Informe trabajo práctico especial

**Integrantes:**

* José María Goicoechea
* Ramiro Maestrojuán

# Objetivo

El objetivo de este trabajo es poder analizar correctamente un dataset, aplicando las herramientas aprendidas durante la cursada para poder buscar relaciones entre las variables del dataset, 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).

El dataset tenía los datos “crudos” (valores repetidos, iguales 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.

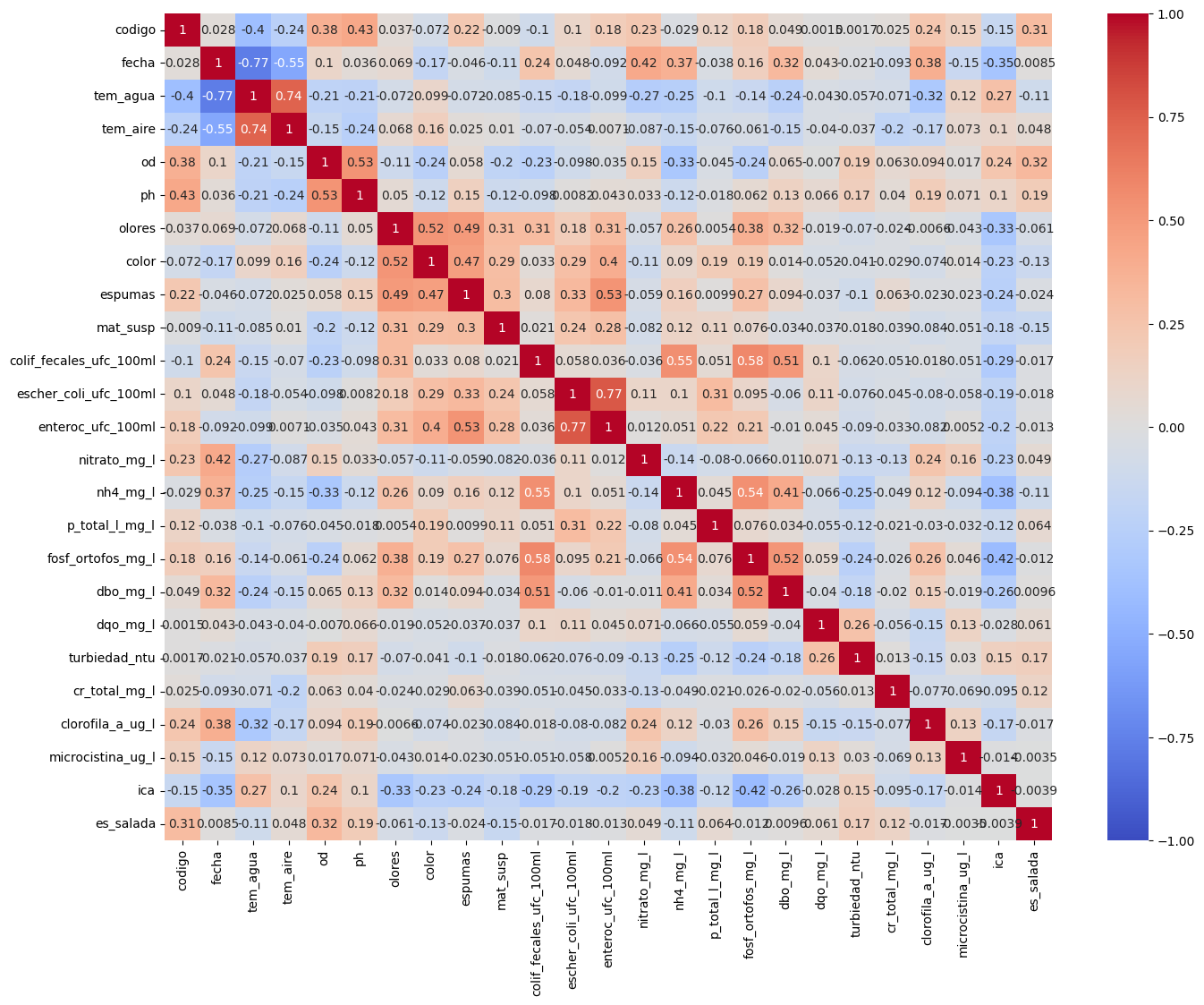
Con el dataset ya limpio comenzamos con el análisis univariado, el cual consiste en analizar cada columna del dataset (cada variable) con diferentes herramientas como gráficos, frecuencia de los datos, valores, promedios e informándonos sobre cada una de esas variables para así poder detectar diferentes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

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.

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 que se decidió agregar una columna para diferenciar las muestras de las zonas de agua salada con las muestras de las zonas de agua dulce y que también se contó con una división de la columna ‘calidad\_de\_de\_agua’ para así poder capaz visualizar más notoriamente ciertos comportamientos según los grupos.



