Maestría en Inteligencia Artificial Aplicada

Ciencia y análitica de datos

Dalina Aidee Villa Ocelotl (A01793258)

Miguel Guillermo Galindo Orozco (A01793695)

Actividad semanal 9: Reto Parte 1

Limpieza, análisis, visualización y k means

Profra. María de la Paz Rico Fernández

15 de noviembre de 2022

Reto: Parte 1

Resumen

El presente trabajo considera la investigación y la aplicación de los conocimientos relacionados a limpieza, análisis de la base de datos, visualización y agrupamiento de datos, sobre aspectos técnicos en gestión del agua subterránea. La mejora del conocimiento del recurso natural mejorará su gestión y su uso.

Introducción

En el análisis e investigación de problemas aplicados la calidad y fuente de datos es muy importante pero sobre todo el tipo de análisis y la forma en que se aborde el tratamiento de los datos, un gran aporte es el comportamiento de los datos por ello en esta primer parte se realizar la limpieza y análisis descriptivo de los datos.

La importancia de este problema radica en que el conocimiento de las características generales de una rea permitirá conocer si se puede haber agua en su interior, y con ello la capacidad que se obtiene de extracción de agua, así mismo se mide la infiltración de agua a través de precipitaciones, evaporación estimada y el caudal que llevan los ríos aledaños.

Contexto de aplicación

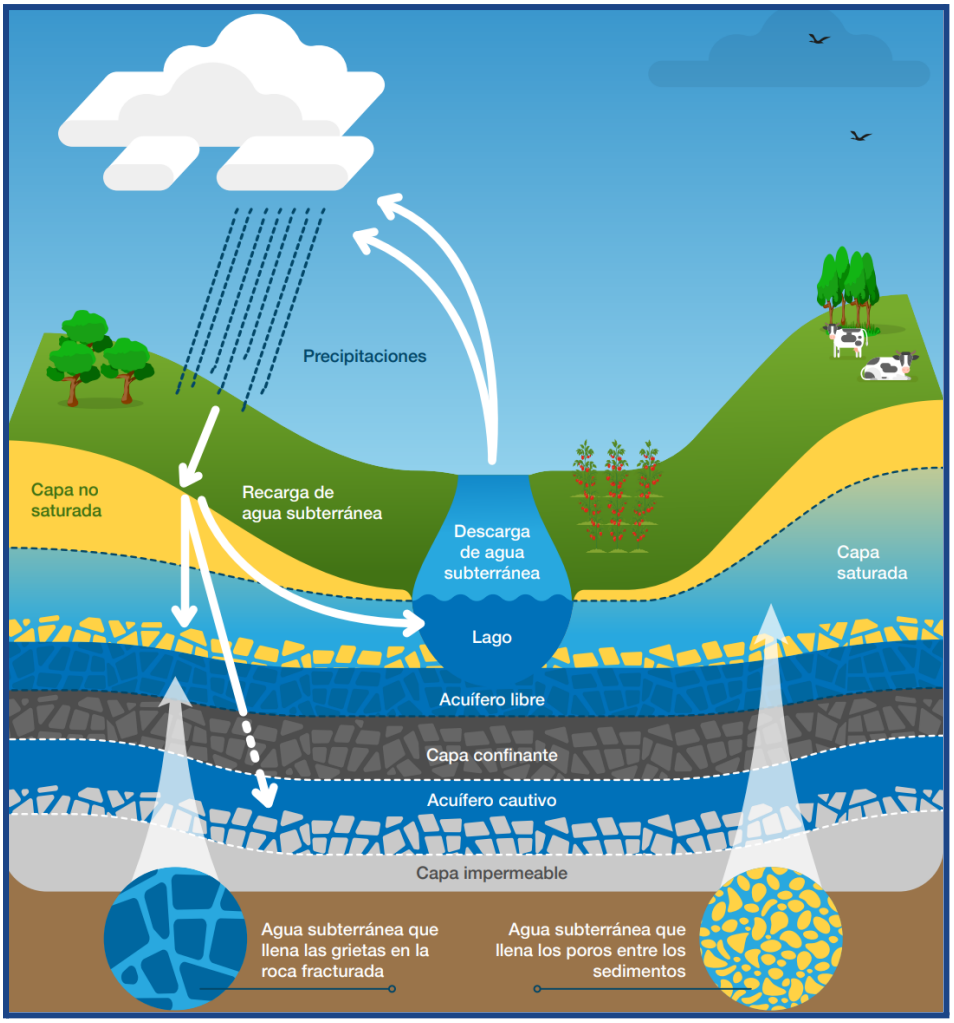
Como en toda aplicación es importante el contexto de aplicación acerca del tipo de aguas, en este caso se decidió estudiar aguas subterráneas.

Aguas subterráneas

Las aguas subterráneas son aquellas que habitan debajo de la tierra, tan solo en México se tienen identificados 653 acuíferos y el 38.7% proviene de esas fuentes por lo que su explotación es importante para satisfacer las necesidades de la sociedad mexicana actual.

Una de las principales características es que proviene del agua de lluvia y que constantemente se encuentra en un proceso de limpia y purificación, denominado infiltración de la lluvia, este proceso se desarrolla de manera más lenta que encima de la superficie. Esta agua se almacena de forma natural en depósitos debajo de la superficie terrestre, conocidos como mantos acuíferos subterráneos,

Una observación importante es el constante flujo del agua subterránea, eso se debe a la porosidad y permeabilidad de los mantos acuíferos y los ríos que los conectan entre ellos, que se expanden por todas las cuencas hidrográficas completas.



Calidad del agua subterránea

Se refiere a la temperatura del agua, la cantidad de sólidos disueltos y a la ausencia de contaminantes tóxicos y biológicos, para ellos es necesario conocer sus condiciones fisicas, quimicas y microbiologicas, donde las concentraciones de agentes tóxicos se dan por aportes externos y no a condiciones naturales.

