TPE Fundamentos de la Ciencia de Datos

Integrantes:

Maximo Reynoso
Francisco Nelli
Kevin Carrera

Introducción

El propósito de este informe es analizar los datos provenientes de un dataset sobre la calidad del agua en el Río de La Plata durante el año 2023, el cual consta de un conjunto de 169 mediciones que abarcan diversos parámetros ambientales y de contaminación.

Este análisis incluye una exploración detallada de las variables de interés como oxígeno disuelto, coliformes fecales, y otros indicadores de contaminación, con el objetivo de caracterizar su comportamiento, identificar posibles anomalías, y evaluar la salud general del agua en diferentes estaciones del año.
A través de este informe, se busca proporcionar una comprensión profunda de la situación actual de la calidad del agua y validar o explorar hipótesis relevantes utilizando todos los conocimientos y conocimientos aprendidos durante la cursada.

Materiales

Para poder llevar adelante los análisis detallados en la notebook de jupyter hicimos uso de las siguientes librerías:

matplotlib: Biblioteca de gráficos 2D para visualización de datos.
numpy: Biblioteca para cálculos numéricos y matrices multidimensionales
pandas: Procesamiento y análisis de datos en estructuras tipo tabla.
scipy: Herramientas de cálculo científico.
seaborn: Visualización de datos, construida sobre matplotlib.
statsmodels: Modelos estadísticos y pruebas en Python.

Métodos utilizados

Limpieza de datos:

A la hora de empezar con el análisis sobre el dataset nos topamos con que el mismo no estaba en las mejores condiciones, presentaba muchos valores nulos o que no tenían mucho sentido, datos repetidos,variables que no tenían utilidad alguna o estaban en un formato que no se podía utilizar para el análisis, por lo que tuvimos que realizar una limpieza de los datos.

Para los valores faltantes optamos por utilizar una técnica de imputación llamada K-Nearest Neighbors el cual completa los valores faltantes basándose en los valores más cercanos según la distancia euclidiana.

Con respecto a las variables sin utilidad optamos por eliminarlas del dataset, ya que la propia definición de la variable no tenía utilidad o bien los valores que tomaba la volvía inutil, por ejemplo había una variable que en la mayoría de los casos tomaba el mismo valor.

Con el objetivo analizar los outliers utilizamos los gráficos conocidos como box-plot para tener una visión general y decidir qué hacer con estos valores.

Para las variables que estaban en un formato con el cual no podíamos trabajar optamos por cambiar el tipo de variable o bien aplicar alguna de las técnicas aprendidas durante la cursada.

Formulacion de hipotesis:

Una vez teniendo los datos limpios, nos pusimos a analizar qué posibles hipótesis podríamos plantear en base a las definiciones de las variables. Además utilizamos gráficos y algoritmos provenientes de las librerías anteriormente mencionadas para buscar otras no pensadas en base al contexto, herramientas como la matriz de correlación entre las variables:

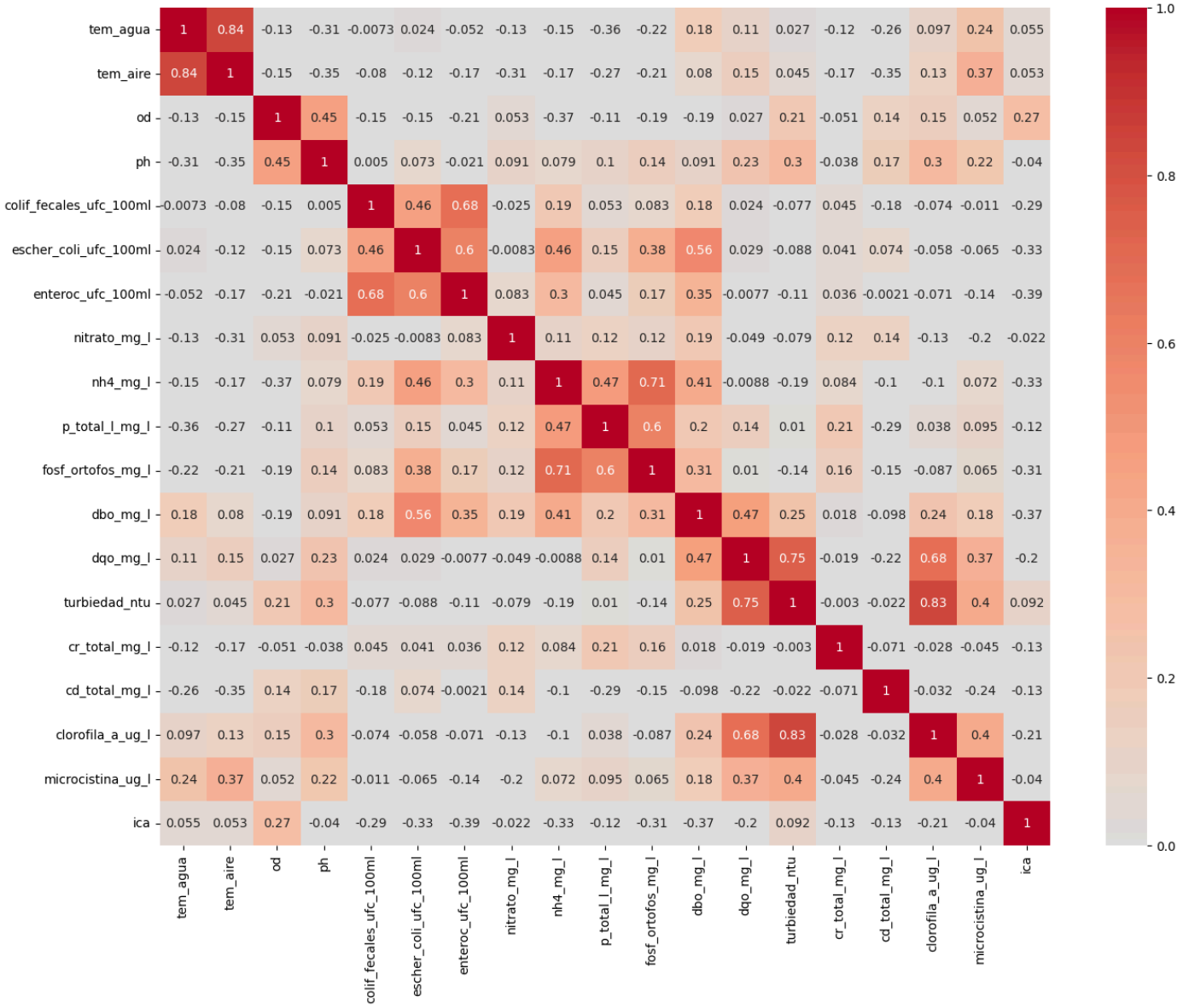

Figura 1: Matriz de correlación de todas las variables cuantitativas.

Completado el análisis exploratorio de los datos, se formularon hipótesis preliminares en base a las características observadas y la información contextual disponible. Estas hipótesis se plantearon con el objetivo de investigar posibles relaciones y patrones entre los distintos atributos del conjunto de datos, los cuales podrían ofrecer una comprensión más profunda del fenómeno estudiado y guiar el análisis posterior.

- Hipótesis cantidad de toxinas está condicionada por diversos factores:

Una de las hipótesis planteadas es que la cantidad de toxinas producidas por las algas podría estar influenciada por varios factores ambientales, tales como la temperatura del agua, la concentración de nitratos, el fósforo total, la concentración de microcistinas, o el pH .

Estas variables podrían ser afectadas por la presencia de toxinas, o bien, podrían influir en el crecimiento de las algas, lo que, a su vez, impactaría la cantidad de toxinas producidas.

Esta hipótesis surgió mediante la interpretación de las variables.

- Hipótesis demanda de oxígeno relacionada con nutrientes y materia fecal:

Los niveles elevados de demanda biológica de oxígeno (dbo_mg_l) pueden estar relacionados con altos niveles de nutrientes como nitrato, fósforo y amonio, además de que los indicadores de materia fecal contribuyen a esta demanda de oxígeno ya que proliferan los microorganismos aeróbicos.

Esta hipótesis surge al interpretar las variables y pre-suponer estas relaciones.

- Hipótesis del aumento de amoniaco según la estación del año:

Utilizando la tabla de correlaciones entre las variables y observando diferentes distribuciones de amoniaco entre las mediciones según la época donde se realizó, vimos que su concentración podría estar siendo afectada por la estación del año.

Claramente esta hipótesis fue desarrollada tras el análisis exploratorio de los datos.

- Hipótesis del aumento de ortofosfatos según la estación del año:

Similar al caso del amoniaco, utilizando la tabla de correlaciones entre las variables y observando diferentes distribuciones de ortofosfatos entre las mediciones según la época donde se realizó, vimos que su concentración podría estar siendo afectada por la estación del año.

Nuevamente, esta hipótesis fue desarrollada tras el análisis exploratorio de los datos.

- Hipótesis de calidad del agua discriminada según grupos:

Considerando las variables pertinentes y el índice de calidad de agua, vimos que podría considerarse la existencia de una sectorización entre las mediciones donde la calidad de agua difiere según el grupo del que se hable.

Esta hipótesis fue obtenida pasando el análisis exploratorio y diversas operaciones mediante la interpretación de los resultados obtenidos en ellas.

- Hipótesis de variación de oxígeno disuelto según época:

Diferenciando los datos según la época donde se realizó la medición, caliente (verano/primavera) o fría (invierno/otoño) vimos indicios de diferencia en el oxígeno disuelto.

Esta hipótesis la obtuvimos mediante la interpretación de las variables.

- Hipótesis la presencia de sustancias fecales afecta el Índice de Calidad del Agua (ICA):

Observando tres valores relacionados a la materia fecal y sus altas correlaciones entre sí, se nos ocurrió que puede haber una relación entre ellos y el índice de la calidad de agua, o que por lo menos lo afectan.

Esta hipótesis surgió durante la interpretación de las variables.

Evaluación de las hipótesis:

- Hipótesis del aumento de amoniaco según la estación del año:

Lo primero que se realizó fue la anteriormente mencionada matriz de correlaciones, y de ellas nos llamó la atención la alta correlación entre las concentraciones de amoniaco y ortofosfatos. Por lo que hicimos un scatter-plot comparando los valores de cada observación, enfrentandolas según su concentración de amonio y ortofosfatos.