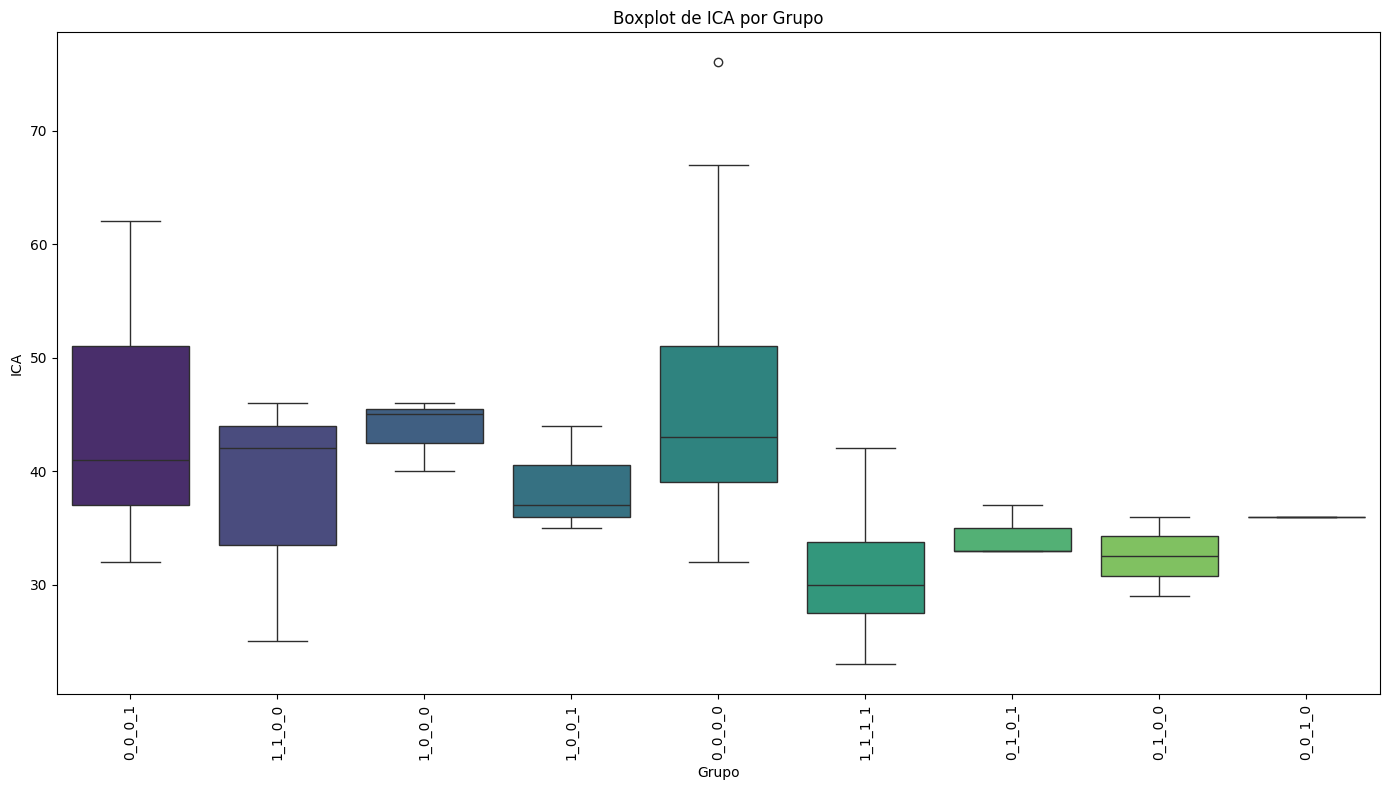
# Hallazgos

Con el análisis mencionado anteriormente pudimos pudimos hacer hallazgos interesantes sobre la calidad del agua en relación sobre que variables en conjunto pueden o no afectarla, estos hallazgos los realizamos mediante el planteo de las siguientes hipótesis:

## Hipótesis 1: La ausencia de olores, color, materia suspendida y de espumas indica un aumento significativo en el índice de calidad de agua.

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.

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 dio un p-valor menor a 0.05 lo cual indica que se rechaza la hipótesis nula, esto quiere decir que nuestra hipótesis se valida.

