.05  # Nivel de significancia
if p > alpha:
    print("No hay suficiente evidencia para rechazar la hipótesis nula.")
    print("No hay una diferencia significativa en el índice de calidad del agua entre las muestras de Invierno con las muetras de las d
else:
    print("Se rechaza la hipótesis nula.")
    print("Existe una diferencia significativa en el índice de calidad del agua entre las muestras de Invierno con las muetras de las d
```

✓ 0.0s Python

```
Test de Mann-Whitney U para ICA: Estadístico=1163.000, p-valor=0.000
Se rechaza la hipótesis nula.
Existe una diferencia significativa en el índice de calidad del agua entre las muestras de Invierno con las muetras de las demás estaciones
```

HIPÓTESIS 3: La calidad del agua se deteriora en función de la presencia de coliformes fecales y la demanda bioquímica de oxígeno.

Al realizar el análisis bivariado (más específicamente la matriz de correlación) observamos una correlación alta entre las variables mencionadas por lo que decidimos analizar si existía una relación entre el deterioro de la calidad del agua y el aumento de estas variables.

Vimos una correlación entre la presencia de coliformes fecales y la demanda de oxígeno en la matriz de las aguas extremadamente deterioradas. Además, se vió una relación lineal negativa de ICA(coeficiente de calidad del agua) con ambas, que si bien fue de menor coeficiente, son de las mayores relaciones de esta última variable.