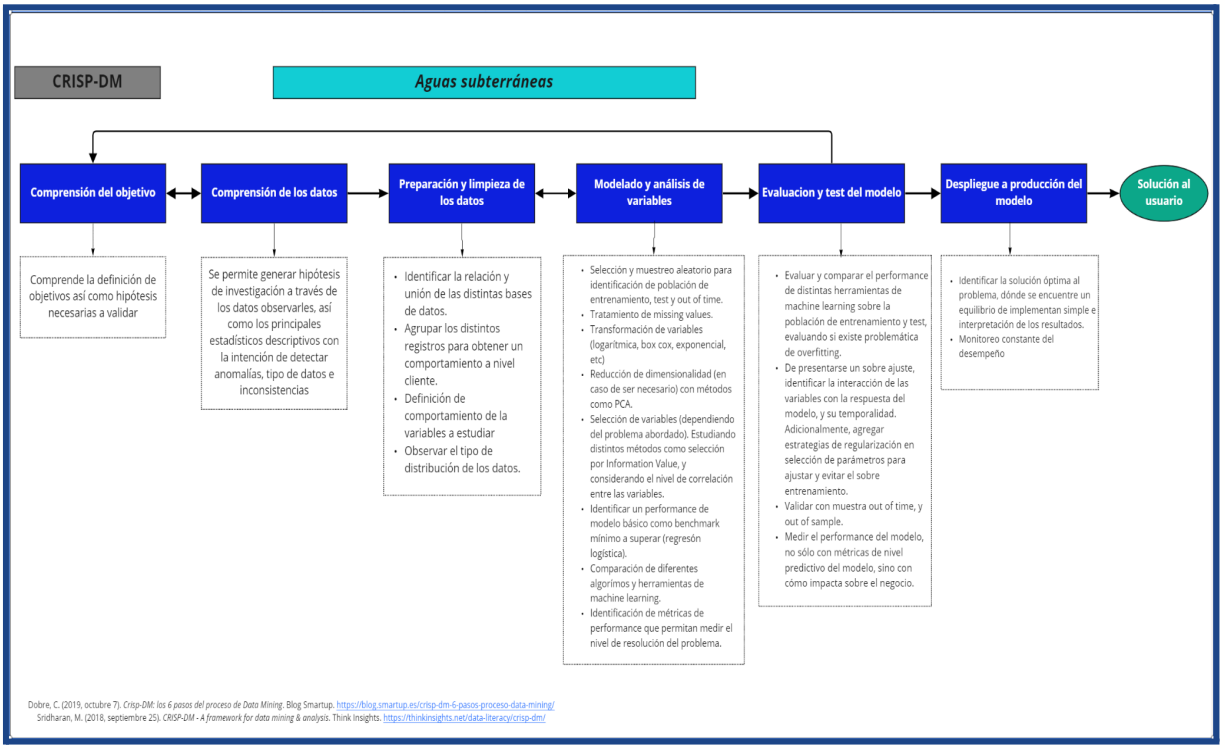
Hidrogeoquímica y calidad del agua subterránea

Tienen su origen en la procedencia de sus iones, provenientes de la disolución de las diferentes formaciones geológicas, se clasificó las familias de agua predominantes, teniendo que en la mayor parte la zona predomina la sódica Clorurada, agua característica de la intrusión por salinidad marina. Hacia la parte poniente y sur de la zona se localizan zonas de familias mixta sódica Clorurada y mixta Cálcica Clorurada.

Metodología de investigación

De acuerdo a las metodologías con las que se puede abordar la investigación, se plantea la metodología de CRISP-DM (Cross-Industry Standard Process for Data Mining), contiene principalmente 6 fases aunque estas pueden ajustarse de acuerdo a la investigación. La importancia de aplicar una metodología reconocida es porque favorece el proceso de

investigación científica, además de que un método se necesita para ordenar, esquematizar, registrar e interpretar datos.



https://miro.com/app/board/uXjVPTib1mw=/

Desarrollo de la investigación

Comprensión del objetivo:

Se tiene una base de datos de información de acuíferos en México su entendimiento es primordial porque garantiza el uso adecuado de los recursos naturales.

La liga del desarrollo de limpieza de datos: https://github.com/PosgradoMNA/actividades-del-projecto-equipo_88/blob/main/Reto_Eq_88_parte_1.ipynb

Comprensión de los datos:

Tenemos una base de datos de calidad del agua de 5000 sitios de monitoreo del año 2020, con 57 columnas y 1068 registros, dentro de las cuales se rastrean las condiciones del agua y los elementos que contienen para determinar su calidad.

Tipo de Calidad	Condicion Excelente
CALIDAD DEL AGUA PARA COLIFORMES FECALES	Agua potable. Agua no contaminada o condicion normal. No hay evidencia de alteracion en los valores de la calidad bacteriologica para el cuerpo de agua subterraneo
CALIDAD DEL AGUA PARA CADMIO	Excelente para riego de todo tipo de cultivos
CALIDAD DEL AGUA PARA ARSENICO	Agua potable. Agua no contaminada o condicion normal
CALIDAD DEL AGUA PARA ALCALINIDAD	Excelente para riego de todo tipo de cultivos
CALIDAD DEL AGUA PARA PLOMO	Agua potable. Agua no contaminada o condicion normal
CALIDAD DEL AGUA PARA NITROGENO DE NITRATOS	Agua potable. Agua no contaminada o condicion normal
CALIDAD DEL AGUA PARA MANGANESO	Agua potable. Agua no contaminada o condicion normal
CALIDAD DEL AGUA PARA MERCURIO	Agua potable. Agua no contaminada o condicion normal
CALIDAD DEL AGUA PARA FLUORUROS	Agua potable. Agua no contaminada o condicion normal
CALIDAD DEL AGUA PARA HIERRO	Agua potable. Agua no contaminada o condicion normal
CALIDAD DEL AGUA PARA DUREZA	Agua potable. Agua no contaminada o condicion normal

Análisis descriptivo de los datos

Para conocer las características más destacables del conjunto de datos y por variable, ejemplo de ello es el conocer el número de observaciones, las medidas de tendencia central. La estadística como rama de las matemáticas permite interpretar información para conocer la variabilidad de los datos, permitiendo su medición numérica.

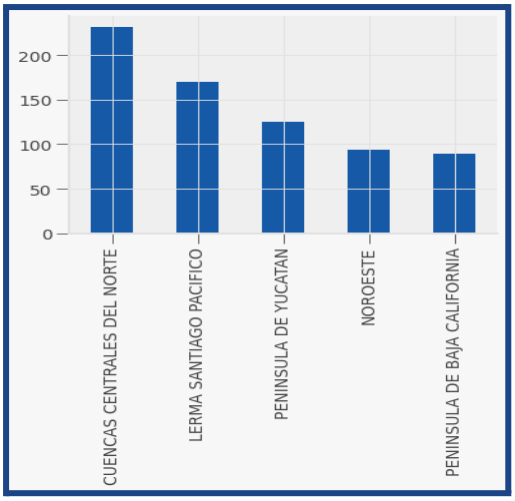
Se validan valores nulos

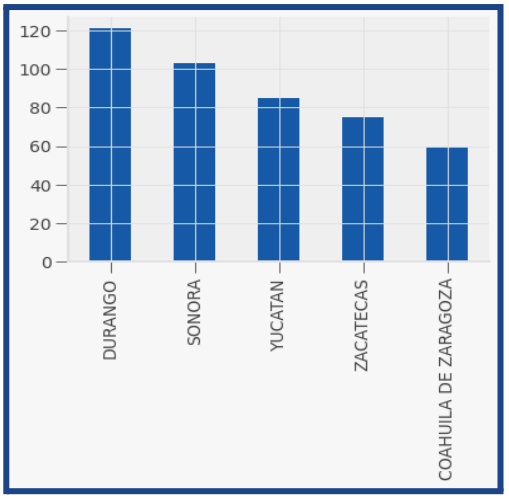
Solo se identificaron 13 variables con valores nulos, en distintos porcentajes.