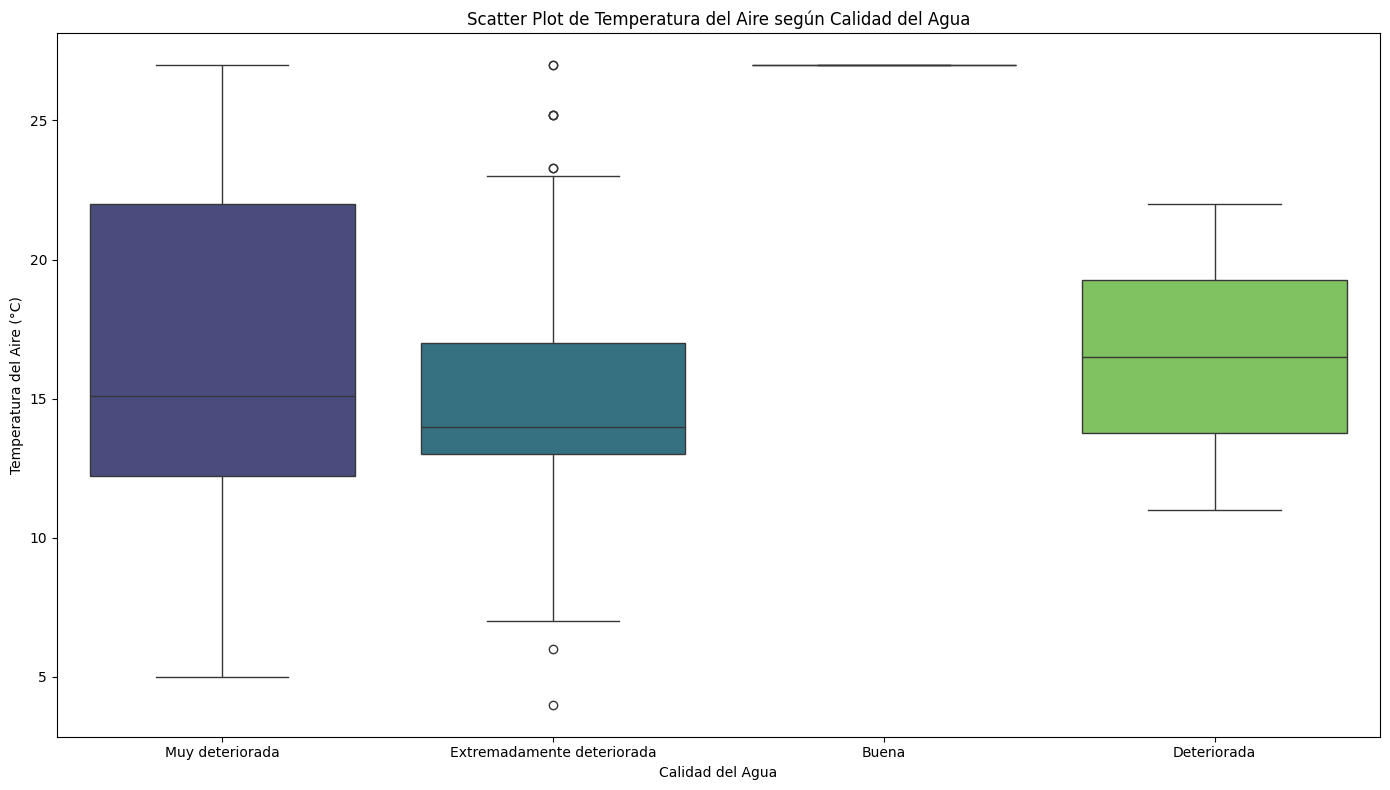
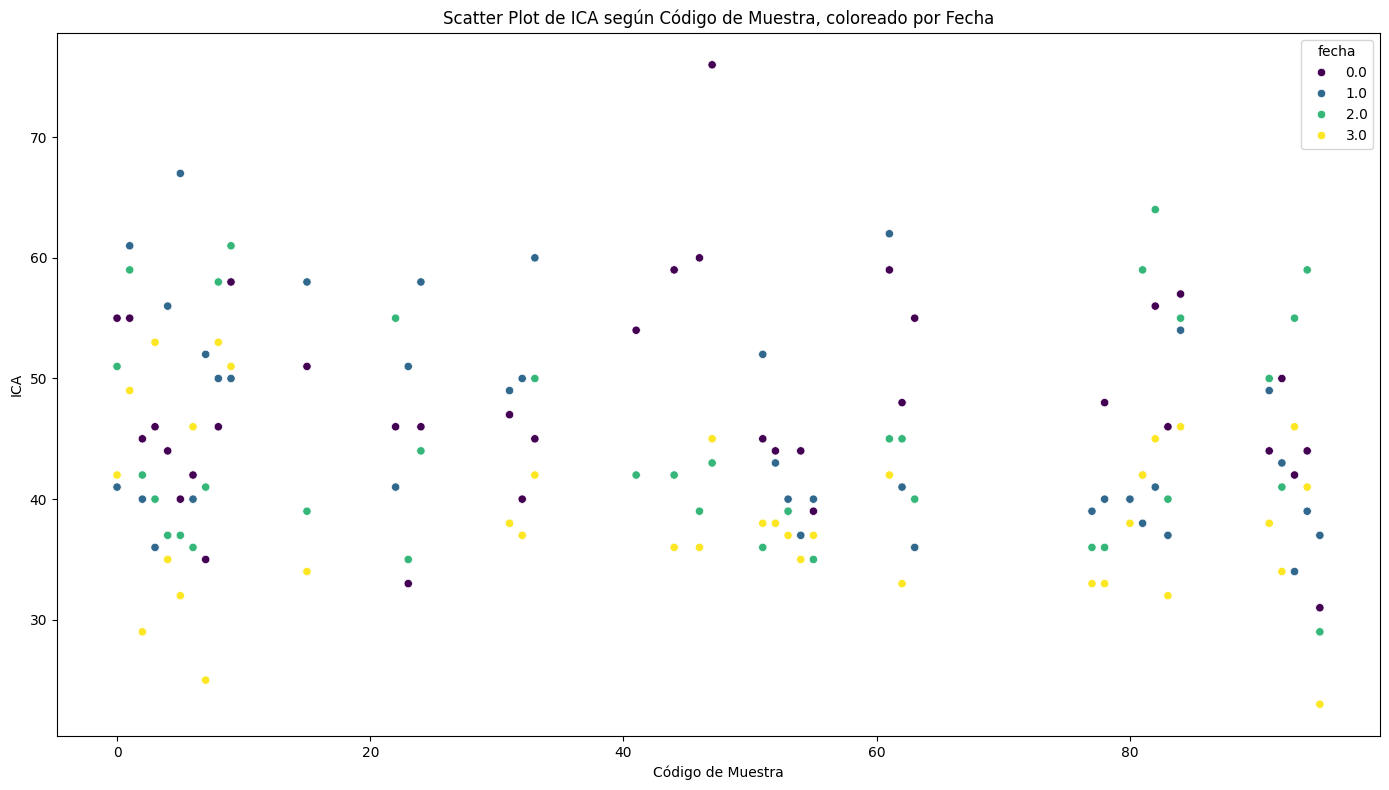