Luego de analizar estos gráficos donde se buscaba esta relación decidimos plantear la hipótesis.

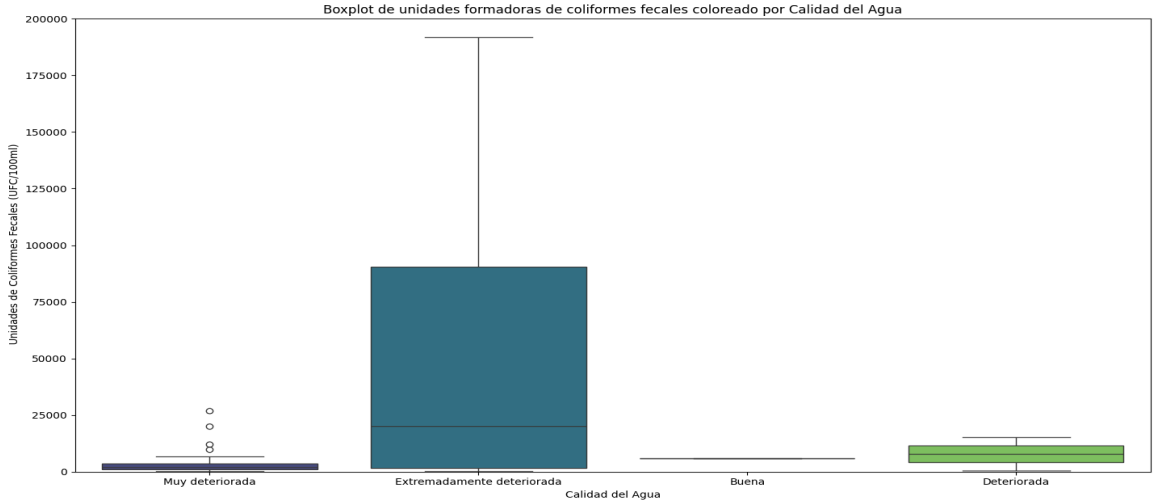
Se realizaron dos tests distintos debido a que Mann-Whitney, el test-t y Kruskal-Wallis nos sirvieron solo para ver el comportamiento en relación a una variable.

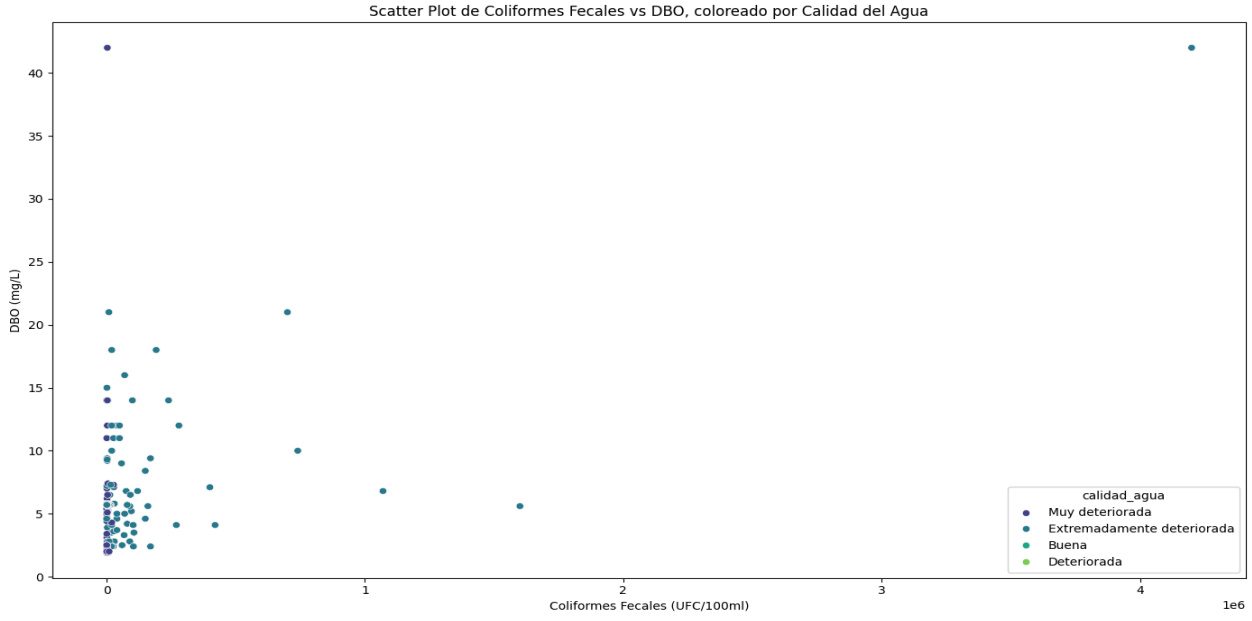
La hipótesis nula para el test de Kruskal-Wallis establece que que las distribuciones de las diferentes muestras son iguales. Específicamente, en este caso, la hipótesis nula establece que la distribución de coliformes fecales es la misma en aguas extremadamente

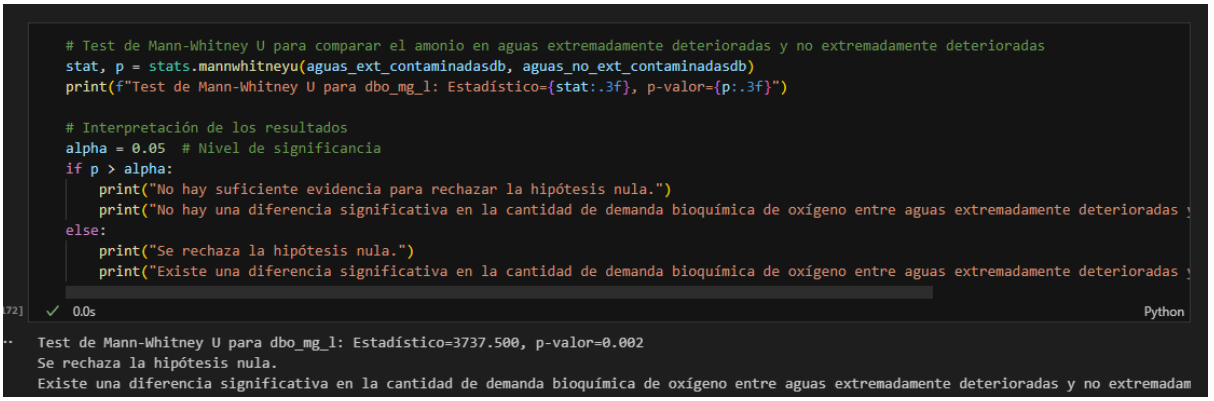
contaminadas y las otras categorías de contaminación del agua (deteriorada y muy deteriorada).

Para la presencia de coliformes fecales no pudimos validar ni la distribución normal de los datos ni la homocedasticidad, por lo que debimos validar la hipótesis por el test de Kruskal. Al validarla nos dio un p-valor a 0.05 por lo que se rechazó la hipótesis nula y validamos nuestra hipótesis de que Existe una diferencia significativa en el colif_fecales_ufc_100ml entre aguas que están extremadamente contaminadas y aguas que no lo estén.

En cuánto a la demanda bioquímica de oxígeno, al dividir de igual manera entre(aguas que están o no extremadamente contaminadas) se pudo ver que estos tenían un comportamiento homocedástico, pero no normal. Se realiza test de Mann-Whitney u.





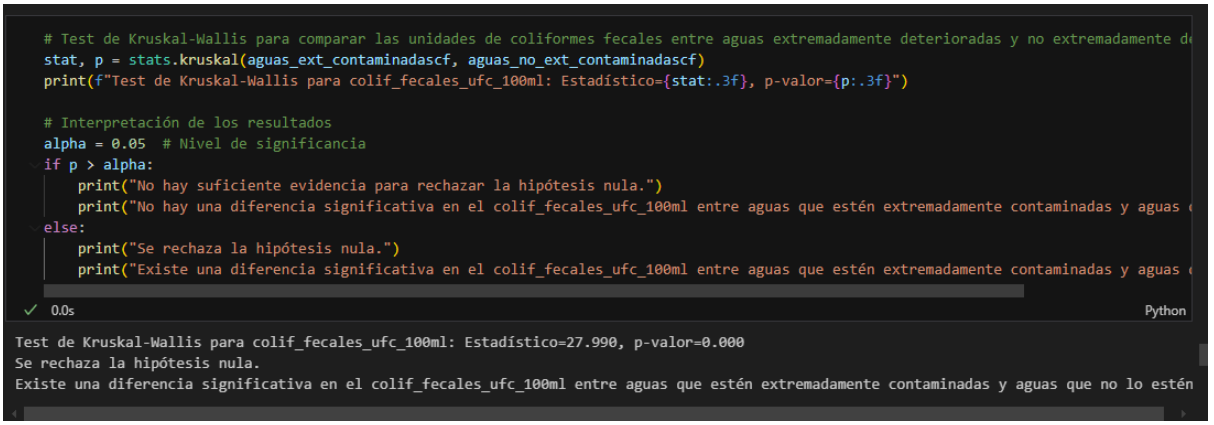