Columna ALC_mg/L tiene 4 valores nulos, que representa 0.38% del total
Columna CALIDAD_ALC tiene 4 valores nulos, que representa 0.38% del total
Columna CONDUCT_mS/cm tiene 6 valores nulos, que representa 0.56% del total
Columna CALIDAD_CONDUC tiene 6 valores nulos, que representa 0.56% del total
Columna SDT_mg/L tiene 1068 valores nulos, que representa inf% del total
Columna SDT_M_mg/L tiene 2 valores nulos, que representa 0.19% del total
Columna CALIDAD_SDT_ra tiene 2 valores nulos, que representa 0.19% del total
Columna CALIDAD_SDT_salin tiene 2 valores nulos, que representa 0.19% del total
Columna DUR_mg/L tiene 1 valores nulos, que representa 0.09% del total
Columna CALIDAD_DUR tiene 1 valores nulos, que representa 0.09% del total
Columna N_NO3_mg/L tiene 1 valores nulos, que representa 0.09% del total
Columna CALIDAD_N_NO3 tiene 1 valores nulos, que representa 0.09% del total
Columna CONTAMINANTES tiene 434 valores nulos, que representa 68.45% del total

Análisis de Variables categóricas

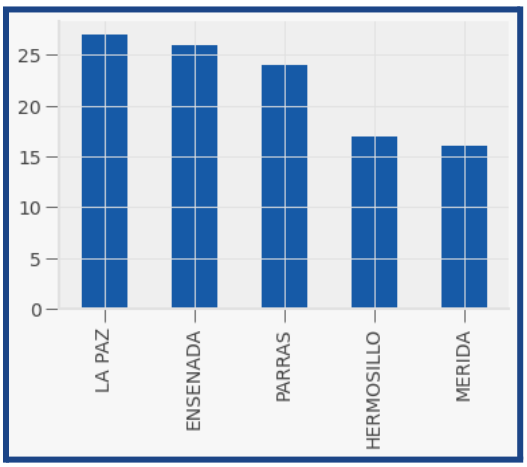
Organismo de Cuenca: De acuerdo al análisis de la primera sección encontramos que se trata de una variable categórica, con 13 valores distintos, y sin valores nulos, donde el valor más repetido es CUENCAS CENTRALES DEL NORTE. que es justamente reservas acuíferas de lugares donde hace más calor por lo que son más profundas.

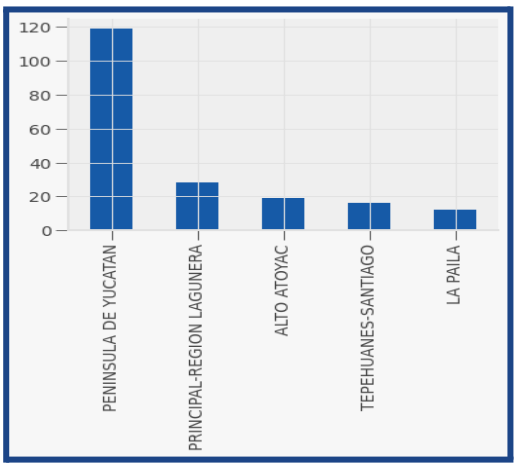




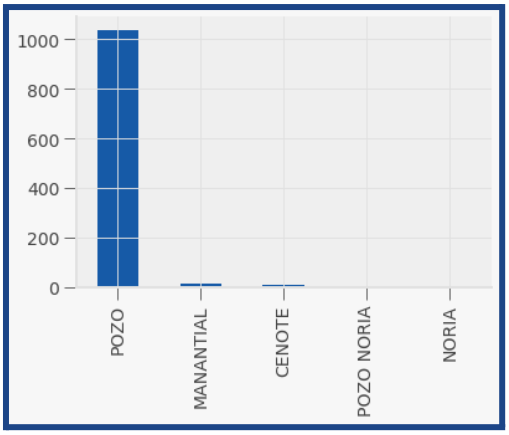
Estado: Variable categórica sobre el estado del registro analizado, sin valores nulos, donde el estado más repetido es DURANGO. Los principales estados donde se encuentran mantos acuíferos son los que han resultado más complicados de explotar.

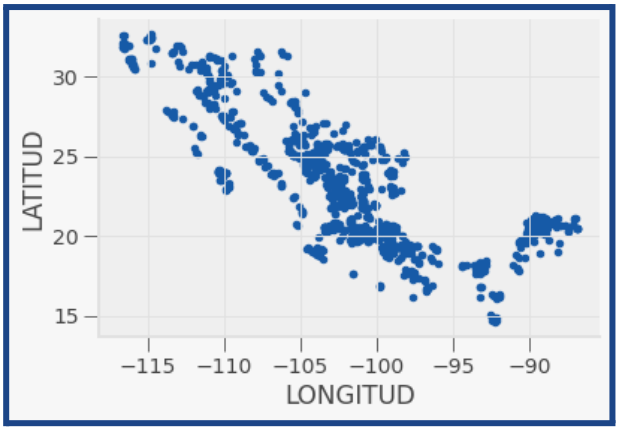
Municipio: Es otra de las variables categóricas que muestran parte de la región en la que se ubican los acuíferos, en este caso el que más se repite es la Paz, en Baja California Sur. El acuífero está sobre cohesionado y se encuentra en condiciones de sobreexplotación debido a que las extracciones anuales han sobrepasado la disponibilidad total del agua, y de acuerdo con la Ley Federal de Derechos de Materia de Agua en 2015 el acuífero se clasificó como zona de disponibilidad 2. La cuenca de La Paz de superficie de 947 km2 , situada en la porción suroriental de la península de BCS, se conforma de cinco subcuencas hidrológicas principales, la de los arroyos El Cajoncito, arroyo La Paz y arroyo El Datilar, La Palma y El Salto.