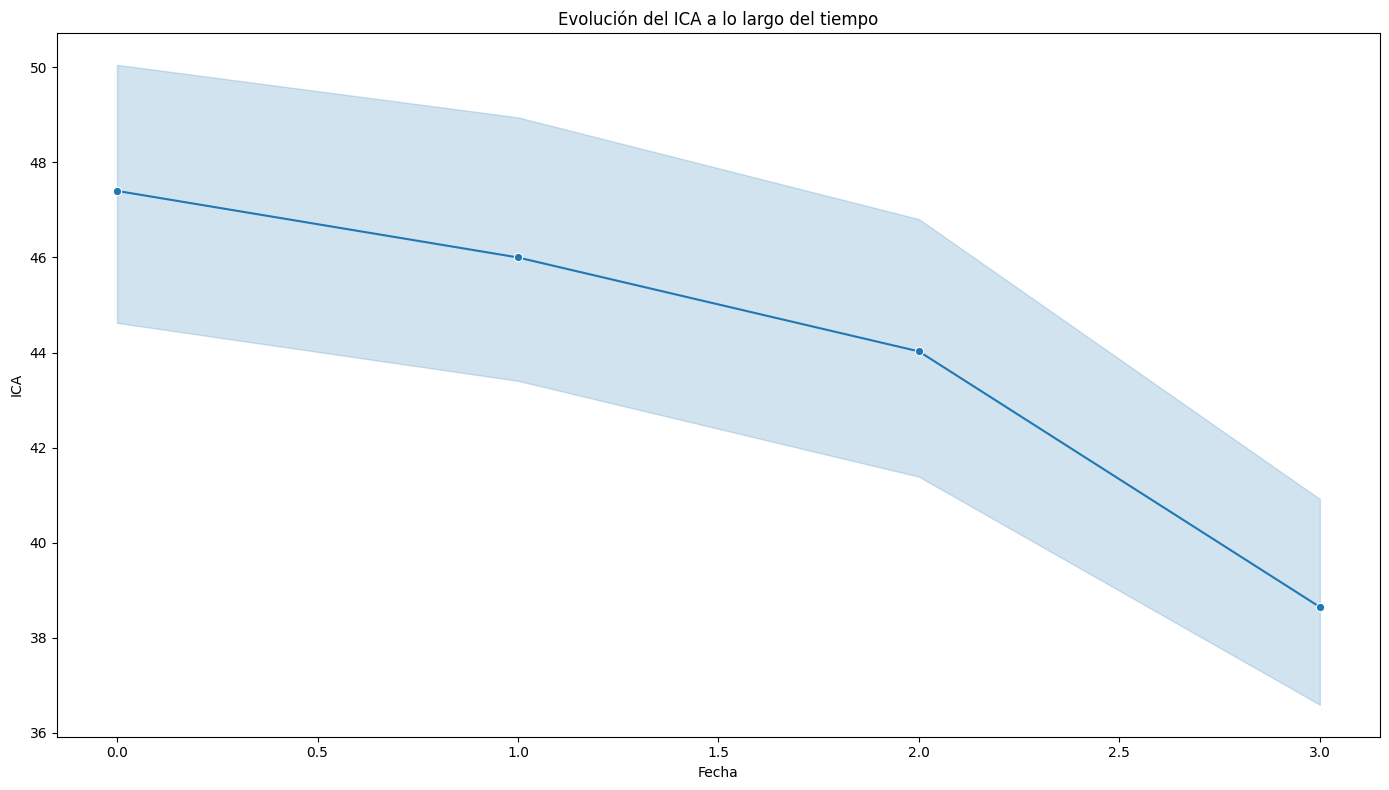


