

Fundamentos de la Ciencia de Datos

Práctico Especial

Alumnos:

Agustín Sequeira (osequeira@alumnos.exa.unicen.edu.ar)

Axel Kjolhede (akjolhede@alumnos.exa.unicen.edu.ar)

Tomás Perco (lperco@alumnos.exa.unicen.edu.ar)

Índice

Resumen.....	2
Introducción.....	2
Descripción de los datos:.....	3
Desarrollo experimental.....	5
EDA.....	6
Entendimiento de las variables.....	6
Limpieza del dataset.....	7
Exploración inicial.....	8
Presentación de las hipótesis:.....	14
H1.....	15
H2.....	17
H3.....	20
H4.....	21
Conclusiones.....	23

Resumen

Este informe presenta la exploración de un conjunto de datos sobre las condiciones del agua del río de la plata, en donde se buscó interpretar qué información nos daban las observaciones registradas. Mediante test adecuados, y con la ayuda de herramientas de visualización, se pusieron a prueba hipótesis ideadas durante el transcurso de la investigación.

Introducción

El Río de la Plata es un recurso natural de gran valor para la región, desempeñando un papel crucial en el ecosistema y en las actividades humanas. Sin embargo, la calidad del agua en esta área ha sido objeto de preocupación debido a la creciente contaminación.


Este informe presenta un análisis exploratorio de un conjunto de datos sobre las condiciones del agua del Río de la Plata. El objetivo principal es interpretar la información obtenida de las observaciones registradas y evaluar la influencia de distintas variables en las condiciones del agua. Para lograr esto, se llevarán a cabo varias hipótesis que se pondrán a prueba mediante técnicas estadísticas adecuadas.

Se emplearán herramientas de visualización de datos para identificar patrones y tendencias en los datos, y se realizarán pruebas de hipótesis para validar las suposiciones hechas durante la investigación. Este enfoque permite una comprensión más profunda de cómo las distintas variables impactan en las condiciones del agua, lo cual es esencial para el desarrollo de estrategias de gestión.

Descripción de los datos:

Las variables que se encuentran presentes en nuestro dataset, según la documentación entregada, son las siguientes:

1. sitios: Localización específica donde se realizó el muestreo del agua.
2. código: Identificador único para cada muestra o estación de muestreo.
3. fecha: Fecha en la que se tomó la muestra de agua.
4. año: Año en que se realizó el muestreo.
5. campaña: Nombre o número de la campaña de monitoreo en la que se realizó el muestreo.
6. tem agua: Temperatura del agua en grados Celsius.
7. tem aire: Temperatura del aire en grados Celsius.
8. od: Oxígeno disuelto, medido en miligramos por litro (mg/L), esencial para la vida acuática.
9. ph: Medida de la acidez o alcalinidad del agua, en una escala de 0 a 14.
10. olores: Presencia de olores en el agua, que puede indicar contaminación.
11. color: Color del agua, que puede ser un indicador de la calidad del agua.
12. espumas: Presencia de espumas en la superficie del agua, que puede ser un signo de contaminación.
13. mat susp: Materia suspendida, que se refiere a partículas sólidas que flotan en el agua.
14. colif fecales ufc 100ml: Unidades formadoras de colonias de coliformes fecales en 100 ml de agua, un indicador de contaminación fecal.
15. escher coli ufc 100ml: Unidades formadoras de colonias de Escherichia coli en 100 ml de agua, otro indicador de contaminación fecal.
16. enteroc ufc 100ml: Unidades formadoras de colonias de enterococos en 100 ml de agua, que también indican contaminación fecal.
17. nitrato mg l: Concentración de nitratos en miligramos por litro (mg/L), que puede indicar contaminación por fertilizantes.
18. nh4 mg l: Concentración de amonio en miligramos por litro (mg/L), que puede ser un indicador de contaminación orgánica.

- 
19. p total l mg l: Fósforo total en miligramos por litro (mg/L), que incluye todas las formas de fósforo en el agua.
 20. fosf ortofos mg l: Concentración de ortofosfatos en miligramos por litro (mg/L), que es un nutriente importante.
 21. 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.
 22. dco 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.
 23. turbiedad ntu: Turbidez del agua medida en unidades NTU (Nephelometric Turbidity Units), que indica la claridad del agua.
 24. hidr deriv petr ug l: Hidrocarburos derivados del petróleo en microgramos por litro ($\mu\text{g/L}$), que indican contaminación por productos petroleros.
 25. cr total mg l: Concentración total de cromo en miligramos por litro (mg/L), un metal pesado que puede ser tóxico.
 26. cd total mg l: Concentración total de cadmio en miligramos por litro (mg/L), otro metal pesado que es tóxico en altas concentraciones.
 27. clorofila a ug l: Concentración de clorofila a en microgramos por litro ($\mu\text{g/L}$), que indica la cantidad de fitoplancton en el agua.
 28. microcistina ug l: Concentración de microcistinas en microgramos por litro ($\mu\text{g/L}$), que son toxinas producidas por ciertas algas.
 29. ica: Índice de calidad del agua, que puede ser un valor calculado para evaluar la calidad general del agua.
 30. calidad de agua: Clasificación general de la calidad del agua basada en los parámetros medidos.

Desarrollo

Cómo se mostró en la sección anterior, para llevar a cabo el análisis contamos con un conjunto de datos sobre distintas mediciones realizadas en varios puntos del Río de la Plata, con observaciones tomadas en cada estación del año para cada punto. En estas mediciones se obtienen datos como la cantidad de distintos químicos y bacterias, la presencia de espuma, materia suspendida, entre otras.

El primer paso para el desarrollo del análisis fue el EDA (Análisis exploratorio de datos) para comprender con qué estábamos tratando en cada variable, posteriormente pasando a la etapa de “limpieza” de los datos (reemplazo por nulos), en la cual se consumió la mayoría del tiempo dedicado debido a la gran cantidad de inconsistencia presentes en el dataset.

Una vez que finalizamos el paso anterior comenzamos a formular diversas hipótesis, basadas en nuestro conocimiento previo de cómo creíamos que se podrían afectar las variables entre sí, y en lo que aprendimos durante la etapa de EDA.

Para finalizar, elegimos aquellas hipótesis que consideramos más relevantes y de mayor apego a nuestra premisa inicial (Entender un poco mejor qué afecta a las condiciones del agua). Comenzamos a comprobarlas haciendo uso de distintas herramientas para poder visualizar las potenciales relaciones entre las variables de estudio y determinar las condiciones de cada subconjunto analizado para, en base a ello, llevar a cabo el test de hipótesis más adecuado.

EDA

Entendimiento de las variables

Para lograr entender más el conjunto de datos con el que estábamos tratando lo primero que hicimos fue informarnos que tipo de datos estaba almacenado en cada variable. En el caso de este dataset, vimos que la gran mayoría no poseía un tipo que se correspondiera mucho a su significado real, por lo que anticipamos que esto requeriría cambios.


A su vez, algunas de las columnas parecían ser redundantes o poco relevantes. Por ejemplo, la variable "Año" no aportaba ninguna información adicional al ya existir la columna "Fecha", y la columna "orden" no sólo no estaba presente en la documentación entregada, sino que tampoco era útil para analizar las condiciones del agua.

De modo similar, el campo que medía los hidrocarburos derivados del petróleo poseía solo un valor a lo largo de todas las muestras, por lo que también era prescindible. También, el atributo que medía los niveles de cadmio en el agua parecía estar distribuido en únicamente dos rangos, lo cuál también conllevaría realizar ajustes sobre la variable.