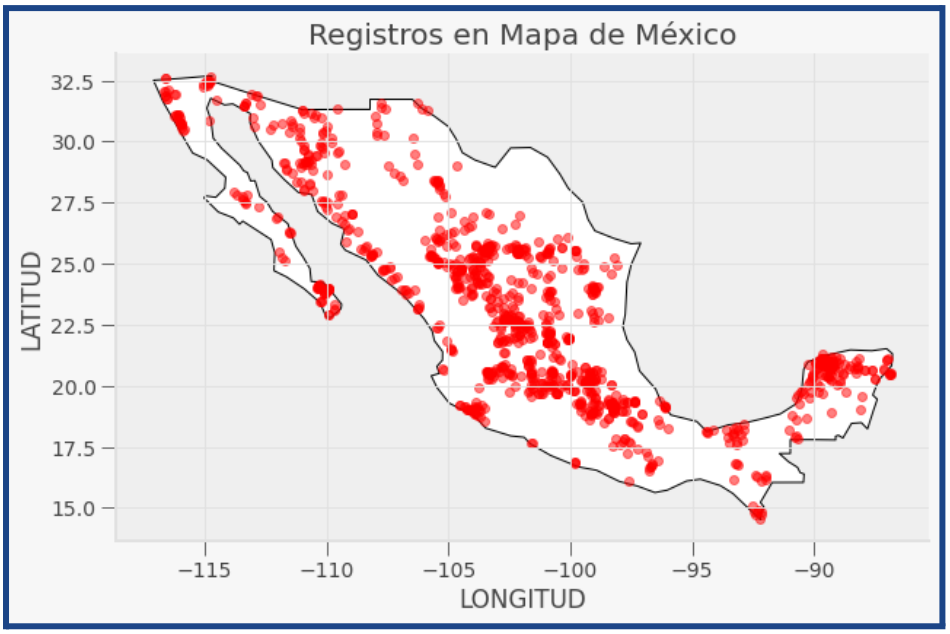
Acuífero: La variable acuífero es del tipo categórica, sin valores nulos, donde el más repetido es la península de Yucatán, este denominado el “El acuífero Península de Yucatán”, clave 3105, está formado por calizas y depósitos de litoral. Se trata de un acuífero libre, costero, kárstico, muy permeable y notablemente heterogéneo con respecto a sus propiedades hidráulicas. Debido a la presencia de la cuña de agua marina que subyace al acuífero, el espesor saturado de agua dulce es reducido, de aproximadamente 30 metros, aunque se incrementa hacia tierra adentro (“Diario Oficial de la Federación”).





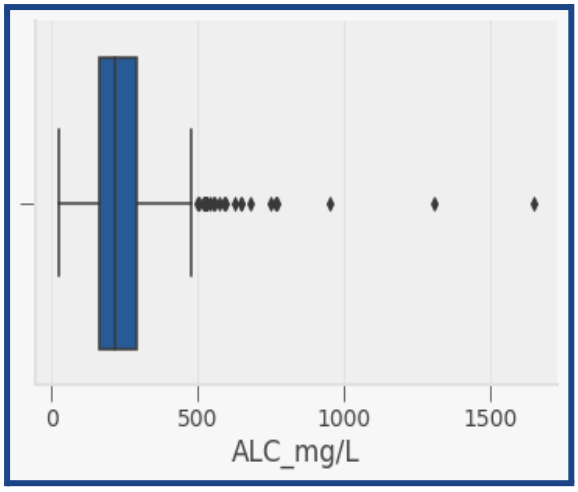
SUBTIPO: La variable que define el subtipo de acuífero para mostrar las características de si es cerrado o no es del tipo categórica, sin valores nulos, donde el más repetido es el subtipo POZO.

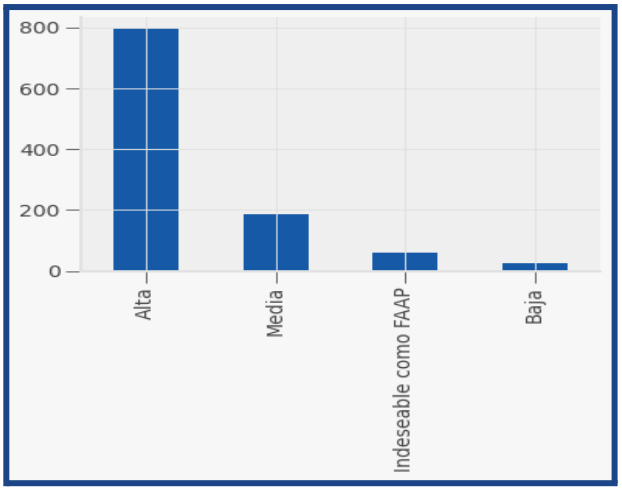
LONGITUD y LATITUD: Podemos observar la ubicación mediante la latitud y longitud de cada uno de los acuíferos y darnos cuenta de su concentración. El país está dividido en 13 regiones hidrológico administrativas (““Estudio de la hidrogeoquímica y calidad del agua subterránea en la zona urbana de Zamora, Michoacán” I.Q. Claudia Alejan”).



.PERIODO: Nos muestra el periodo al que pertenecen los datos de estudio, donde todos los registros pertenecen al 2020

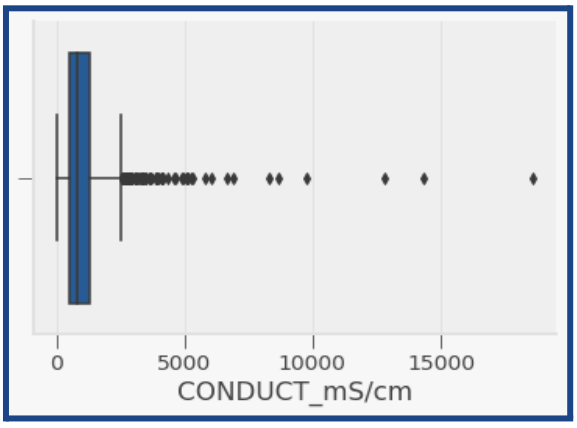
ALC_mg/L : La alcalinidad es una medida de su capacidad para neutralizar ácidos, y puede ser expresada en miliequivalente por litro (meq/L) o concentración de partes por millón (ppm o mg/L) de carbonatos, esta propiedad se debe principalmente a la presencia de sales. Con lo anterior, en esta variable numérica no encontramos diferencias significativas entre la media y mediana, decidimos utilizar la media como forma de asignar valores nulos.

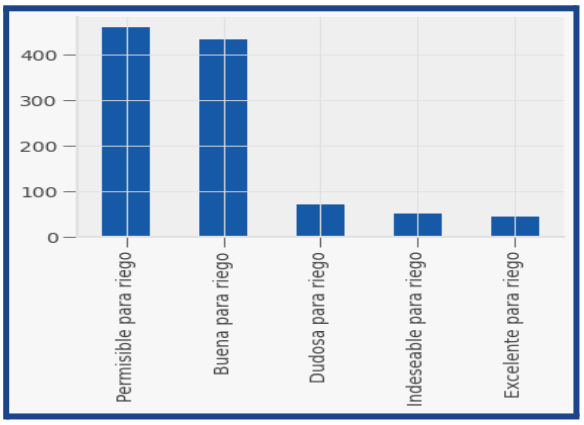




CALIDAD_ALC: Esta variable nos permite conocer si el agua tiene cierto nutrientes presentes en el PH del sustrato, un nivel alto puede deberse a altos niveles de fertilizante, cantidad, componentes del sustrato, y el cultivo, lo que es más probable que afecte a cultivos en sustrato, esto se debe a que la capacidad de amortiguamiento del mismo se agota con el paso del tiempo, por la formación de carbonatos y bicarbonatos en el sustrato ("La Alcalinidad del Agua y su Efecto en los Sustratos"). Al ser una variable categórica, asignamos missing values por la moda, donde observamos que es alta la alcalinidad.