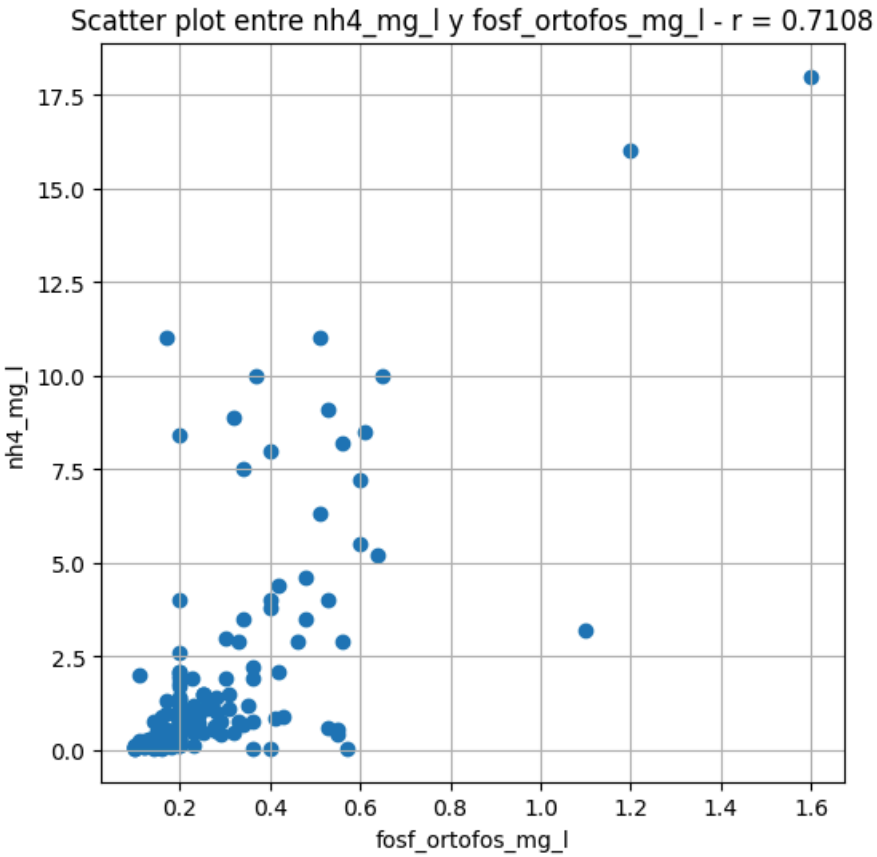

Figura 2: Scatter plot entre nh4_mg_l y fosf_ortosfos_mg_l

No nos decía mucho así tal cual estaba, así que se nos ocurrió diferenciar los valores según la estación de medición a ver si observamos algún comportamiento que valiese la pena.