```
# Test de Mann-Whitney U para comparar el amonio en aguas extremadamente deterioradas y no extremadamente deterioradas
stat, p = stats.mannwhitneyu(aguas_ext_contaminadasdb, aguas_no_ext_contaminadasdb)
print(f"Test de Mann-Whitney U para dbo_mg_l: Estadístico={stat:.3f}, p-valor={p:.3f}")

# Interpretación de los resultados
alpha = 0.05  # Nivel de significancia
if p > alpha:
    print("No hay suficiente evidencia para rechazar la hipótesis nula.")
    print("No hay una diferencia significativa en la cantidad de demanda bioquímica de oxígeno entre aguas extremadamente deterioradas
else:
    print("Se rechaza la hipótesis nula.")
    print("Existe una diferencia significativa en la cantidad de demanda bioquímica de oxígeno entre aguas extremadamente deterioradas
```

✓ 0.0s Python

```
Test de Mann-Whitney U para dbo_mg_l: Estadístico=3737.500, p-valor=0.002
Se rechaza la hipótesis nula.
Existe una diferencia significativa en la cantidad de demanda bioquímica de oxígeno entre aguas extremadamente deterioradas y no extremadam
```



```
# Test de Kruskal-Wallis para comparar las unidades de coliformes fecales entre aguas extremadamente deterioradas y no extremadamente d
stat, p = stats.kruskal(aguas_ext_contaminadascf, aguas_no_ext_contaminadascf)
print(f"Test de Kruskal-Wallis para colif_fecales_ufc_100ml: Estadístico={stat:.3f}, p-valor={p:.3f}")

# Interpretación de los resultados
alpha = 0.05  # Nivel de significancia
if p > alpha:
    print("No hay suficiente evidencia para rechazar la hipótesis nula.")
    print("No hay una diferencia significativa en el colif_fecales_ufc_100ml entre aguas que estén extremadamente contaminadas y aguas
else:
    print("Se rechaza la hipótesis nula.")
    print("Existe una diferencia significativa en el colif_fecales_ufc_100ml entre aguas que estén extremadamente contaminadas y aguas
```

✓ 0.0s Python