CONDUCT_mS/cm : La conductividad hidráulica ("Ley de Darcy. Conductividad hidráulica") que permite conocer el flujo que puede tomar el agua. Esta variable numérica se observa que existen outliers que jalan la media a valores más altos que la mediana, por lo que asignamos missing values a la mediana para no afectar la distribución

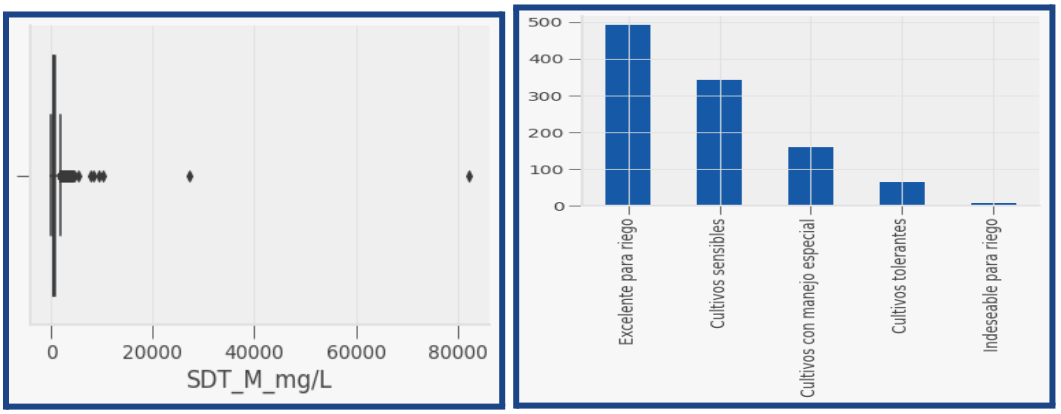




CALIDAD_CONDUC: La calidad con la que puede fluir el agua de manera subterránea, es una variable categórica, asignamos missing values por la moda.

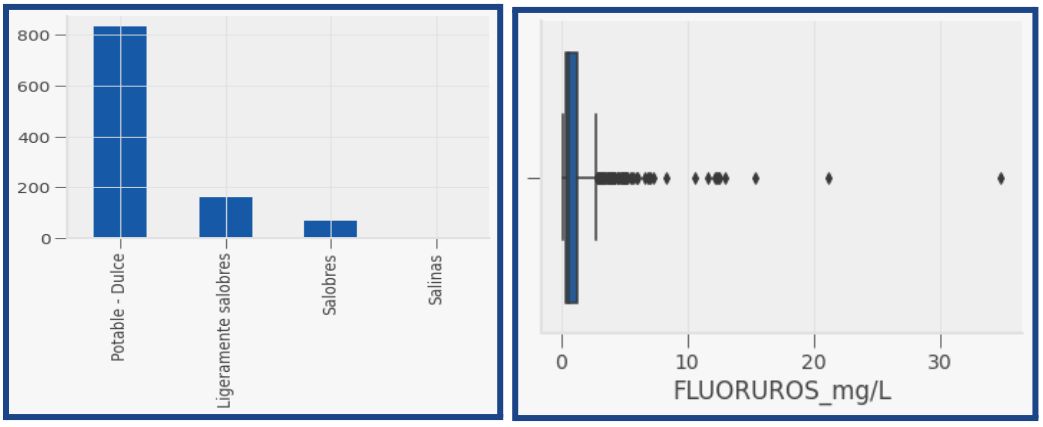
SDT_mg/L: Los sólidos disueltos totales (SDT, o TDS) son el residuo que queda después de evaporar una muestra de agua previamente filtrada a través de un elemento de fibra de vidrio ("Significado de los sólidos disueltos totales en agua (TDS)"). En esta variable todos los valores son nulos, se elimina la variable.

SDT_M_mg/L: Con lo anterior, observamos que existen outliers que jalan la media a valores más altos que la mediana.Asignamos missing values a la mediana para no afectar la distribución



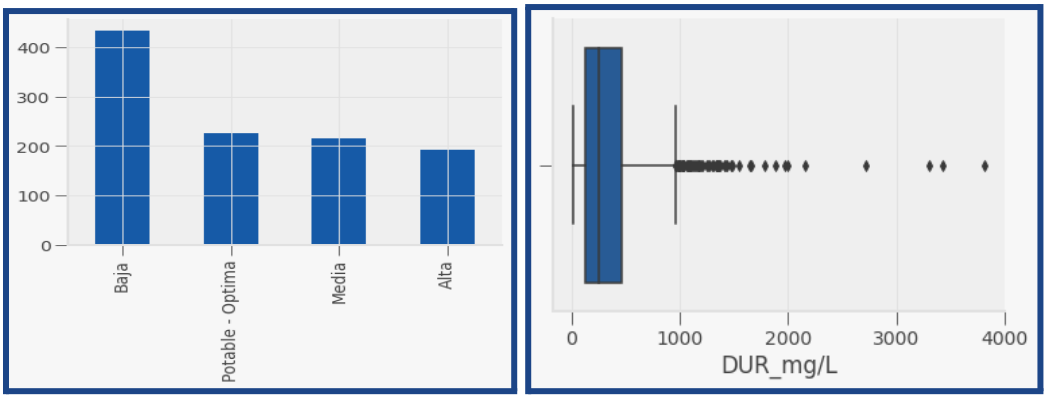
CALIDAD_SDT_ra: La mayoria de los valores muestran que al tener sólidos es un riesgo utilizarla sin procesamiento, al ser una variable categórica, asignamos missing values por la moda.

CALIDAD_SDT_salin: Nos permite conocer si es viable para consumo potable, al ser una variable categórica, asignamos missing values por la moda.



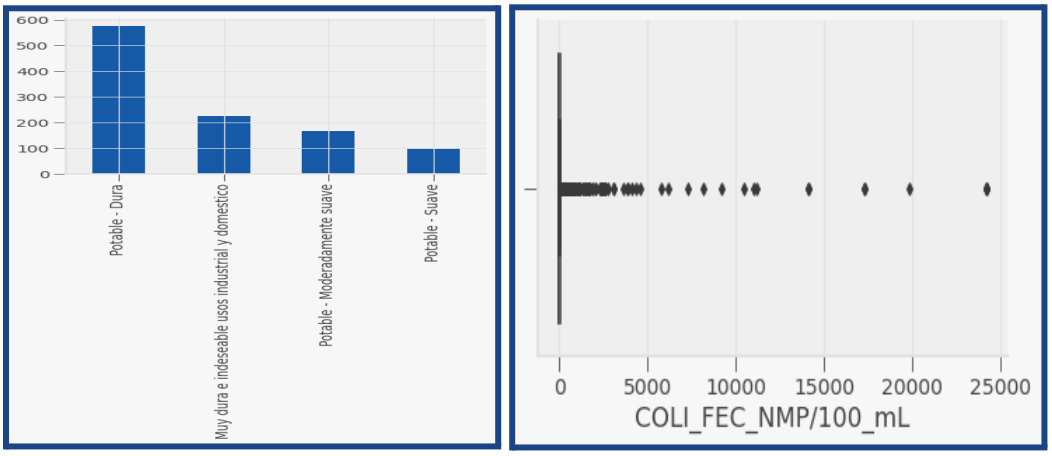
FLUORUROS_mg/L: El límite permisible de cloruros para agua potable en la NOM es de 250 mg L^{-1}, en aguas subterráneas de zonas áridas son frecuentes las concentraciones superiores a los 1000 mg L^{-1} (Davis y Wiest, 1971). Con lo anterior, observamos que existen outliers que jalan la media a valores más altos que la mediana. Asignamos missing values a la mediana para no afectar la distribución.