Limpieza del dataset

Para limpiar el conjunto de datos, lo primero que hicimos fue crear una copia del mismo en la cuál haríamos todos los cambios e ir revisando columna por columna en busca de valores nulos o sin sentido. Por ejemplo, valores como "No se midió" ó "Faltó un frasco" fueron reemplazados por nulos para poder ser tratados posteriormente.

Investigando sobre la fuente de este dataset, descubrimos que para algunas de nuestras variables existían valores que quienes tomaron los datos consideraban "aceptables", Esto nos resultó de utilidad para lograr identificar outliers.



Las columnas que no aportaban información útil (hidrocarburos, orden y año) fueron eliminadas. Además, las filas que tenían muchos nulos (más de 10), fueron removidas por no ser consideradas confiables. Respecto a la variable de la concentración de cadmio en el agua, al investigar los valores de referencia aportados por Res. ACUMAR 46/2017¹, podemos asegurar que los valores medidos se podrían dividir en no preocupantes y preocupantes. Teniendo todo esto en cuenta consideramos apropiado cambiar esta columna por una que tenga como nombre "Alta concentración de Cadmio", de tipo booleana, reflejando los dos grupos representados por las mediciones.

Para reemplazar los valores representados por rangos en columnas que debían tener entradas puntuales, tomamos el criterio de utilizar el mismo valor del rango (por ejemplo, un dato con valor "<2.0" pasaría a ser "2.0").

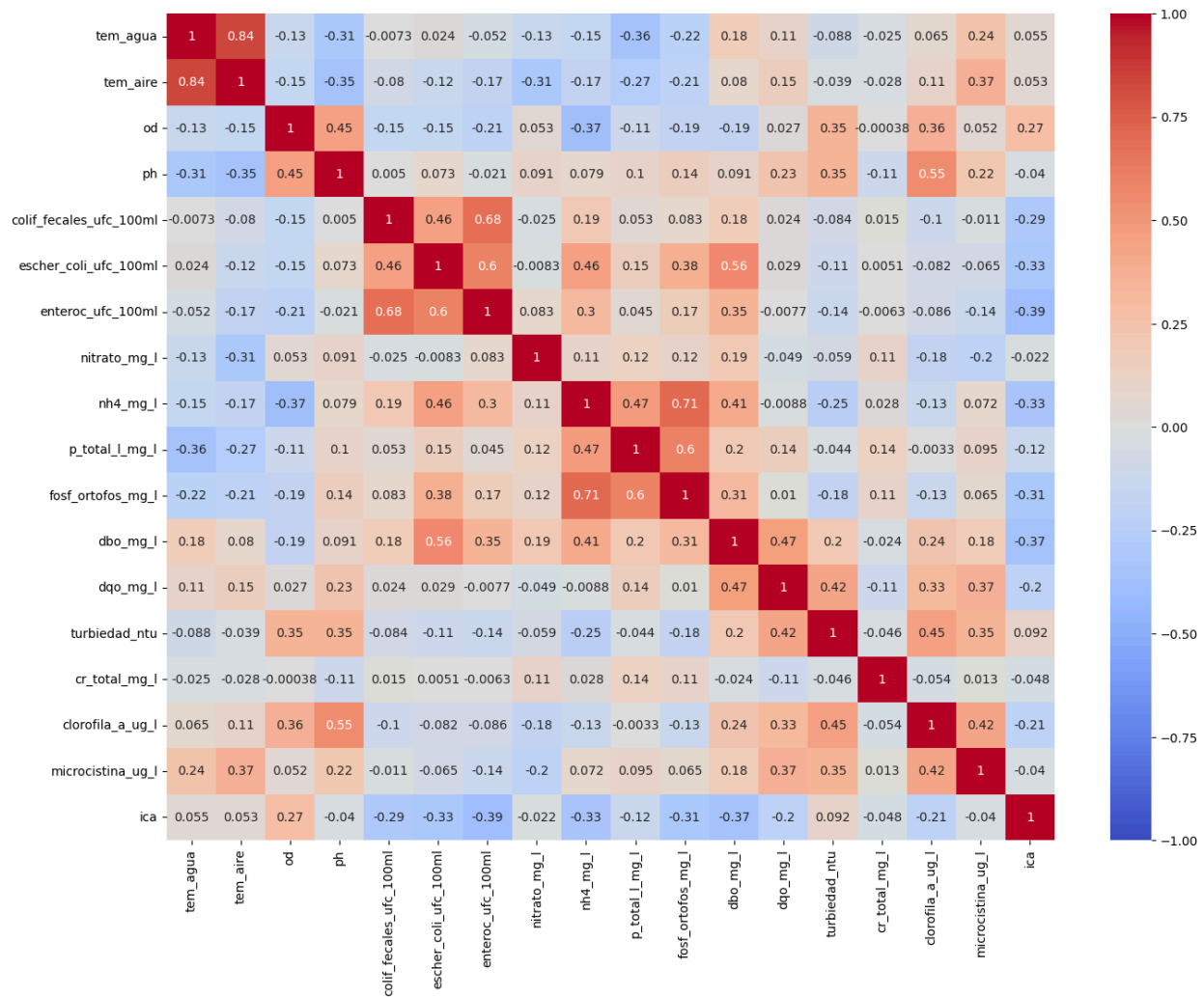
Como se anticipaba en el análisis inicial de las variables, se ajustaron los tipos de las mismas para que sean más representativas; las que medían valores continuos pasaron a ser numéricas, las que parecían ser dicotómicas pasaron a ser booleanas, etc.

Por último, al momento de tratar los datos nulos, llegamos a la conclusión de que no nos podíamos permitir eliminar toda tupla con al menos uno de estos valores ya que perderíamos una gran parte del dataset. Por lo tanto, para las variables continuas decidimos reemplazarlos con la media, siendo la única excepción el ICA, ya que en este caso podíamos utilizar la media agrupada por sitio en lugar de la general dado que la cantidad de valores faltantes en la columna era menor y no se daba el caso de no tener ningún dato válido para un sitio en particular.

¹ Monitoreo de la calidad del agua en el Río de la Plata. Campaña Otoño 2023
(https://ciam.ambiente.gob.ar/images/uploaded/datasets/368/Metadata_Monitoreo_calidad_Rio_de_la_Plata_oto%C3%B1o_2023.pdf)