Teniendo en cuenta que se agrupó por las distintas combinaciones de dichos elementos y que un 1 denota que sí están presentes, y un 0 denota que no lo están se puede detallar aquí que el coeficiente aumenta en base a que disminuya la cantidad de los atributos mencionados.

## HIPÓTESIS 2: La calidad del agua empeora notablemente 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. Para validar la hipótesis por el test-t buscamos la normalidad de los datos la cual no encontramos, luego buscamos la homocedasticidad de los datos lo cual pudimos validar, entonces planteamos el test de Mann-Whitney U, 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.

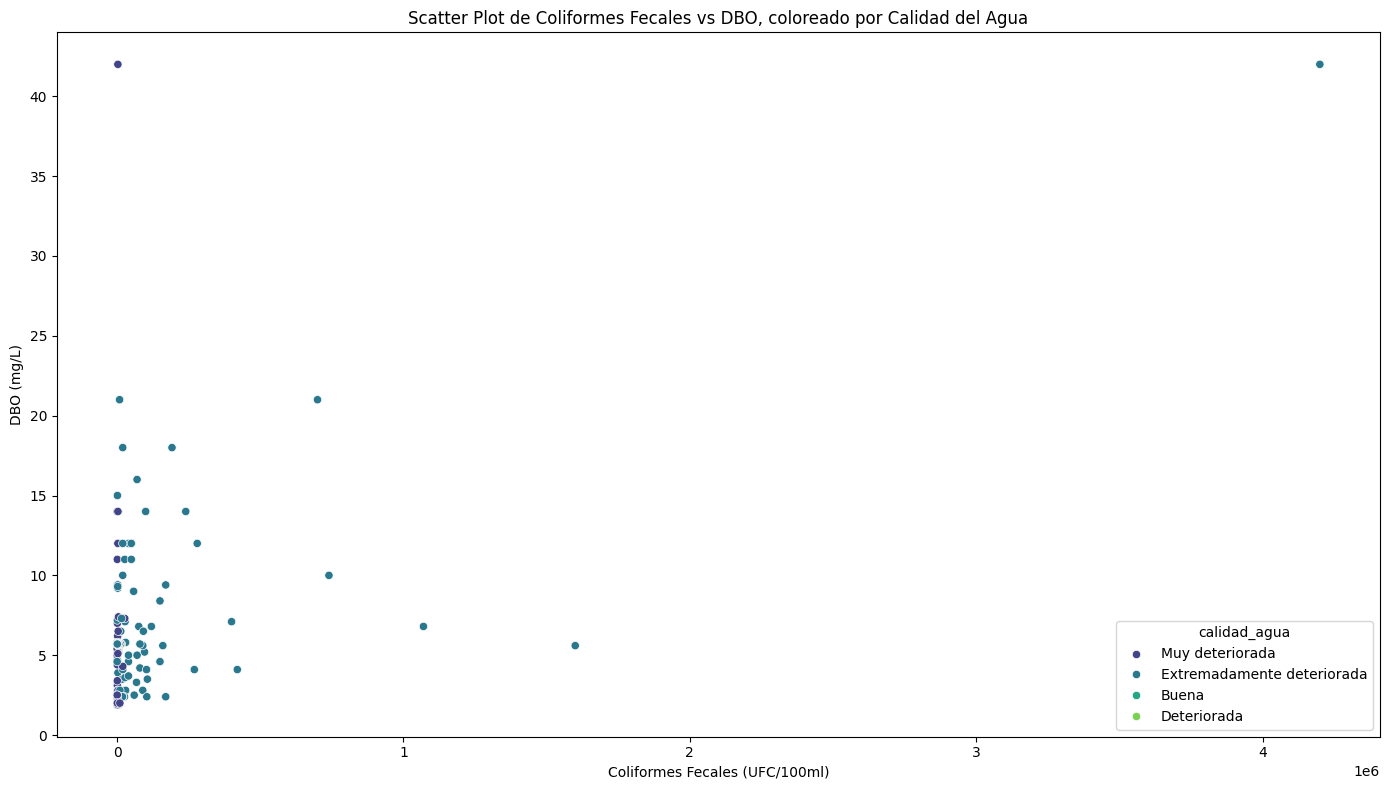
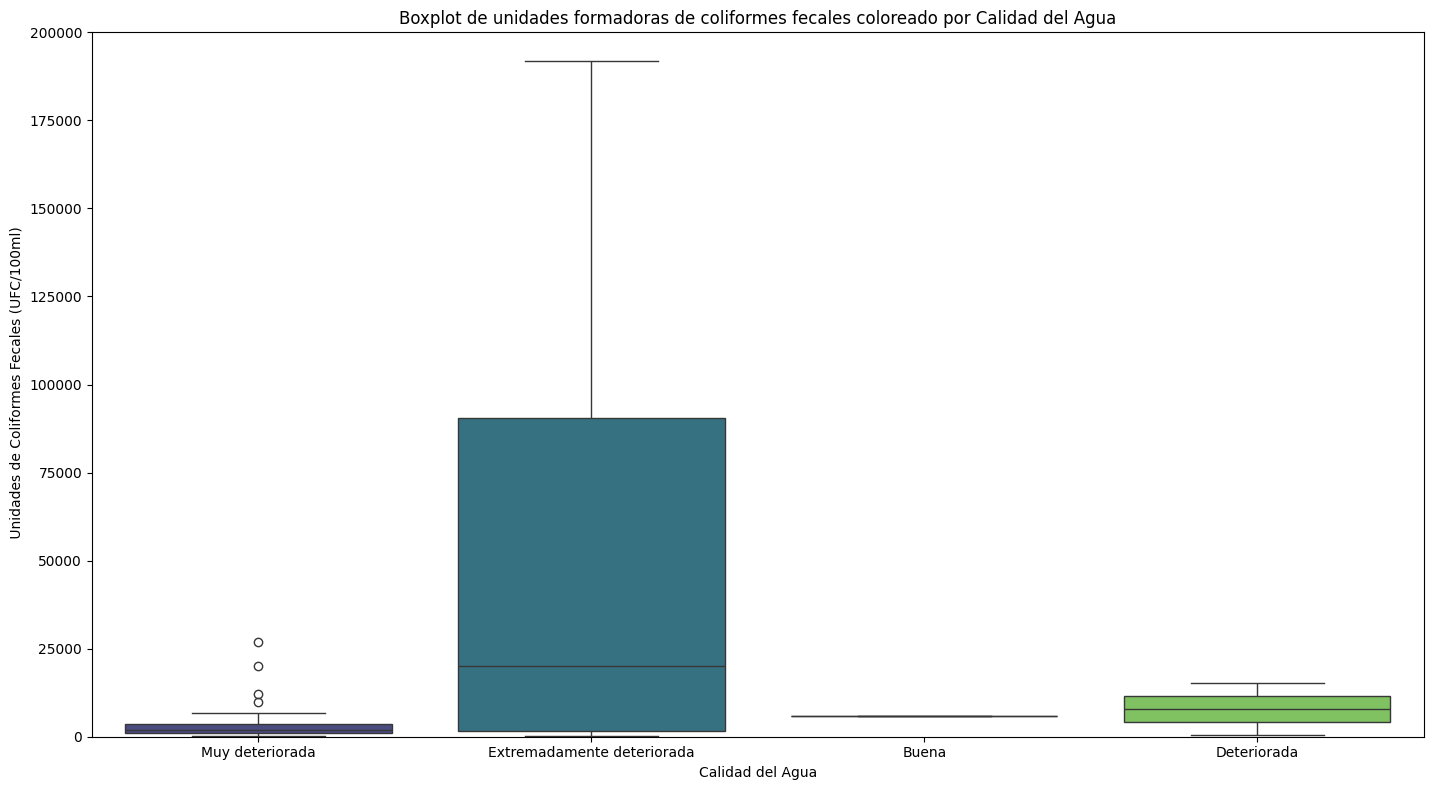