CALIDAD_FLUO: Al ser una variable categórica, asignamos missing values por la moda.



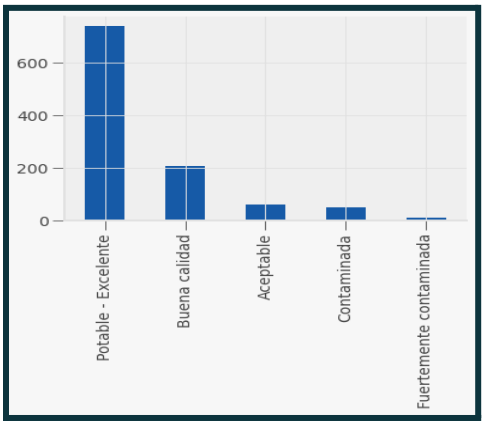
DUR_mg/L : Se denomina dureza del agua a la concentración de compuestos minerales que hay en una determinada cantidad de agua, en particular sales de magnesio y calcio (“La dureza del agua”). Con lo anterior, observamos que existen outliers que jalan la media a valores más altos que la mediana. Asignamos missing values a la mediana para no afectar la distribución

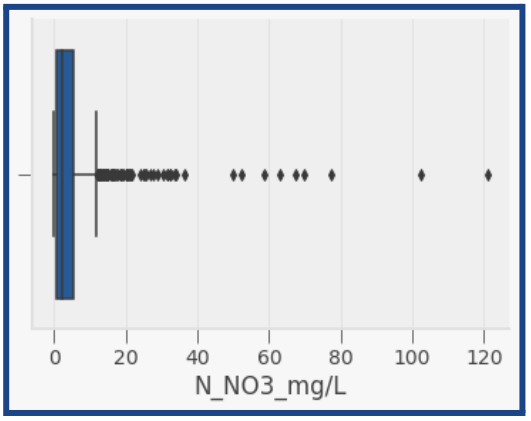
CALIDAD_DUR: Al ser una variable categórica, asignamos missing values por la moda.



COLI_FEC_NMP/100_mL : Con lo anterior, observamos que existen outliers que jalan la media a valores más altos que la mediana. Asignamos missing values a la mediana para no afectar la distribución.

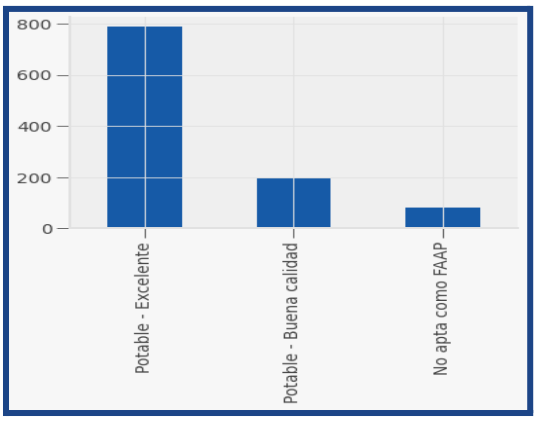
CALIDAD_COLI_FEC: Al ser una variable categórica, asignamos missing values por la moda.

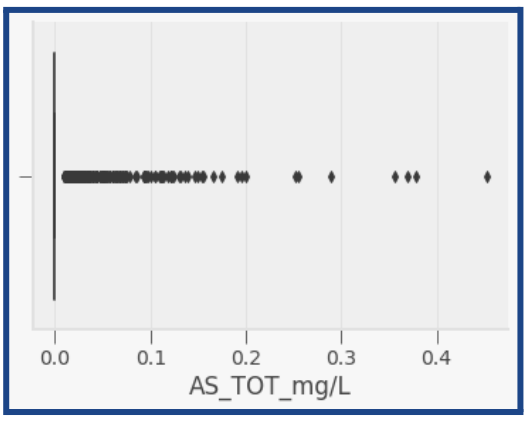




N_NO3_mg/L: Nos permite conocer la concentración de nitratos en agua, nos permite conocer la calidad del agua, pues son contaminantes móviles en el agua subterránea pues no son absorbidos por los materiales del acuífero y no precipitan como un mineral (“ANÁLISIS DE LA CONCENTRACIÓN DE NITRATOS (N-NO3 (mg/L).”). Se observan pocos valores atípicos.

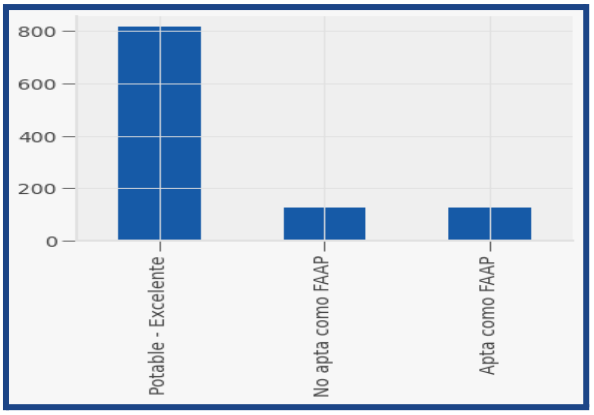
CALIDAD_N_NO3: La calidad nos permite conocer si es agua que puede ser considerada como potable.

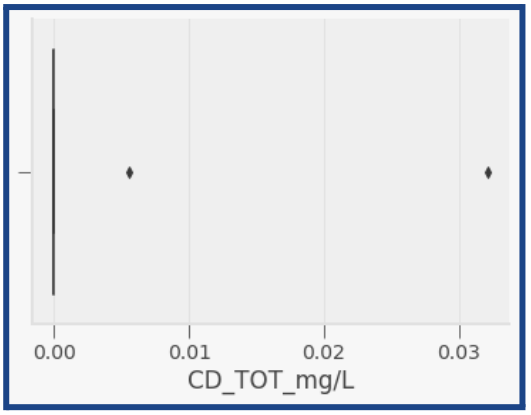




AS_TOT_mg/L: Con lo anterior, observamos que existen outliers que jalan la media a valores más altos que la mediana. Asignamos missing values a la mediana para no afectar la distribución.