Exploración inicial

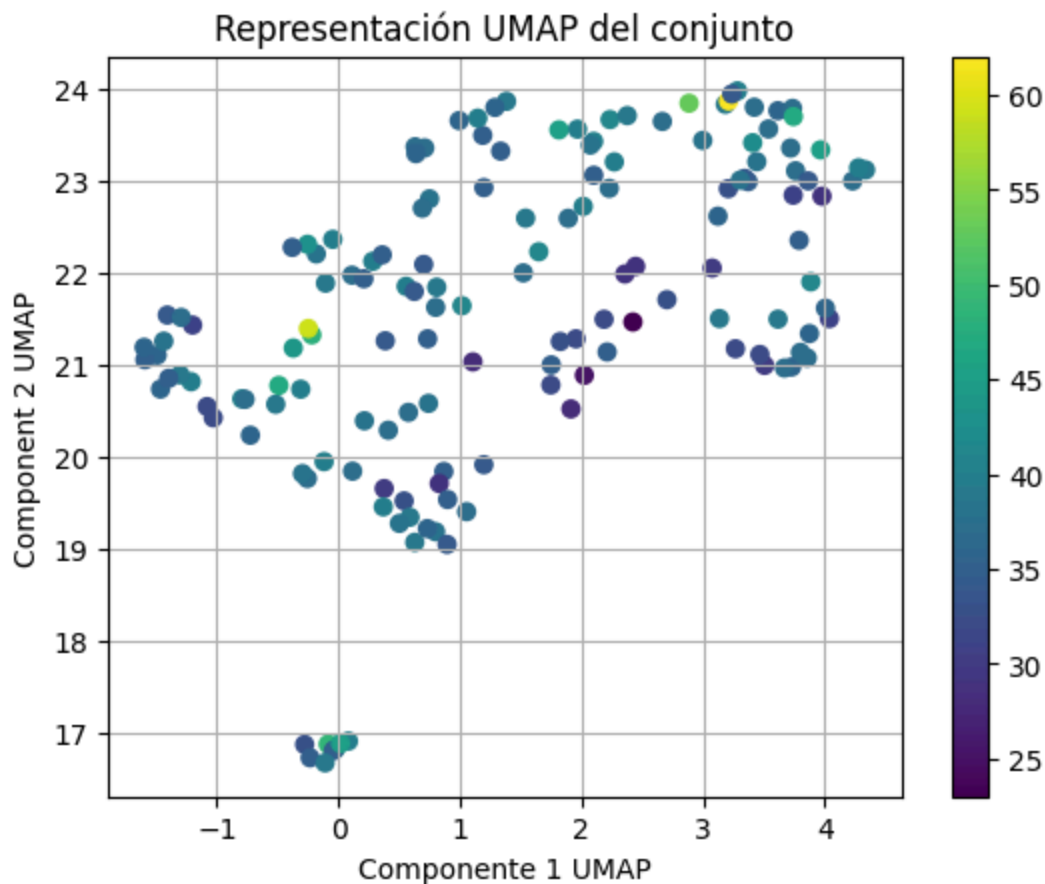
Desde un principio nuestro objetivo fue comprender las condiciones del agua, por lo que lo primero que decidimos fue comprobar si alguna de las variables tenía una fuerte correlación con el índice de calidad de agua (ICA) que entendimos como la más cercana a nuestro objetivo de estudio.



[Gráfico 1: Matriz de correlación entre las variables cuantitativas del conjunto de datos.]

Pero, como observamos que no hay ninguna que destaque de la manera que buscamos (Como se ve en el gráfico, no hay una correlación alta entre ICA y otra variable), consideramos que quizá esto se deba a que todas en conjunto influyen sobre el valor de ICA asociado a cada muestra.

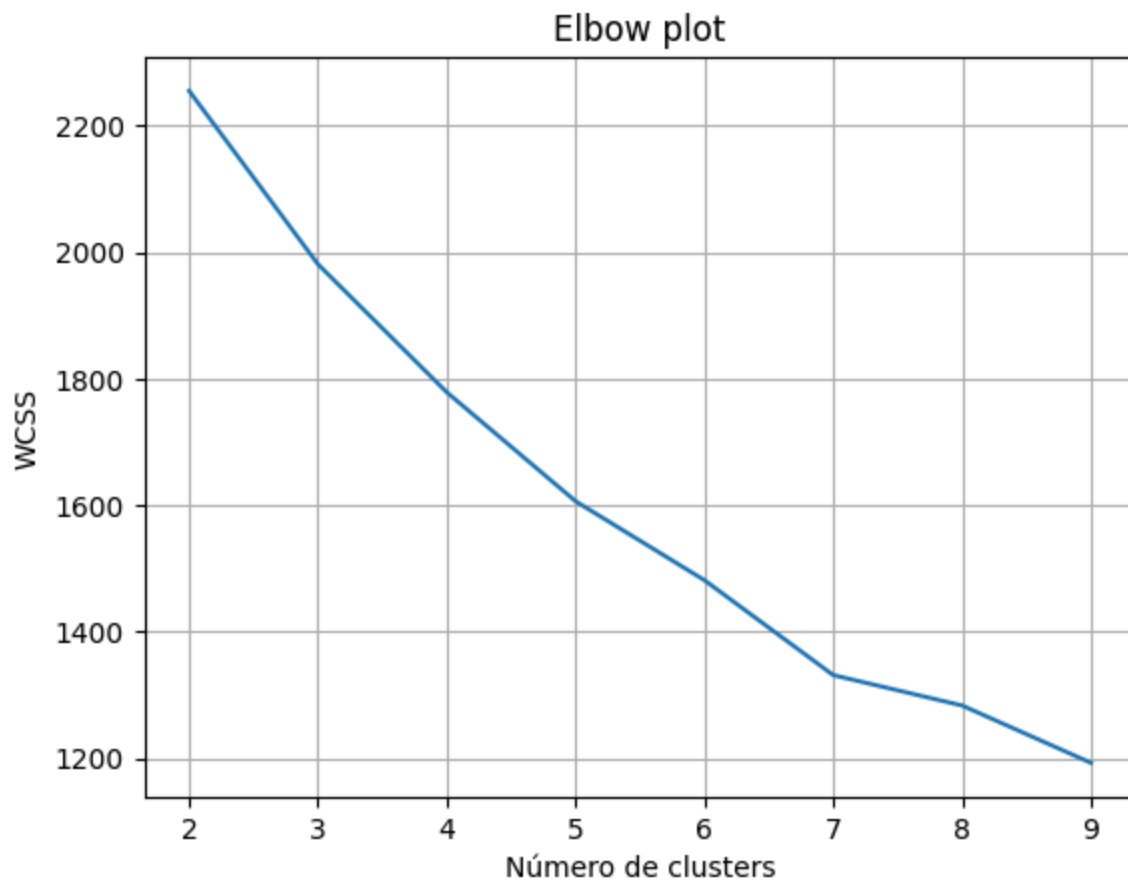
Ya desde este punto pudimos ver que no se hacían evidentes agrupaciones, y el hecho de que la varianza explicada sea solo del 35%, hizo que buscáramos grupos con otro método de reducción que, a diferencia de este, no sea lineal. Por ello replicamos el procedimiento anterior pero en su lugar utilizamos el método UMAP, el cuál nos entregó el siguiente agrupamiento:



[Gráfico 3: Agrupamiento utilizando UMAP y coloreando con los valores de ICA correspondientes a cada observación]

Esta nueva gráfica nos mostró una distribución más “amigable”, sin embargo a la hora de colorear las observaciones con los valores de ICA no pudimos encontrar grupos distinguibles. Es por esto que como un intento final decidimos que en lugar de colorear de esta manera podríamos dejar que un algoritmo de clustering agrupe las muestras y nosotros revisar qué características distinguen a cada grupo en caso de existir.