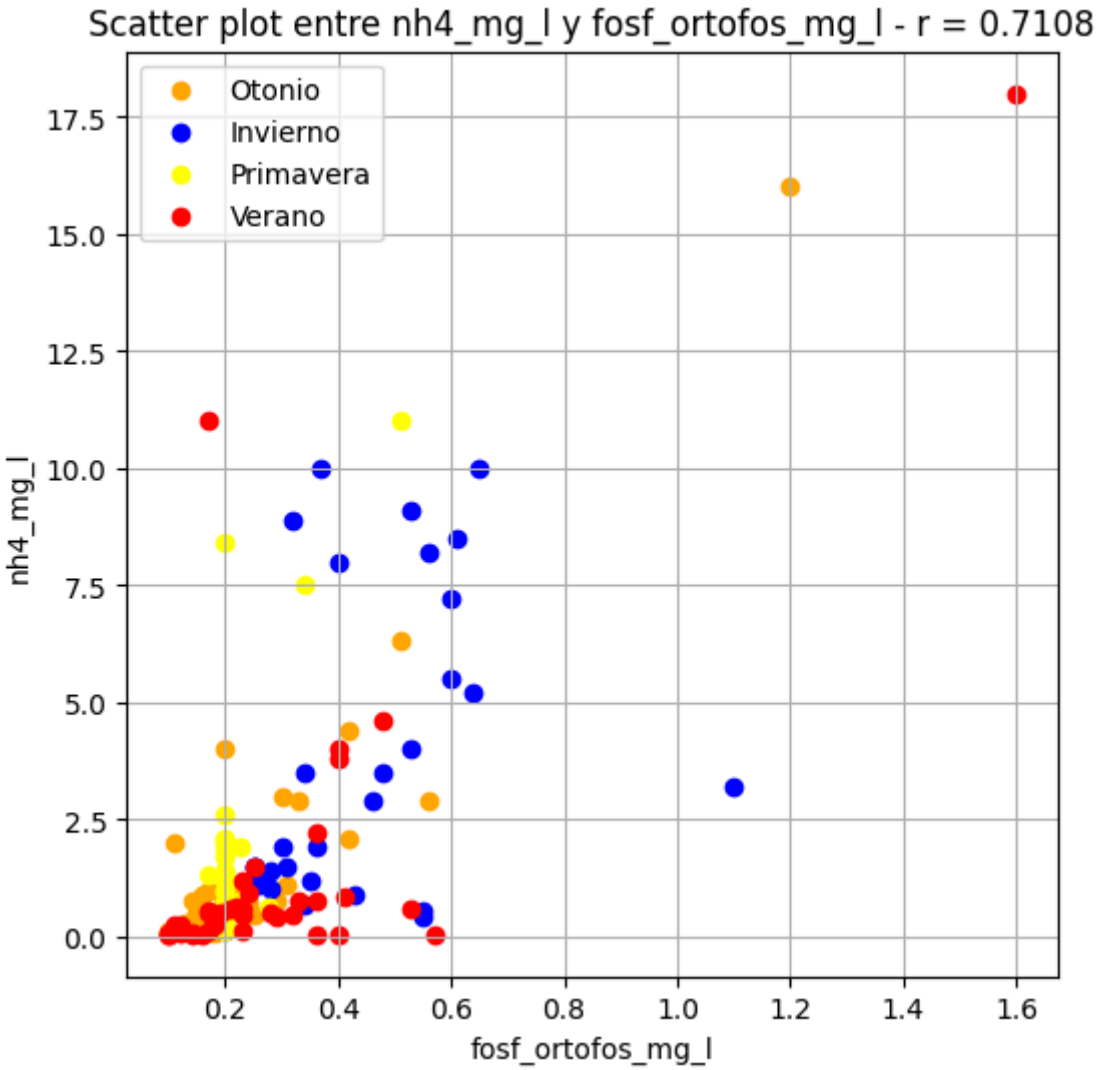

Figura 3: Scatter plot entre nh4_mg_l y fosf_ortosfos_mg_l separado por estacion

Observando este gráfico ya diferenciado, podemos empezar a ver comportamiento diferenciado según la estación del año. Sigamos chequeando la diferencia de valores en el caso del amoniaco.

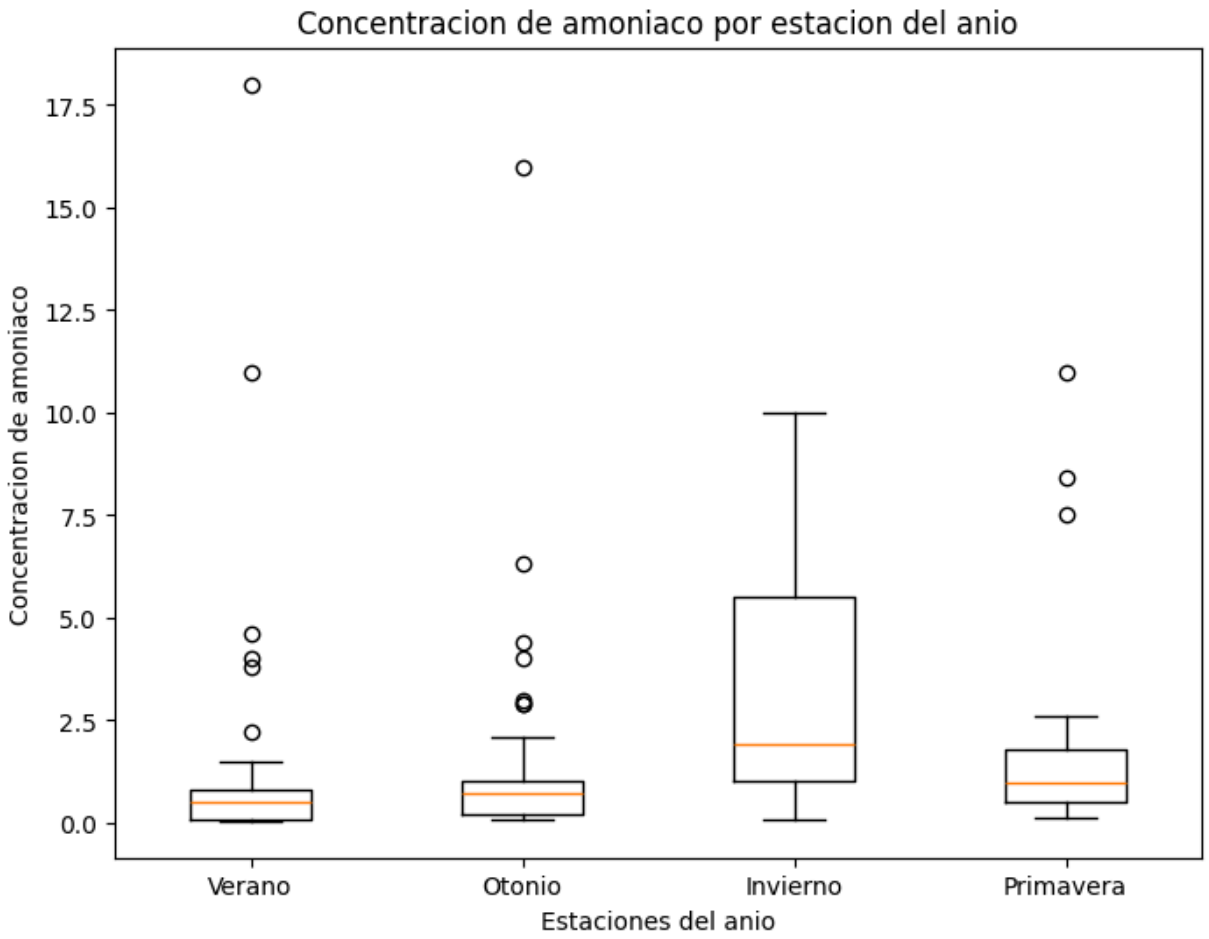

Figura 4: boxplot de concentraciones de amoniaco por estación.

Viendo un claro aumento de los valores en primavera tenemos el precedente para empezar a tratar a la hipótesis como tentativa, lo que significa que vale la pena testearla. Dicha prueba fue realizada siguiendo los criterios pertinentes y obtuvimos los siguientes resultados:

Se rechaza la hipótesis nula.
Existe una diferencia significativa en la concentración de amoniaco por estación del año cuando se midió.

- Hipótesis del aumento de ortofosfatos según la estación del año:

Volviendo al scatter-plot anterior, veamos la distribución de valores de ortofosfatos según la estación del año a ver si vale la pena considerar la hipótesis en cuestión.