## HIPÓTESIS 3: La calidad del agua se deteriora en función de la presencia de coliformes fecales y la demanda bioquímica de oxígeno.

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.

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

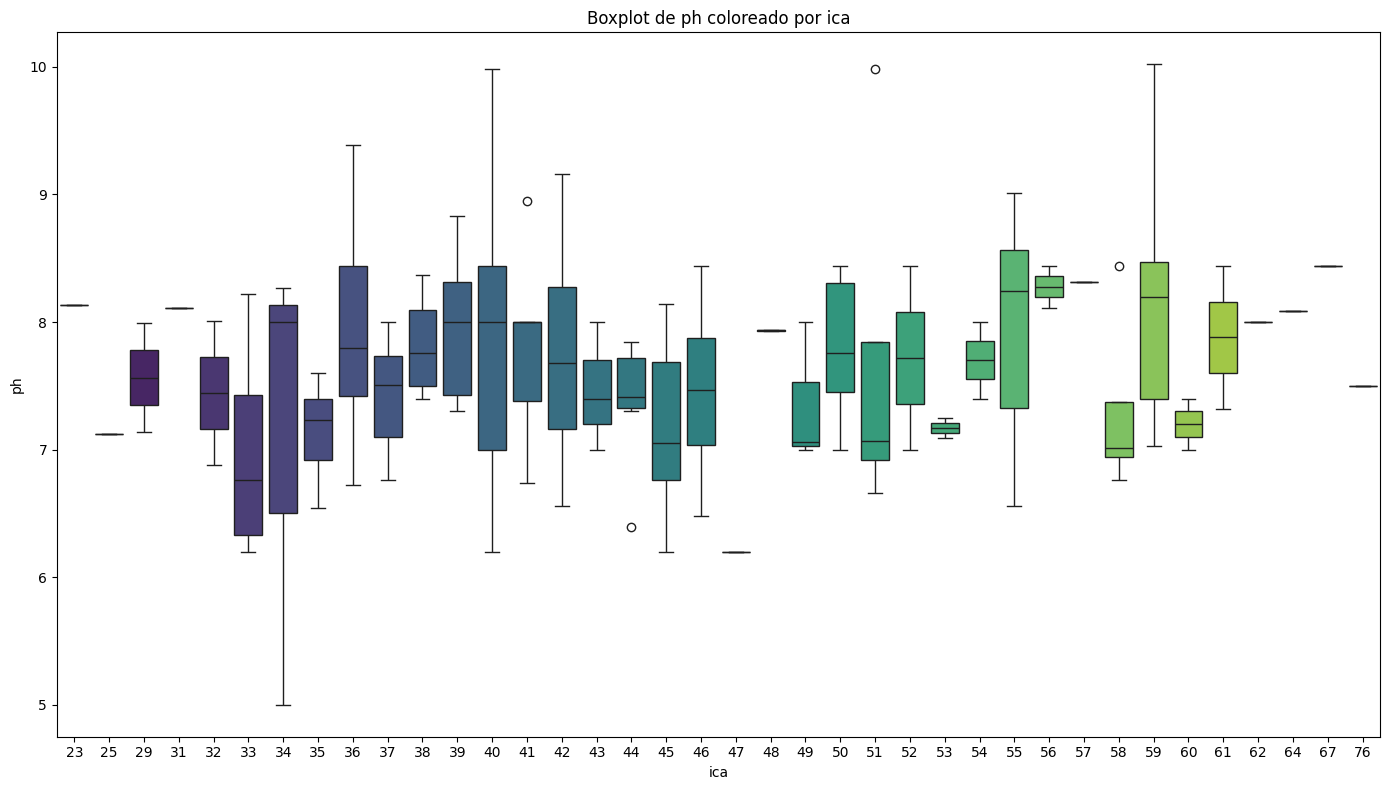
Para la presencia de coliformes fecales no pudimos validar ni la distribución normal de los datos ni la homocedasticidad, por lo que debimos validar la hipótesis por el test de Kruskal. Al validarla nos dio un p-valor a 0.05 por lo que se rechazó la hipótesis nula y validamos nuestra hipótesis de que Existe una diferencia significativa en el colif\_fecales\_ufc\_100ml entre aguas que están extremadamente contaminadas y aguas que no lo estén.