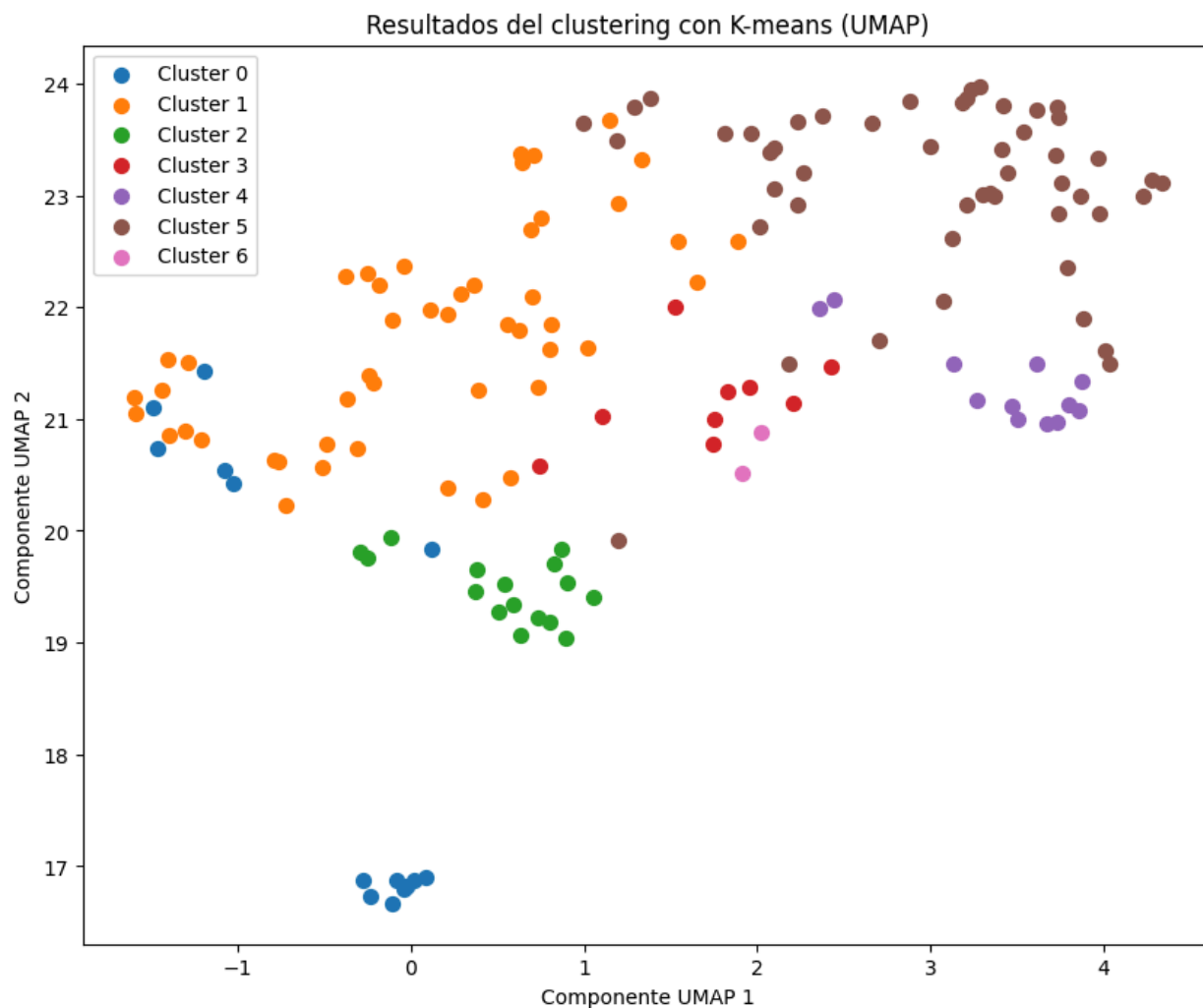
Decidimos utilizar el algoritmo de agrupamiento K-means pero como a este se le debe indicar cuántos grupos queremos que complete lo primero que hicimos fue un “Elbow-plot”, gráfica que nos puede llegar a orientar sobre cuál es la cantidad de grupos (clusters) más adecuada.



[Gráfico 4: Elbow Plot mostrando la dispersión interna ($WCSS^2$) para distintas cantidades de clusters en el análisis K-means]

Como observamos en la gráfica, la variación de la dispersión interna decrece a partir de 7 clusters por lo que es este número el que usamos al realizar K-Means. Una vez que este último generó los grupos consideramos adecuado asignar un color a cada uno y pintarlos sobre el gráfico que nos generó UMAP.

² WCSS : mide la suma de las distancias al cuadrado entre los puntos de datos dentro de un mismo clúster y su centroide, cuanto menor es el WCSS, más cercanos están los puntos de datos a sus centroides



[Gráfico 5: Agrupamiento utilizando UMAP coloreado de acuerdo a cada cluster encontrado por K-Means]

Aunque esta visualización era más cercana a lo que esperábamos ver, seguía teniendo problemas, como lo son el hecho de que algunos grupos estén mezclados o la gran disparidad en el tamaño de los mismos. Es por esto que decidimos modificar el rumbo de nuestro análisis, y tratamos de enfocarnos en preguntas/hipótesis más puntuales que surgieron durante la exploración del dataset.

Presentación de las hipótesis:

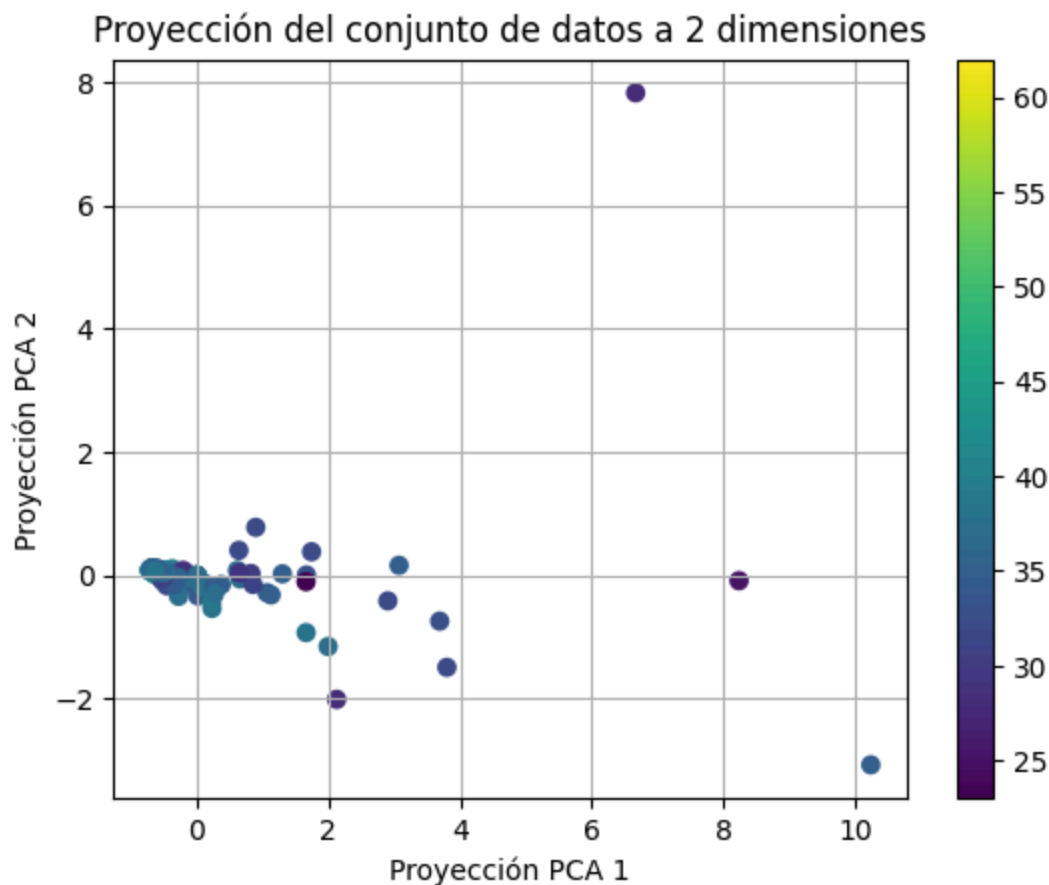
- H1: "Las variables relacionadas a la materia fecal repercuten negativamente sobre la calidad del agua". La presencia de altos niveles de coliformes fecales y otros indicadores de materia fecal tiende a asociarse con clasificaciones de calidad de agua más bajas. Queremos comprobar si este es el caso ya que servirá para señalar la necesidad de controles sanitarios en fuentes y sistemas de tratamiento.
- H2: "La presencia de cadmio influye en la calidad del agua."
Comprobar que los sitios con alta presencia de cadmio tienen una de calidad de agua más baja en el índice ICA podría revelar que hay zonas críticas con riesgo de contaminación por metales pesados en las cuales el agua se ha deteriorado en consecuencia a esto.
- H3: "La época del año influye en la presencia de materia suspendida. (Primavera - Verano) y (Otoño - Invierno)":
Es posible que, en las épocas del año en las que predomina el calor (primavera y verano), haya una diferencia significativa en la presencia de materia suspendida respecto al resto de estaciones. De determinarse que esto es cierto, podría servir para hacer notar la importancia de la implementación de un protocolo de limpieza más intensivo o de concientizar a la población durante estas épocas del año.
- H4: "El valor de oxígeno disuelto es menor cuando hay presencia de materia suspendida".
A mayor densidad de materia suspendida, el oxígeno disuelto podría ser menor, indicando que el aumento de sólidos suspendidos reduce la capacidad del agua para intercambiar gases con la atmósfera, lo cual puede ser perjudicial para la vida acuática.
- H5: "Relación entre temperatura del agua y niveles de oxígeno disuelto (OD)":
Las altas temperaturas del agua están asociadas con niveles más bajos de oxígeno disuelto. Confirmar esta relación en el Río de la Plata permitiría evidenciar cómo el calentamiento del agua impacta negativamente en su capacidad para sostener la vida acuática.