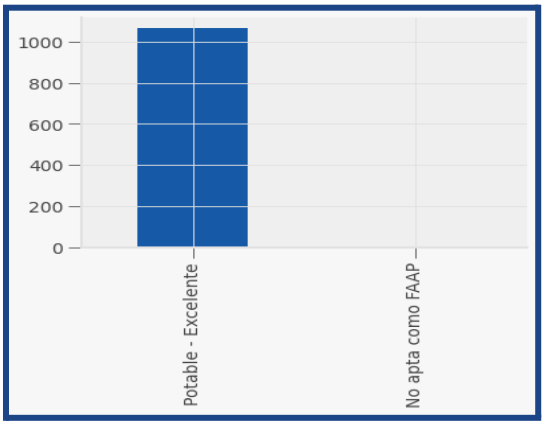
CALIDAD_AS: Al ser una variable categórica, asignamos missing values por la moda

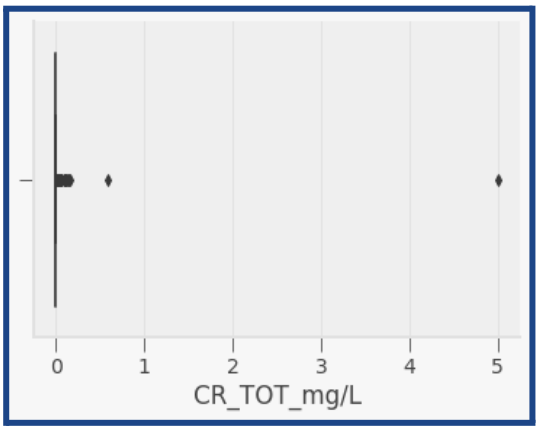




CD_TOT_mg/L: Con lo anterior, observamos que existen outliers que jalan la media a valores más altos que la mediana. Asignamos missing values a la mediana para no afectar la distribución

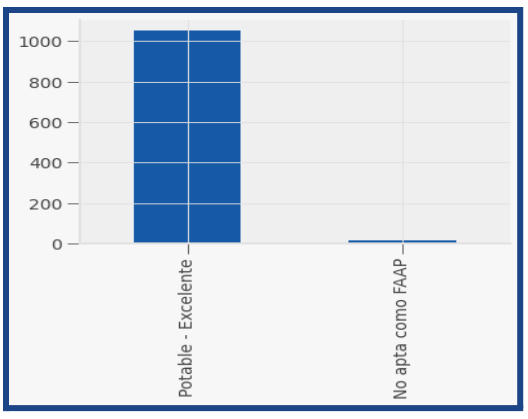
CALIDAD_CD: Al ser una variable categórica, asignamos missing values por la moda

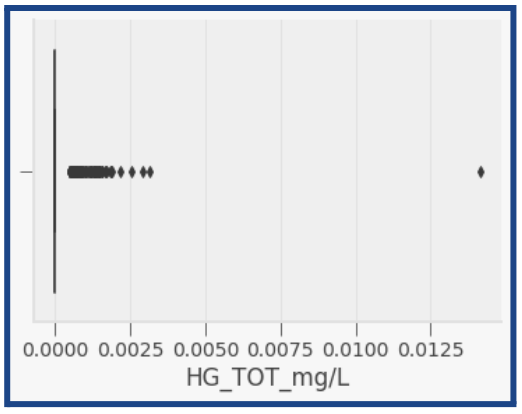




CR_TOT_mg/L: Con lo anterior, observamos que existen outliers que jalan la media a valores más altos que la mediana. Asignamos missing values a la mediana para no afectar la distribución

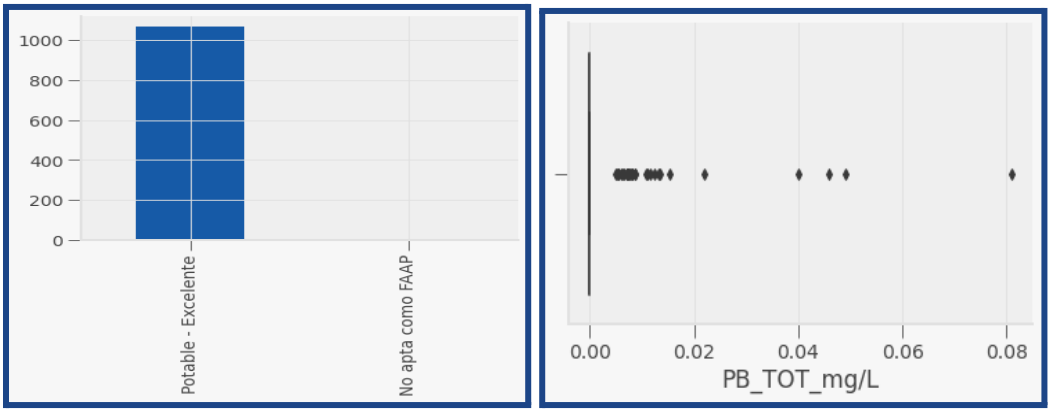
CALIDAD_CR: Al ser una variable categórica, asignamos missing values por la moda.





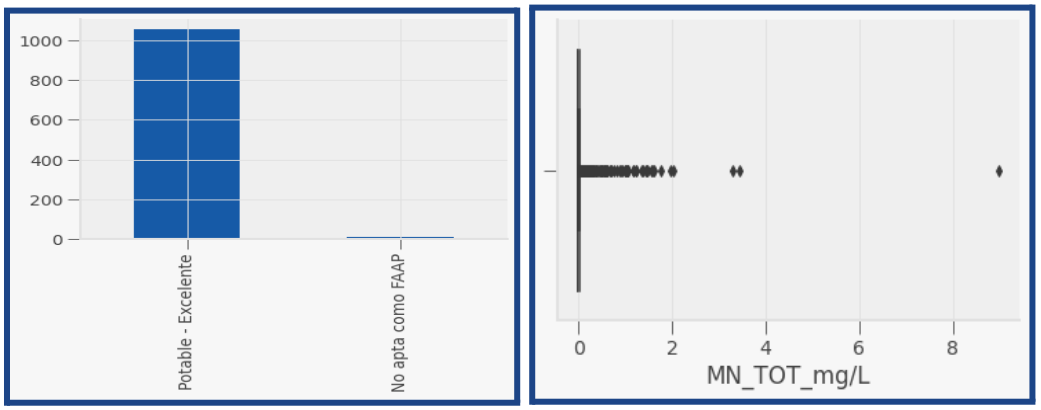
HG_TOT_mg/L: Con lo anterior, observamos que existen outliers que jalan la media a valores más altos que la mediana. Asignamos missing values a la mediana para no afectar la distribución.

CALIDAD_HG: Al ser una variable categórica, asignamos missing values por la moda.