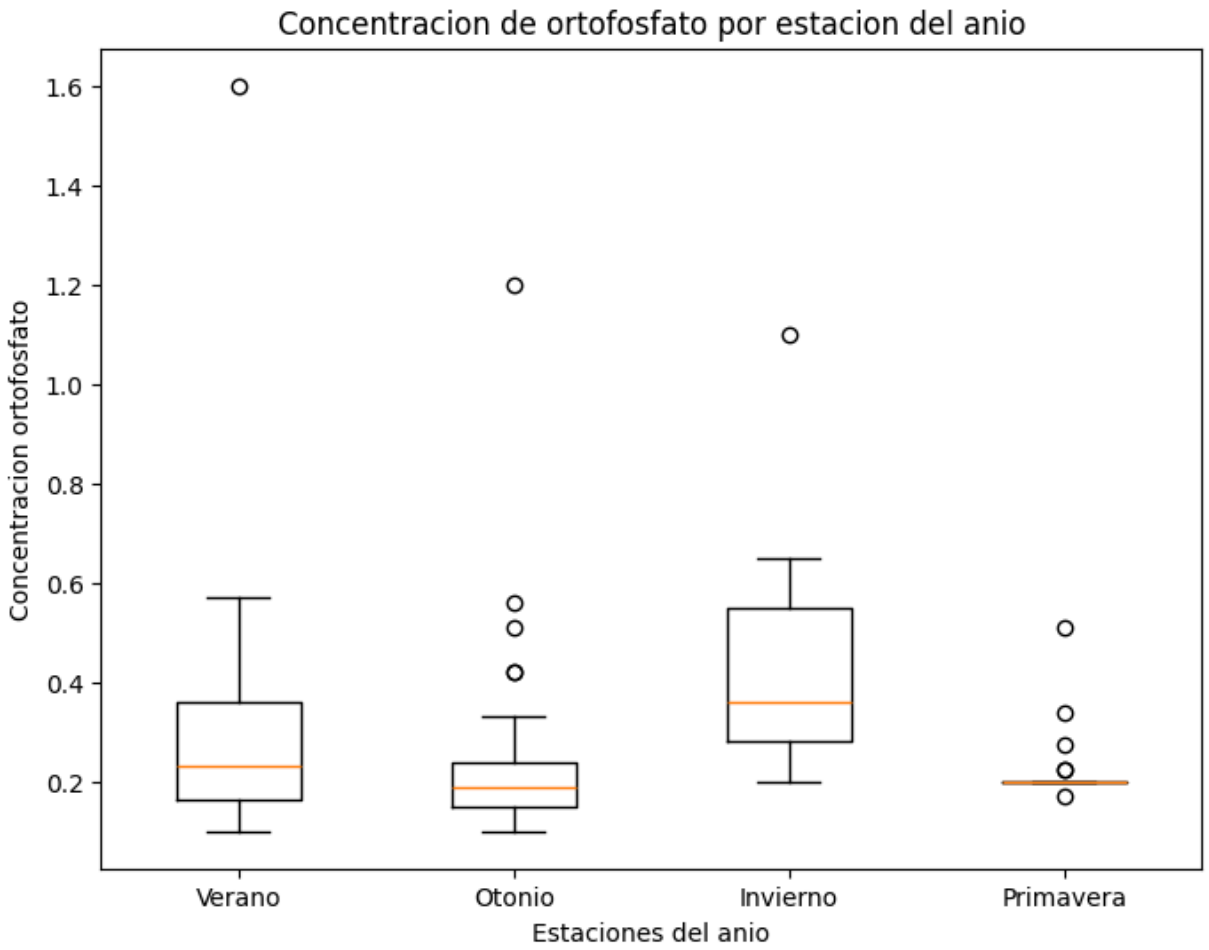

Figura 5: boxplot de concentraciones de ortofosfato por estación.

En este caso también vemos diferencias notables en las concentraciones de ortofosfatos según la estación del año, por lo que es justificado probar la hipótesis, dicha prueba fue realizada siguiendo los criterios pertinentes y obtuvimos los siguientes resultados:

Se rechaza la hipótesis nula.
Existe una diferencia significativa en la concentración de ortofosfato por estación del año cuando se midió.

- Hipótesis de calidad del agua discriminada según grupos:

Para este caso juntamos todas las variables de las mediciones salvo el índice de calidad de agua para ver cómo se distribuía, obtuvimos lo siguiente.