- H6: “Los niveles de contaminación fecal se ve afectado por las distintas estaciones del año”:

Dado que en la hipótesis anterior buscamos comprobar si la contaminación fecal afecta la calidad del agua (H1), dependiendo de su conclusión podríamos obtener información útil al ver si las estaciones del año también juegan un rol sobre la magnitud del problema, lo que ayudaría a poder enfocar de una manera aún mejor los recursos para la solución.

H1

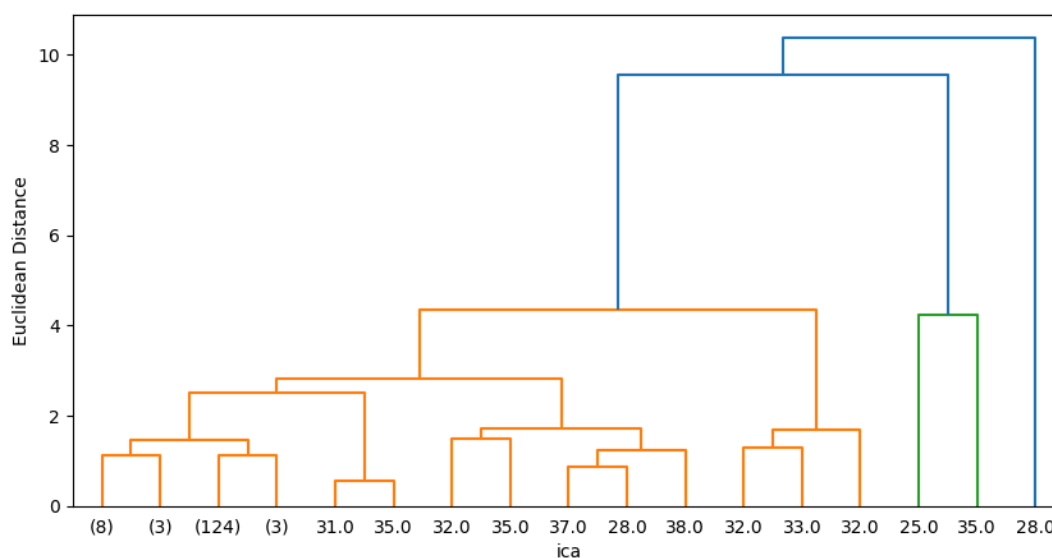
Comenzamos planteando un agrupamiento entre las variables de bacterias en materia fecal (colonias de coliformes fecales, colonias de *Escherichia coli* y colonias de enterococos), para luego hacer un análisis en conjunto y ver que pasaba entre ellas utilizando reducción de dimensionalidad con PCA.



[Gráfico 6: Agrupamiento utilizando PCA sobre bacterias fecales y coloreado de acuerdo a valores de ICA]

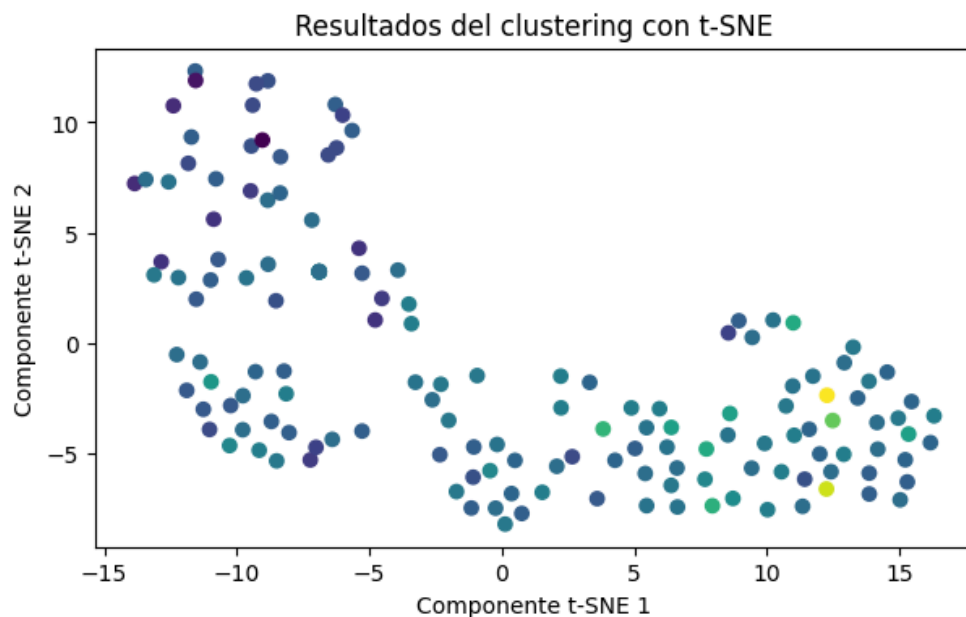
Con dicha gráfica (Gráfico 6) observamos cómo estas variables afectan al ICA intentando encontrar una separación distinguible mediante el coloreo.

También se buscó otra forma de visualización haciendo clustering jerárquico, en donde se ve claramente lo marcados que están los 3 grupos (ver gráfico PCA)



[Gráfico 7: Agrupamiento utilizando clustering jerárquico entre bacterias fecales agrupadas por valores de ICA]

Al ver que el comportamiento no era el esperado y que no obtuvimos mucha información de la gráfica, optamos por hacer un t-SNE (método no lineal) en busca de una relación significativa entre el conjunto de bacterias fecales y el ICA.