```
Test de Kruskal-Wallis para colif_fecales_ufc_100ml: Estadístico=27.990, p-valor=0.000
Se rechaza la hipótesis nula.
Existe una diferencia significativa en el colif_fecales_ufc_100ml entre aguas que estén extremadamente contaminadas y aguas que no lo estén
```

HIPÓTESIS 4: La relación entre la acidez del agua y el oxígeno disuelto en la misma NO influyen directamente en la calidad del agua.

Al igual que en la anterior hipótesis luego de realizar la matriz de correlación observamos una correlación interesante entre el ph del agua y el oxígeno disuelto en la misma, pero a su vez notamos que la correlación de cada columna con el ICA era cercana a 0 lo cual sugiere que si bien puede existir una relación entre el aumento o disminución del ph y el oxígeno disuelto, no necesariamente esto infiere en la calidad del agua.

La hipótesis nula para el test mann whitney plantea que no hay diferencias significativas entre los grupos comparados. Para nuestra hipótesis sería que no hay diferencia significativa en la calidad del agua (índice ICA) cuando varían los valores de oxígeno disuelto en agua y el ph.

Luego tomamos esa hipótesis nula como verdadera para nuestra hipótesis, por lo que la hipótesis nula sería que hay diferencias significativas entre los grupos comparados.

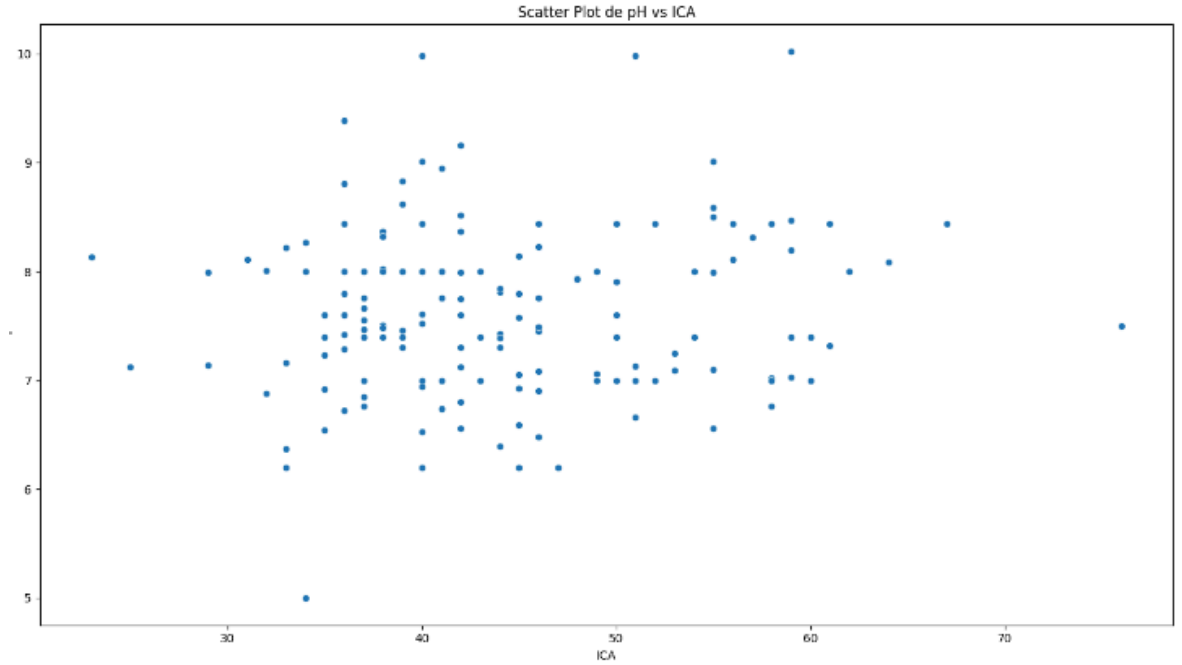