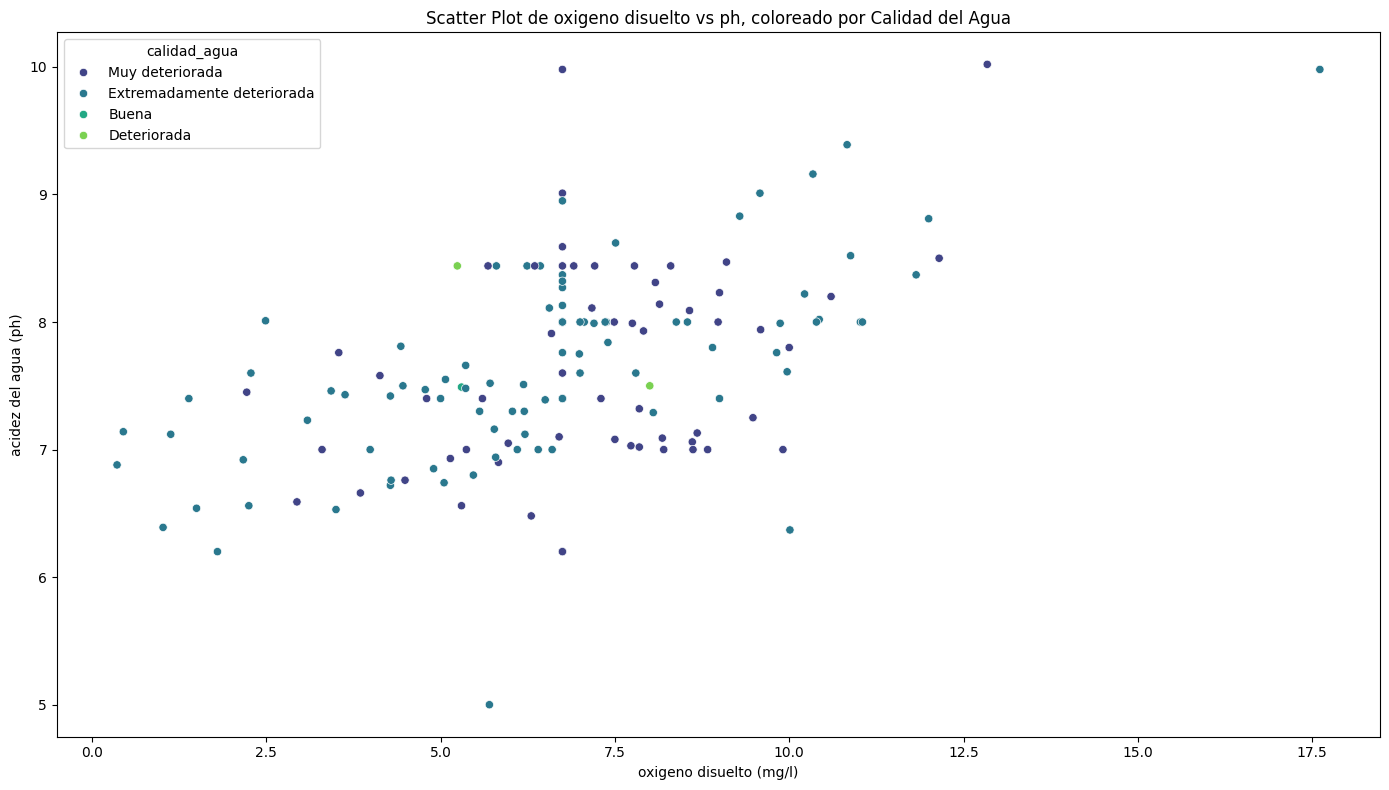
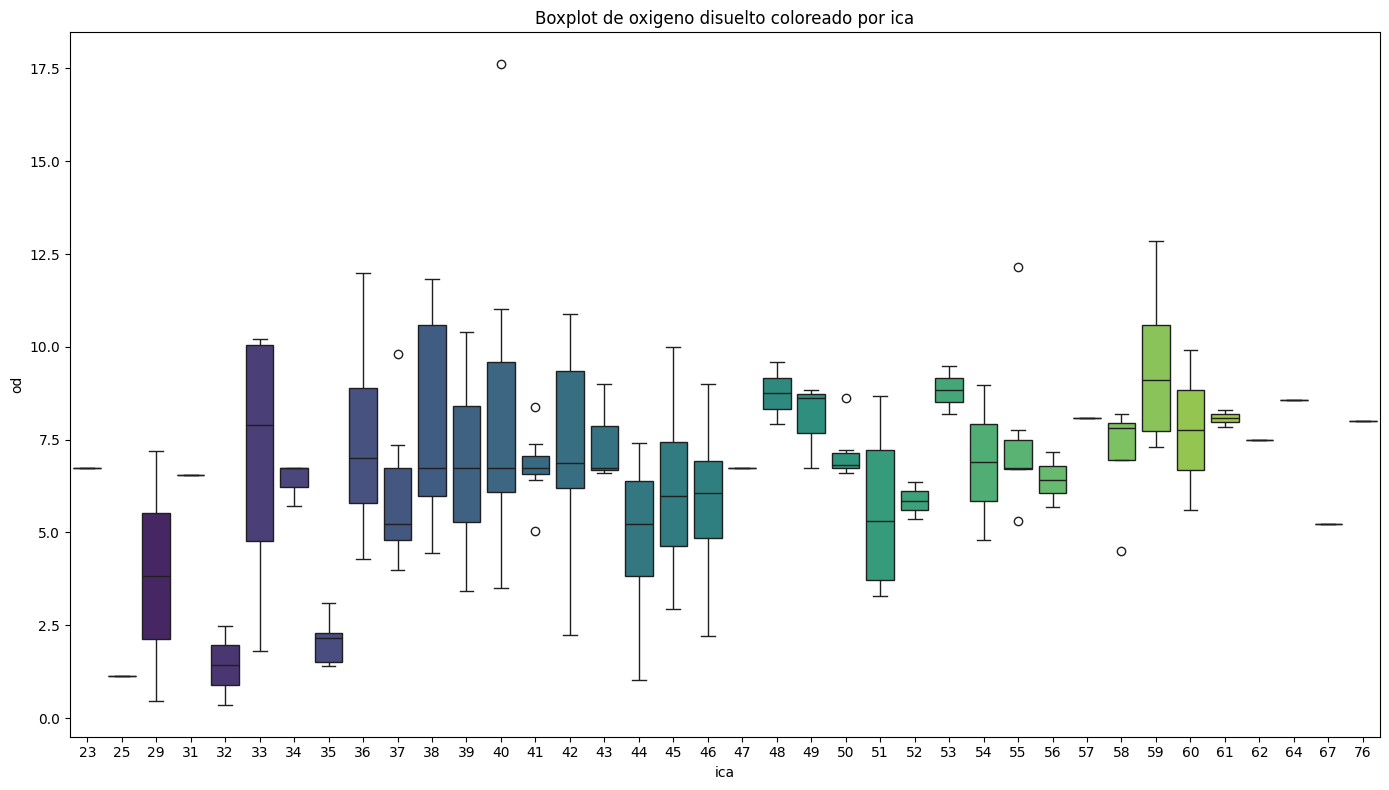


## HIPÓTESIS 4: La relación entre la acidez del agua y el oxígeno disuelto en la misma NO influyen directamente en la calidad del agua.

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.

Decidimos entonces plantear la hipótesis. Para está hipótesis decidimos que para hacer válida la misma, el p-valor del test que utilicemos debería ser mayor a 0.05. Lo que buscábamos con esto es que se acepte la hipótesis nula (nuestra hipótesis) y que la hipótesis que se rechazara fuera la de que existe una relación entre el ph, el od y el deterioro del agua (hipótesis que se podría deducir de la matriz de correlación).

Aclarado esto, luego de aplicar el test de Mann-Whitney U nos dio un p-valor mayor a 0.05 (que era lo que buscábamos), 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.



# 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.