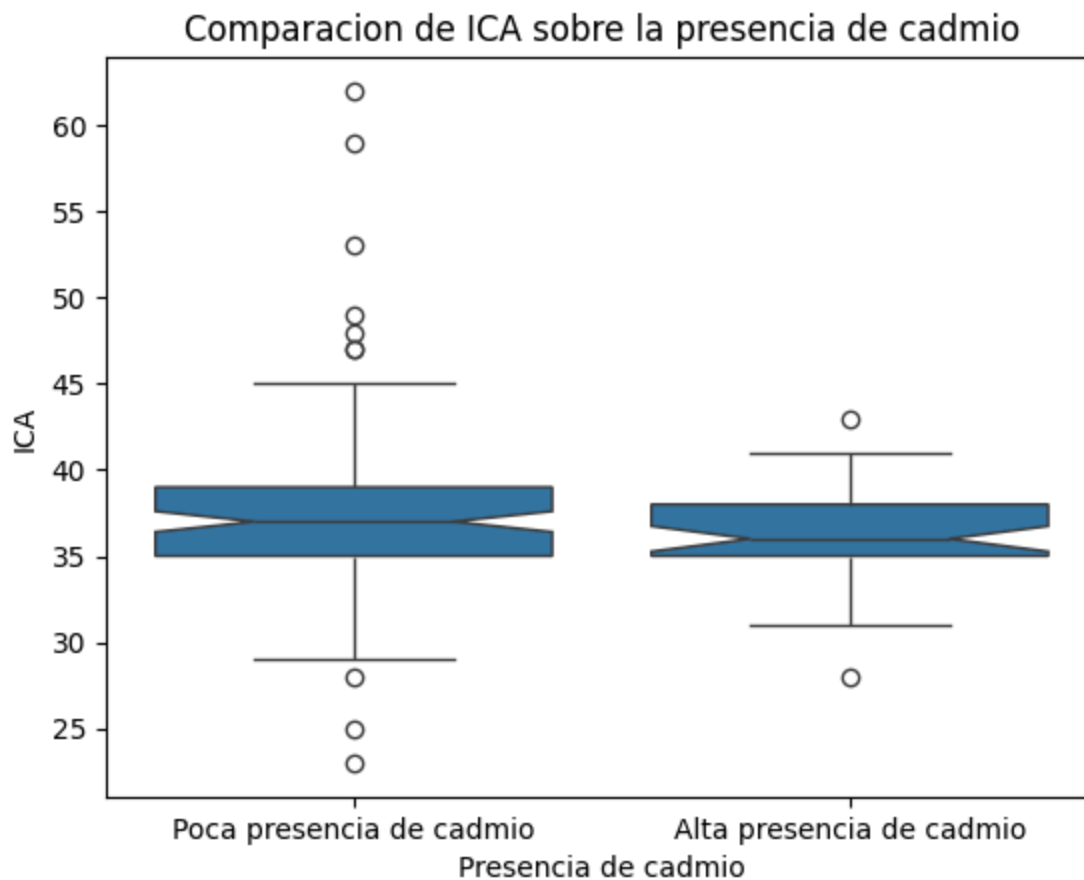
[Gráfico 8: Agrupamiento por t-SNE de bacterias fecales medidas por ICA]

Como era de esperarse, el gráfico arrojado mostró que no había un comportamiento de grupo significativo y por ende decidimos que continuar con un análisis de test de hipótesis iba a ser innecesario/no tendría sentido.

H2

Como mencionamos anteriormente, una de las hipótesis que teníamos en mente era la de que los valores de ICA eran distintos dada según los niveles de cadmio presentes durante la medición.

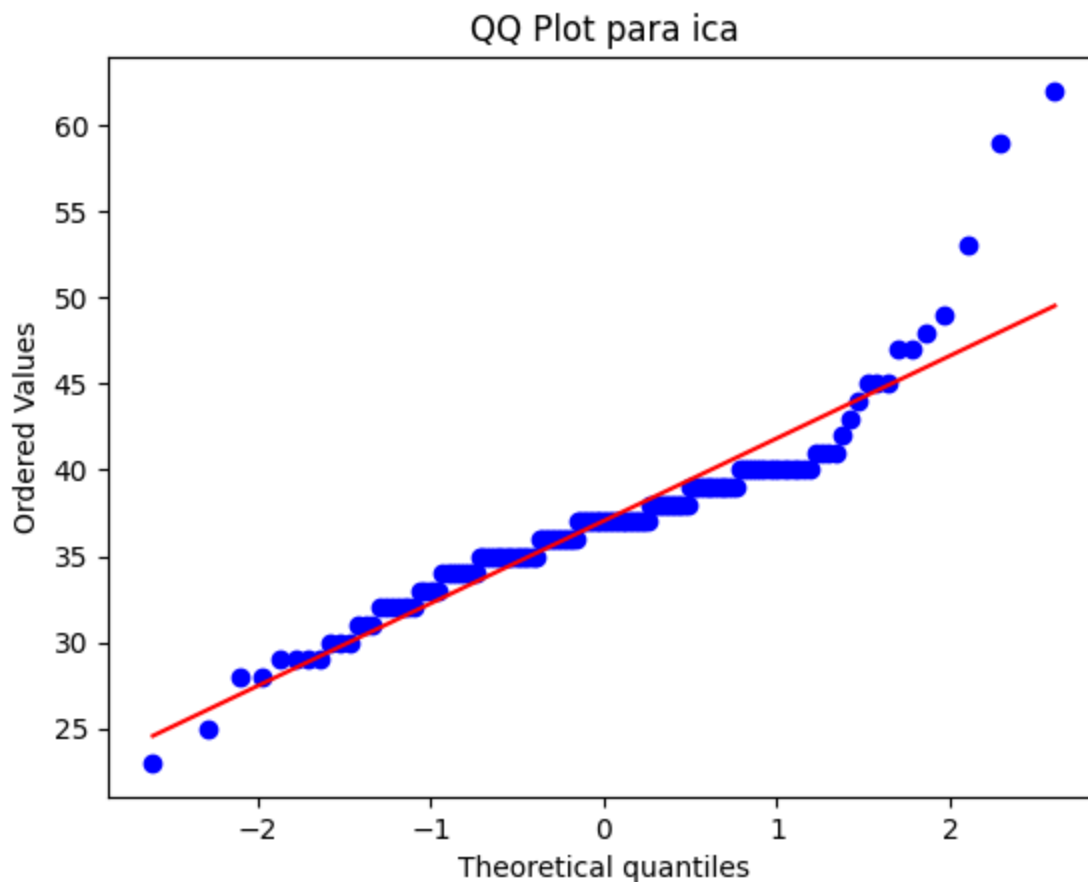
El primer acercamiento que realizamos fue un Box-Plot para visualizar de manera sencilla cómo era la dispersión de los valores del ICA en ambos grupos (Presencia de cadmio y ausencia del mismo), esperando encontrar alguna pista.



[Gráfico 9: Box-plot en donde se muestra la dispersión de los valores de ICA en cada grupo junto a la densidad de valores en cada cuartil]

Ver que la mediana de ambos grupos es similar ya nos daba una pista sobre que quizás no habría la influencia que esperábamos encontrar. Sin embargo, hasta no realizar un test de hipótesis no podríamos concluir nada. Por ello continuamos analizando la normalidad y homocedasticidad de la variable ICA, ya que dependiendo de cuales de estos dos supuestos se cumpla se debe elegir entre un test de hipótesis u otro.