Para nuestra hipótesis sería que no hay diferencia significativa en la calidad del agua (índice ICA) cuando varían los valores de oxígeno disuelto en agua y el ph.

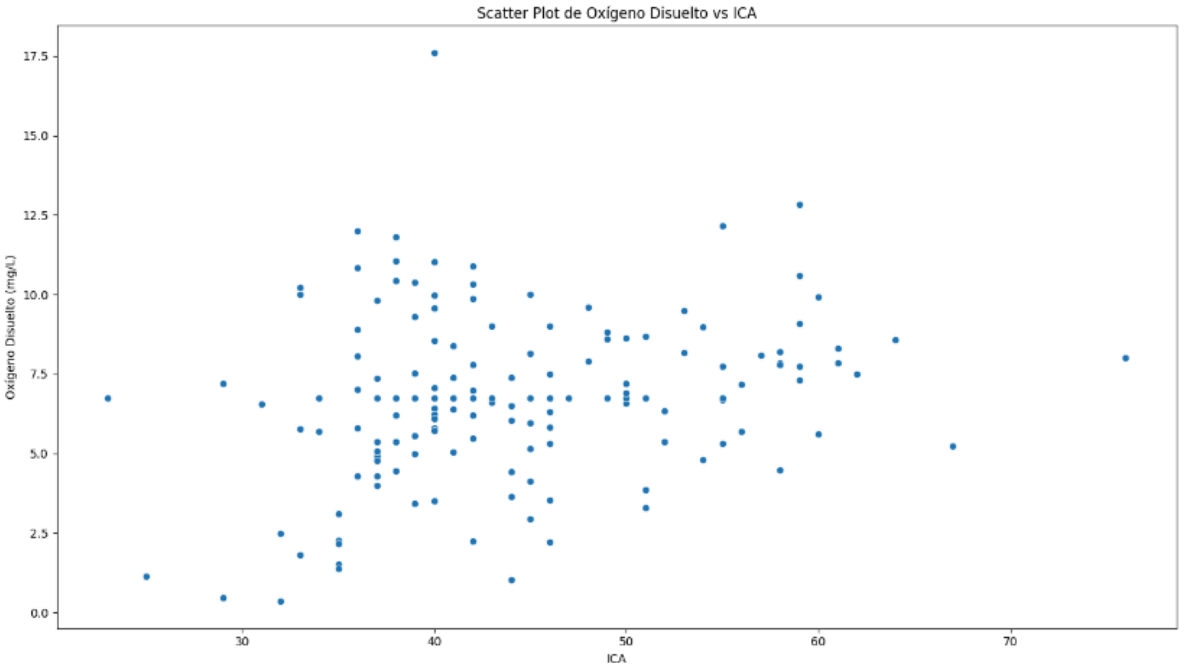
Realizamos un test para ver si la acidez del agua influye en el ICA, y otro distinto para ver si el oxígeno disuelto influye también. Si ambas hipótesis se validan, se valida la hipótesis 4.

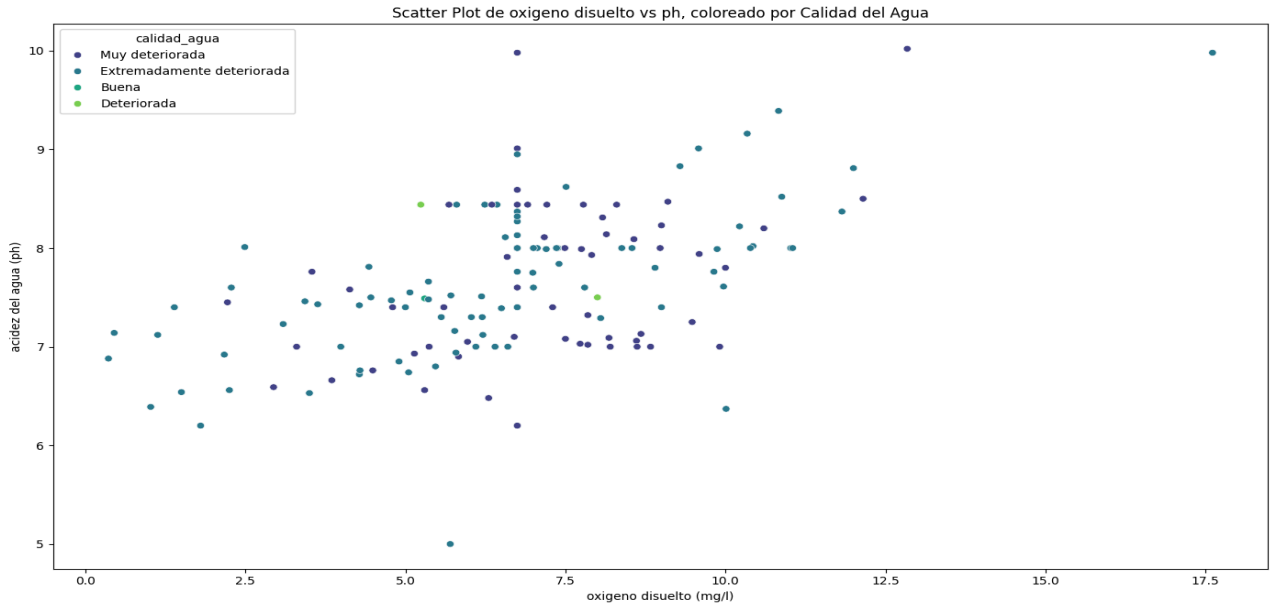
Para hacer válida la hipótesis, el p-valor del test que utilicemos debería ser mayor a 0.05 tanto en el test de la acidez del agua como en el test del oxígeno disuelto. Lo que buscábamos con esto es que la hipótesis nula sea que hay una diferencia significativa entre las aguas extremadamente contaminadas y las demás. Que se rechazara la hipótesis daría

lugar a que la relación entre el ph y el od no inciden en el deterioro del agua (hipótesis que se podría deducir de la matriz de correlación).

Aclarado esto, luego de aplicar el test mann whitney u con dos grupos (aguas extremadamente deterioradas y aguas no extremadamente deterioradas) nos dió un p-valor mayor a 0.05 (que era lo que buscábamos) en ambos casos, por lo que validamos nuestra hipótesis de que la correlación entre la acidez del agua y el oxígeno disuelto en el a misma no influyen directamente en la calidad del agua. Se puede ver en los siguientes gráficos que no hay grandes variaciones en los valores de sus resultados.