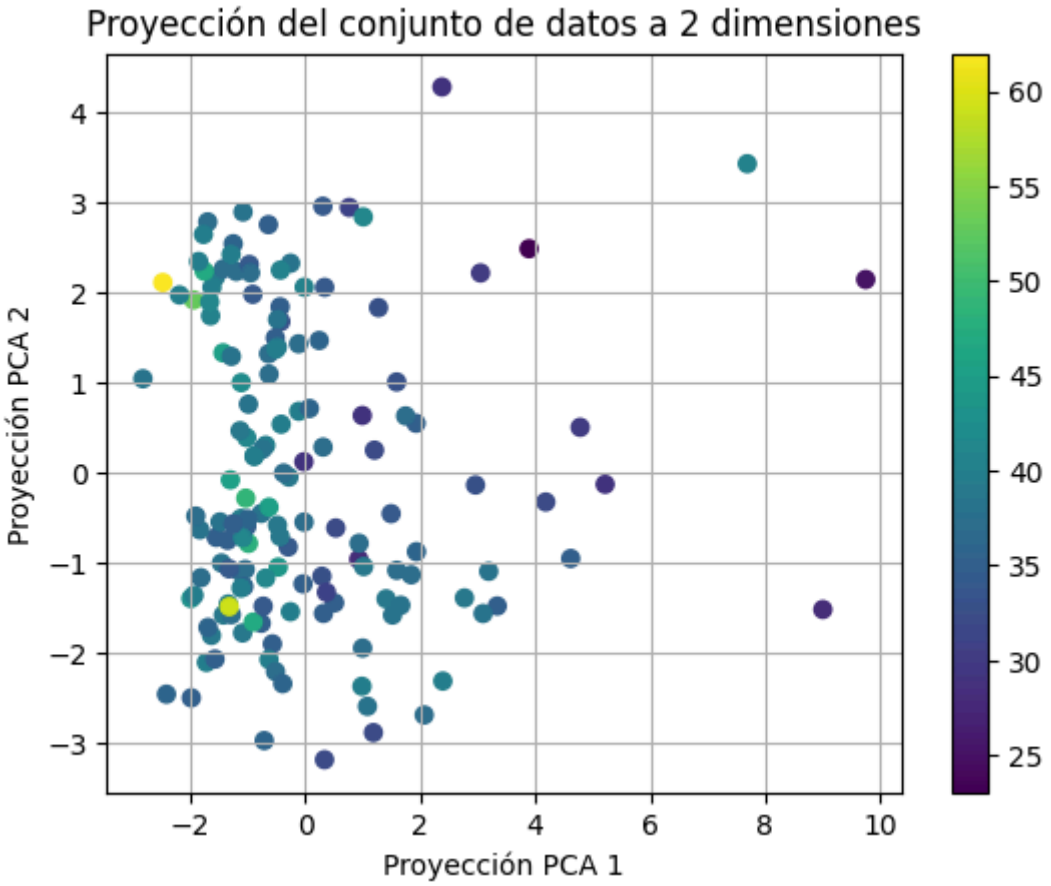

Figura 6: proyección en dimensión reducida.

Siendo los colores representaciones del valor del índice de calidad de agua, podemos ver unos grupos diferenciados, por lo menos 4, la cantidad de grupos óptima según nuestras mediciones. A estos grupos luego les evaluamos las concentraciones en para ver si se diferenciaban y daban validez a nuestra hipótesis.

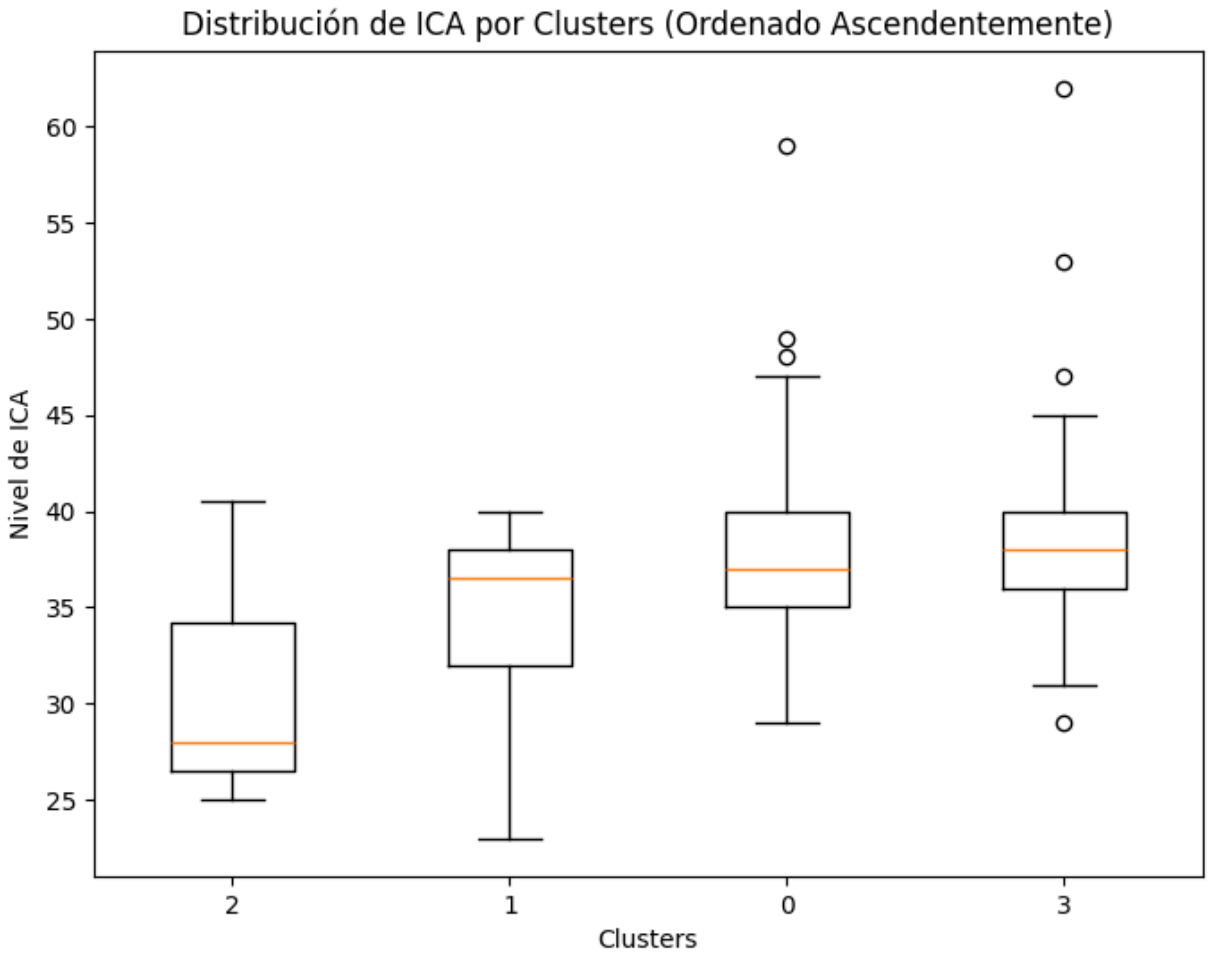

Figura 7: boxplot de distribución de ica por clusters.