Aplicando Shapiro-Wilk verificamos que la variable ICA no era normal ($p\text{-valor} = 0.000$) y por ende los subconjuntos de la misma tampoco lo serán.



[Gráfico 10: QQ-plot³ que muestra los puntos se desvían significativamente de la línea diagonal, indicando que los datos no siguen una distribución normal]

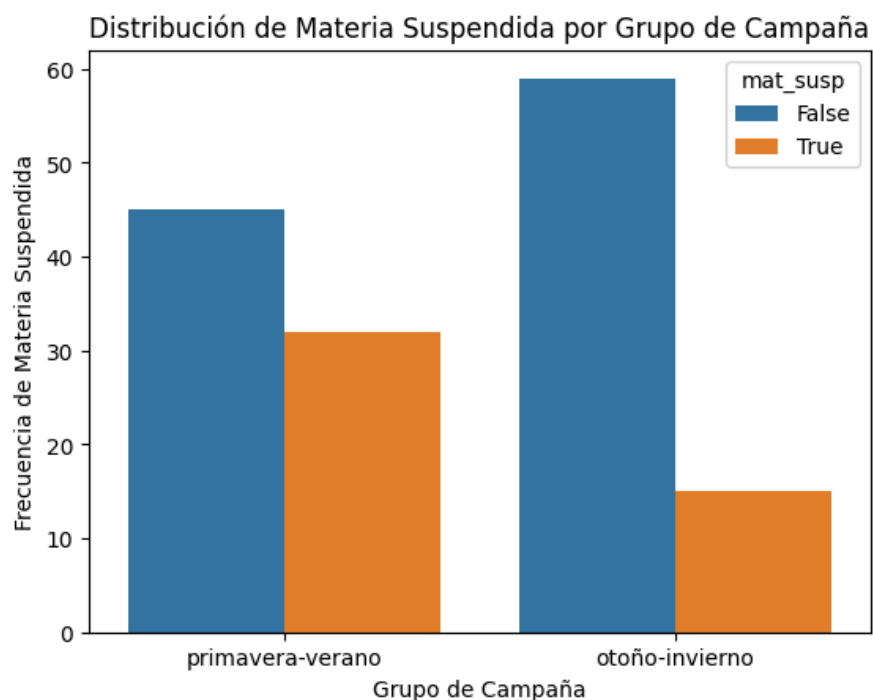
Y utilizando el test de Levene vemos que tampoco se cumple la homocedasticidad (p -valor = 0.049), algo que nos podíamos esperar al ver que en el box-plot los dos conjuntos tienen una varianza de valores significativamente distinta (El conjunto de poca presencia de cadmio cuenta con valores más dispersos).

Ahora que sabemos que la variable ICA no se distribuye normalmente y que los conjuntos a comparar no son homocedásticos, tenemos que optar por el test de Kruskal-Wallis para verificar la veracidad de nuestra hipótesis al ser el más adecuado. Al realizar el test obtenemos que se acepta la hipótesis nula (p -valor = 0.163), ya que no hay evidencia suficiente para rechazarla. Por lo tanto se concluye que no hay una diferencia significativa en el valor del ICA entre las muestras con alta presencia de cadmio y poca.

³ El QQ-plot (gráfico de cuantiles-cuantiles) compara los cuantiles de la distribución de nuestros datos con los cuantiles de una distribución teórica normal. Si los datos siguen una distribución normal, los puntos se alinean a lo largo de la línea diagonal.

H3

Para esta hipótesis planteamos la idea de analizar la materia suspendida en el agua por estaciones, pensamos que entre temporadas de calor (primavera-verano) y temporadas de frío (otoño-invierno) habría una diferencia significativa en la presencia de materia suspendida en el agua. Así sin más, comenzamos la exploración tomando los datos (presencia/ausencia de `mat_susp`) de las estaciones de frío y calor, y analizamos con el algoritmo de chi-cuadrado⁴ la veracidad de nuestra hipótesis. Haciendo cuentas obtuvimos un $p\text{-valor} \approx 0.008$, lo cual implica rechazar nuestra hipótesis nula ("No existe una dependencia directa entre la materia suspendida y la temporada del año"), por lo tanto existiendo una dependencia entre la materia suspendida y las estaciones.



[Gráfico 11: Gráfico de barras que muestra las frecuencias obtenidas para la presencia y ausencia de materia suspendida por temporada frío/calor]

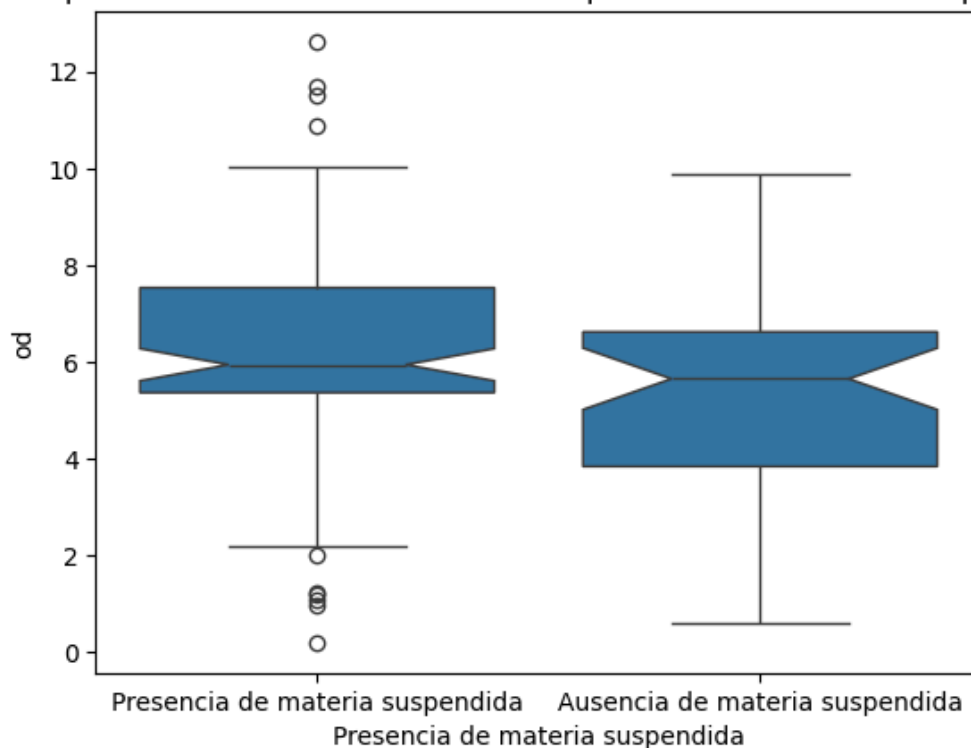
⁴ chi-cuadrado : se utiliza en estadística para evaluar si existe una asociación significativa entre dos variables categóricas o si la distribución de una variable observada se ajusta a una distribución esperada.

H4

La hipótesis que planteamos en este caso surge del sentido común, de pensar que si hay más materia suspendida en la superficie, la cantidad de oxígeno disuelto en el agua debería ser menor por la dificultad que supondrán estos materiales para el intercambio de oxígeno. Sin embargo aún no teníamos certeza de ello, por lo que lo pusimos a prueba usando los datos que nos proporcionó el dataset.