```python
# Test de Mann-Whitney U para comparar el amonio en aguas extremadamente deterioradas y no extremadamente deterioradas
stat, p = stats.mannwhitneyu(aguas_ext_contaminadasod, aguas_no_ext_contaminadasod)
print(f"Test de Mann-Whitney U para od: Estadístico={stat:.3f}, p-valor={p:.3f}")

# Interpretación de los resultados
alpha = 0.05  # Nivel de significancia
if p < alpha:
    print("No hay suficiente evidencia para rechazar la hipótesis nula.")
    print("Existe una diferencia significativa en la cantidad de oxigeno disuelto entre aguas extremadamente deterioradas y no extremada
else:
    print("Se rechaza la hipótesis nula.")
    print("No hay una diferencia significativa en la cantidad de oxigeno disuelto entre aguas extremadamente deterioradas y no extremada
```

0.0s Python

```
Test de Mann-Whitney U para od: Estadístico=2365.500, p-valor=0.052
Se rechaza la hipótesis nula.
No hay una diferencia significativa en la cantidad de oxigeno disuelto entre aguas extremadamente deterioradas y no extremadamente deterior
```

```python
# Test de Mann-Whitney U para comparar el ph en aguas extremadamente deterioradas y no extremadamente deterioradas
stat, p = stats.mannwhitneyu(aguas_ext_contaminadasph, aguas_no_ext_contaminadasph)
print(f"Test de Mann-Whitney U para ph: Estadístico={stat:.3f}, p-valor={p:.3f}")

# Interpretación de los resultados
alpha = 0.05  # Nivel de significancia
if p < alpha:
    print("No hay suficiente evidencia para rechazar la hipótesis nula.")
    print("Existe una diferencia significativa en la medida de acidez o alcanilidad entre aguas extremadamente deterioradas y no extrema
else:
    print("Se rechaza la hipótesis nula.")
    print("No hay una diferencia significativa en la medidad de acidez o alcanilidad entre aguas extremadamente deterioradas y no extre
```

0.0s Python

```
Test de Mann-Whitney U para ph: Estadístico=2923.000, p-valor=0.929
Se rechaza la hipótesis nula.
No hay una diferencia significativa en la medidad de acidez o alcanilidad entre aguas extremadamente deterioradas y no extremadamente deter
```

Conclusión

En base a lo que pudimos observar en nuestro análisis, el río se ve afectado por varias causas, pero a pesar de lo pensado previamente hay ciertos elementos que no han ejercido una caída en la calidad de las muestras. Se buscó ver qué factores incidían en la calidad de las muestras. Se vió que la calidad varía nomás si esta es en invierno, con relación a otras relaciones. Por otra parte, se logró validar el hecho de que factores

químicos del agua generan una variación en el comportamiento de la calidad de las muestras, aunque esto no haya sucedido siempre, tal es el caso del oxígeno disuelto y de la acidez del agua. Se buscó ver la incidencia del comportamiento según la zona del muestreo, pero no se vió indicios de que estos tengan una mayor influencia.