Como claramente se ven estamentos diferenciados, por lo que hicimos el testeo de nuestra hipótesis y obtuvimos el siguiente resultado:

Se rechaza la hipótesis nula.
Existe una diferencia significativa en el índice de calidad de agua respecto a los clusters encontrados.

- Hipótesis cantidad de toxinas está condicionada por diversos factores/ Hipótesis demanda de oxígeno relacionada con nutrientes y materia fecal:

El enfoque que tuvimos para estas dos hipótesis fue casi el mismo, ya que planteamos que una variable se ve afectada por un grupo de estas.
Como primer paso utilizamos estadísticas descriptivas para analizar los valores que toma y tener una comprensión más profunda de los datos. Además estudiamos las distribuciones y mediante box-plots los valores atípicos.

También analizamos las correlaciones entre las variables en busca de algo que nos motive a seguir con el análisis.

Luego para poder graficar y ver qué pasaba con los datos utilizamos PCA para poder proyectar todas esas variables de interés a una dimensión con la que podamos trabajar para posteriormente ver si se relacionaba con la variable de interés de alguna manera, como muestra las siguientes imágenes.

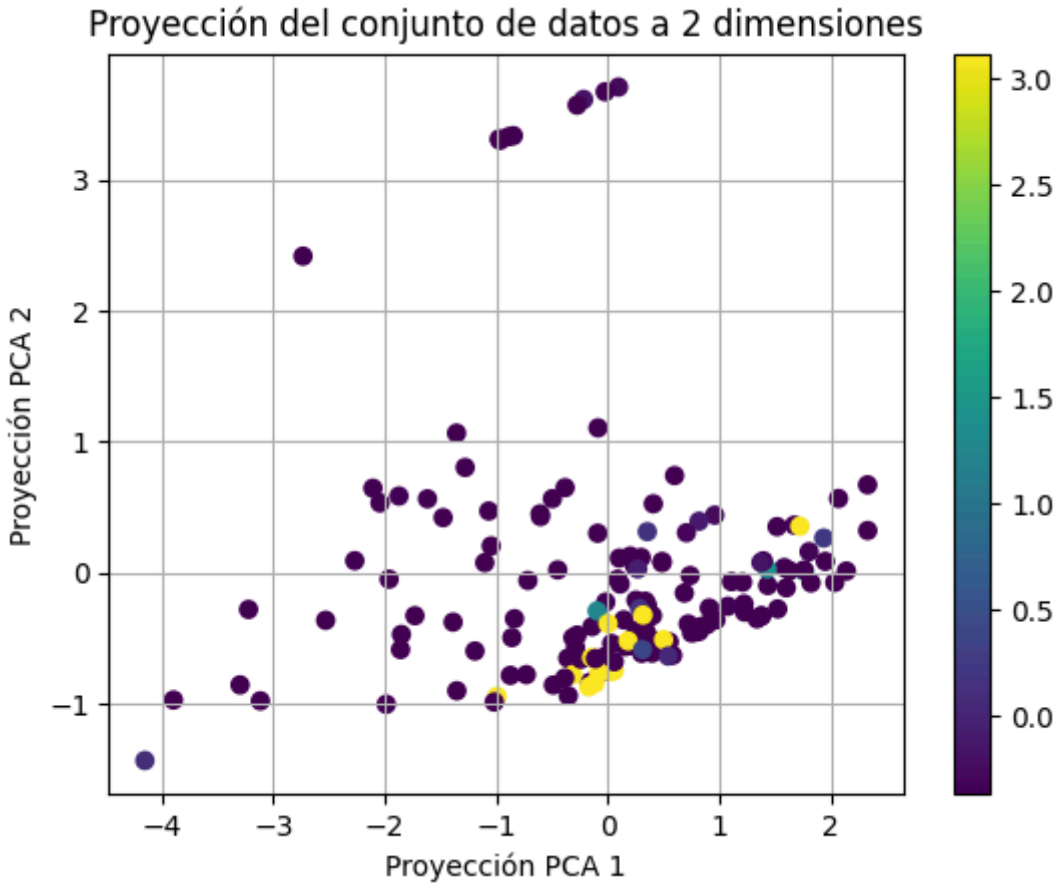

Figura 8: proyección de reducción de dimensionalidad con toxinas/microcistinas en color.