Para comenzar, separamos los valores registrados de oxígeno disuelto (OD) en dos grupos, las muestras tomadas con presencia de materia suspendida y las que no. Luego realizamos un Box-Plot con estos conjuntos para ver fácilmente cómo es que se dispersan los valores dentro de los mismos.

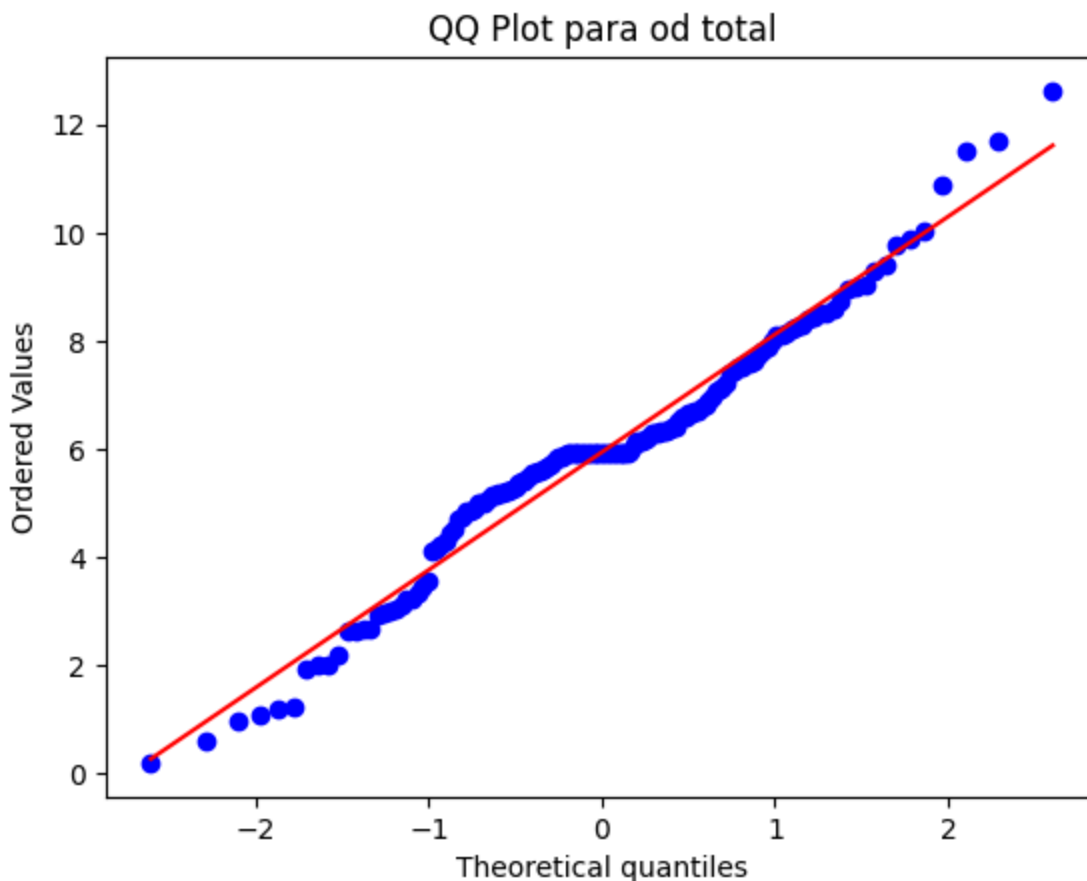
Comparacion de valores de od sobre la presencia de materia suspendida



[Gráfico 12: Box-plot en donde se muestra la dispersión de los valores de OD en cada grupo junto a la densidad de valores en cada cuartil]

Al observar el Box-Plot podíamos sospechar que quizás no había diferencia suficiente entre ambos conjuntos como para determinar siquiera que la presencia de materia suspendida afectaba los valores que pudiera tomar la variable de OD, e incluso si la influencia estuviera presente, nos sorprendió que la tendencia parecía ser contraria a nuestros pensamientos iniciales. Pero una vez más, hasta no hacer el test de hipótesis no podíamos sacar conclusiones, por eso comenzamos a verificar la normalidad y homocedasticidad de la variable OD con el fin de elegir el test de hipótesis indicado.

Para la normalidad hicimos uso de Shapiro-Wilks, obteniendo que la distribución de la feature no era normal ($p\text{-valor} = 0.003$) y por ende sus subconjuntos tampoco.



[Gráfico 13: QQ-plot que muestra que los puntos en los extremos se desvían de la línea diagonal lo suficiente para que los datos no sigan una distribución normal]

Con uno de los supuestos ya verificados solo nos restaba verificar la homocedasticidad de los dos conjuntos a analizar, y para ello usamos una vez más el test de Levene. Una vez realizado, verificamos que la homocedasticidad entre los conjuntos se cumplía ($p\text{-valor} = 0.969$), algo esperable al ver como en el box-plot la dispersión de valores dentro de cada conjunto era bastante similar entre sí (Los valores extremos de cada clase, dejando de lado los outliers, es bastante similar).

Ahora que ya sabíamos que solo cumplimos el supuesto de la homocedasticidad, debíamos usar Mann-Whitney U, ya que era el apropiado para esta situación. Al realizar el test obtuvimos como resultado que se rechazó la hipótesis nula ($p\text{-valor} = 0.01$) y por lo tanto verificamos que el valor de oxígeno disuelto es menor con presencia de materia suspendida.

Conclusiones

En definitiva, se utilizó un conjunto de mediciones en distintos puntos y épocas del año del Río de la Plata para interpretar las condiciones del agua. El estado inconsistente del dataset implicó una limpieza que demandó la mayoría del tiempo empleado en el trabajo. Sin embargo, el análisis permitió formular varias hipótesis, algunas de las cuales no resultaron como esperábamos.

Por ejemplo, en H1, esperábamos que las variables relacionadas a la materia fecal en conjunto explicaran diferencias en el Índice de Calidad del Agua (ICA). Nos sorprendió no encontrar un agrupamiento significativo al analizar estas variables junto con el ICA.

H2 también resultó inesperada; la alta presencia de cadmio no explicó una diferencia en el comportamiento de los valores del ICA, indicando que la contaminación por metales pesados no era tan influyente como suponíamos.

En contraste, H3 y H4 confirmaron nuestras expectativas iniciales. En H3, se observó que la época del año influye en la presencia de materia suspendida. De hecho, en el gráfico 11 se aprecia un aumento durante la época Primavera-Verano, con lo cual podría deducirse que durante estos momentos del año hay mayor presencia de materia suspendida e incluso se podría suponer que ese aumento se atribuye al incremento del turismo y la cantidad de desechos arrojados por consecuencia del mismo. Con este planteamiento, se podrían dedicar recursos a la limpieza de materia suspendida en estas temporadas de calor en específico. H4 mostró que la cantidad de oxígeno disuelto (OD) impacta significativamente en los niveles de OD, lo cual es un aspecto a considerar por el efecto que podría tener sobre la vida acuática. De hecho, siguiendo la deducción realizada a raíz de lo comprobado en la hipótesis anterior, se podrían proponer campañas de concientización en dichas épocas del año utilizando como argumento el efecto que tiene sobre el oxígeno en el agua y por ende en el ecosistema del río.

En conclusión, los aspectos medidos en el dataset dan lugar a varios análisis que pueden ser de mucha utilidad. Sería ideal aumentar la rigurosidad en la recolección de datos, de modo que la credibilidad de las hipótesis aumente al no verse tan afectadas negativamente debido a las inconsistencias en el muestreo.