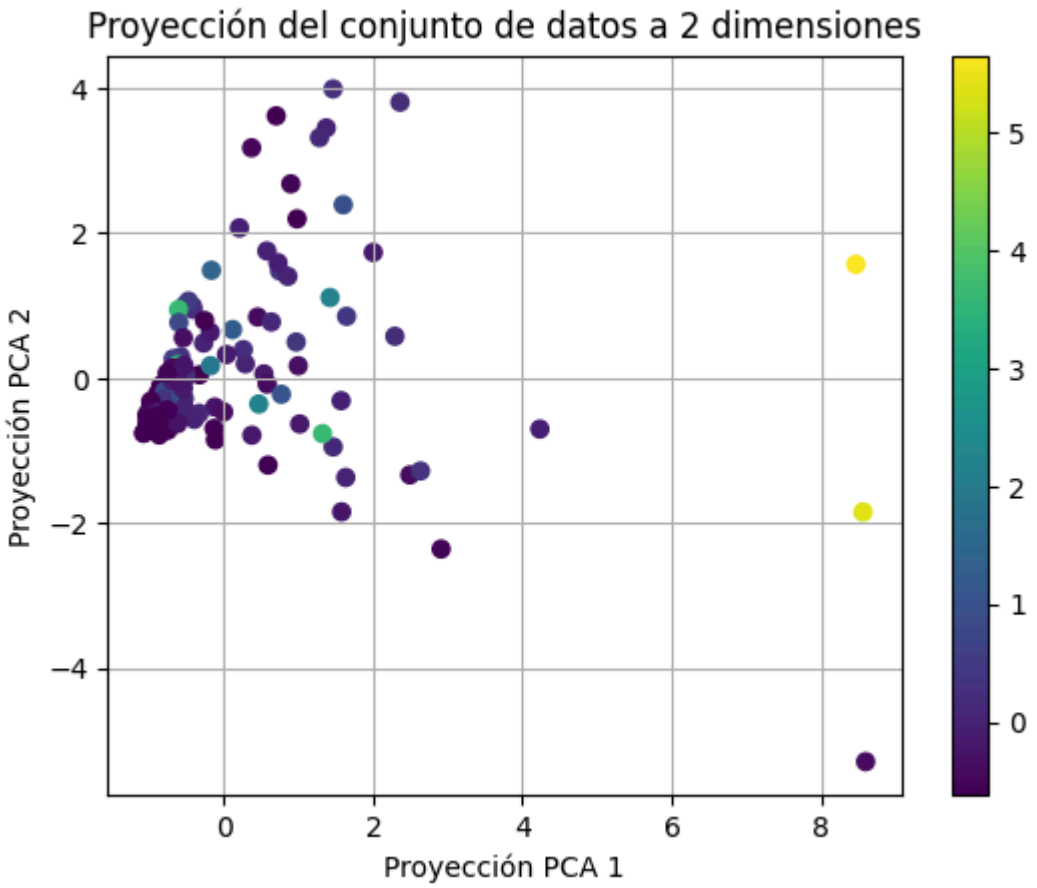

Figura 9: proyección de reducción de dimensionalidad con oxígeno en color.

Lamentablemente ambas hipótesis quedaron descartadas ya que al someterlas a un análisis no se noto alguna especie de relación que nos motive a realizar algún test.

-Hipótesis de variación en concentraciones de oxígeno según época:

Para empezar a analizar esta hipótesis lo que hicimos fue armar dos grupos, uno con los datos de las estaciones cálidas junto con el oxígeno disuelto y otro con los de las estaciones frías.

Sin embargo, obtuvimos que no había suficiente evidencia para afirmar que hubiese una diferencia significativa en los niveles de oxígeno disuelto entre las distintas estaciones.

- Hipótesis la presencia de sustancias fecales afecta el Índice de Calidad del Agua (ICA):

Durante el análisis de los datos, surgió la idea de investigar la relación entre la presencia de sustancias fecales, específicamente enterococos, Escherichia coli y coliformes fecales, y el ICA. Para ello, decidimos reducir la dimensionalidad de estas tres variables, de modo que pudiéramos representar su variación en un plano y observar así su impacto sobre el ICA, para ello llegamos a un resultado alentador mediante la técnica de *t-distributed Stochastic Neighbor Embedding.*

Usando el método t-SNE, observamos que existe una aparente relación entre las variables de sustancias fecales y el ICA. Sin embargo, debido a que esta relación no es lineal, actualmente no contamos con las herramientas necesarias para profundizar en su análisis de manera efectiva, al no tratarse de una relación no lineal. Por lo que no podemos asumir que afectan activamente el índice de calidad del agua.

Conclusiones

A partir de lo obtenido mediante los análisis que realizamos, pudimos sacar las siguientes conclusiones:

- La concentración de amoniaco aumenta según la estación del año.
- La concentración de ortofosfatos aumenta según la estación del año.
- Las mediciones son agrupables en grupos diferenciados según su índice de calidad de agua.
- No se puede afirmar que la microcistina se encuentra condicionada por factores ambientales (considerando pH, la temperatura del agua, la concentración de nitratos y el fósforo total).
- No se puede afirmar que la demanda biológica de oxígeno está correlacionada con altos niveles de nutrientes como nitrato, fósforo y amonio, y los indicadores de materia fecal.

- No hay una evidencia significativa para diferenciar la concentración de oxígeno según las estaciones cálidas o frías.

- No podemos afirmar que la presencia de índices relacionados a la materia fecal alteren significativamente el índice de la calidad del agua.

Referencias

Documentación:
https://seaborn.pydata.org/
https://matplotlib.org/stable/contents.html
PCA con Python
Clustering con Python

Información:
Gestión del Aguas Industriales: la Relación Entre el pH y las Algas
Reducir la demanda biológica de oxigeno (DBO) en aguas residuales
Eliminación del fósforo presente en aguas residuales
La química del agua: Cómo la composición de los lagos afecta a la vida acuática - Atlas de Ecosistemas