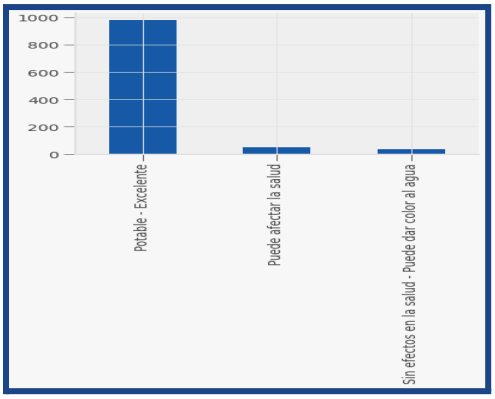
PB_TOT_mg/L : Con lo anterior, observamos que existen outliers que jalan la media a valores más altos que la mediana. Asignamos missing values a la mediana para no afectar la distribución.

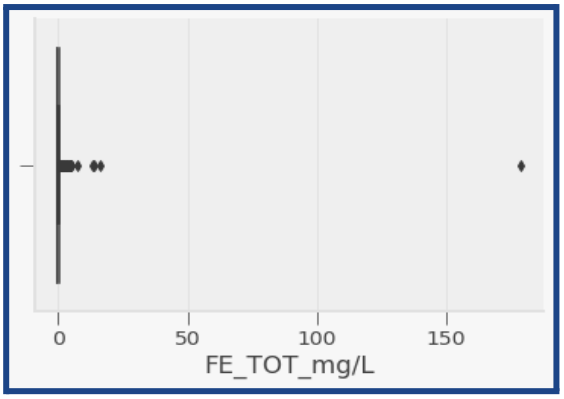
ALIDAD_PBC : Al ser una variable categórica, asignamos missing values por la moda.



MN_TOT_mg/L : Con lo anterior, observamos que existen outliers que jalan la media a valores más altos que la mediana. Asignamos missing values a la mediana para no afectar la distribución.

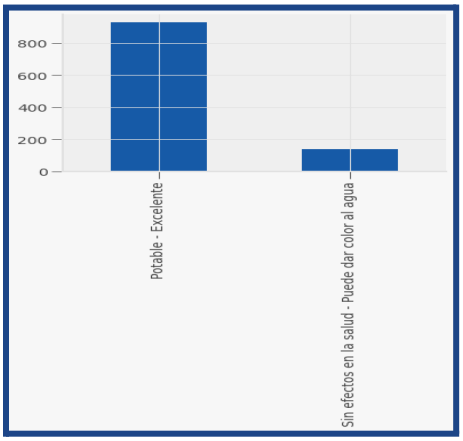
CALIDAD_MN: Al ser una variable categórica, asignamos missing values por la moda.

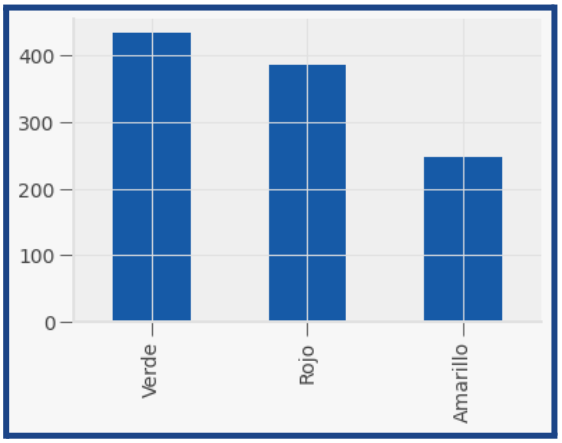




FE_TOT_mg/L: Con lo anterior, observamos que no existen diferencias significativas entre la media y la mediana. Asignamos missing values a la media.

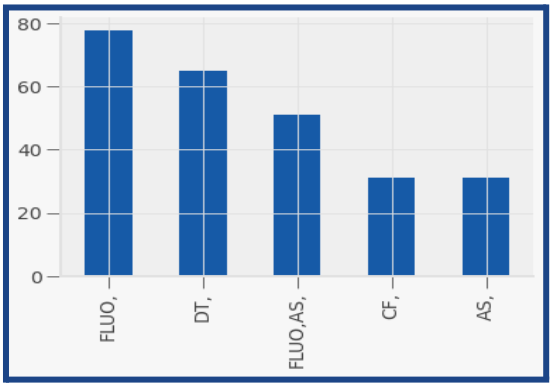
CALIDAD_FE: Al ser una variable categórica, asignamos missing values por la moda

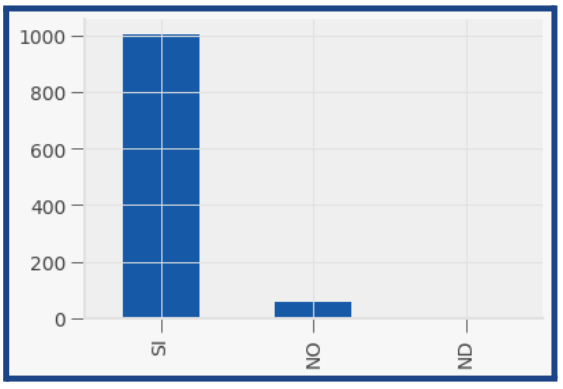




SEMÁFORO: Nos permite distinguir la clasificación de los tipos de acuíferos es decir el tipo de calidad del agua subterránea que se puede tener. Al ser una variable categórica, asignamos missing values por la moda

CONTAMINANTES: Nos permite conocer si se tienen contaminantes en el agua lo que la clasifique como no potable. Al ser una variable categórica, asignamos missing values por la moda.



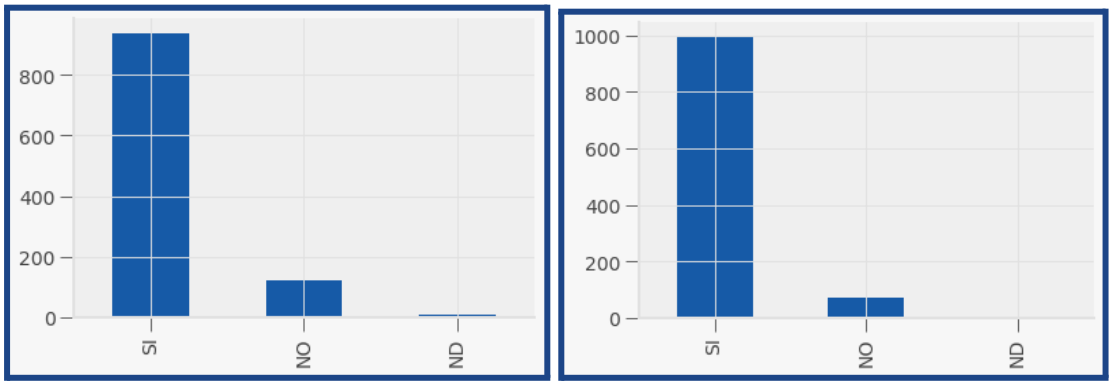


Variables binarias

Se agregan las variables binarias que permiten realizar el análisis de acuerdo a cada elemento presente en el agua y ver si cumplen.

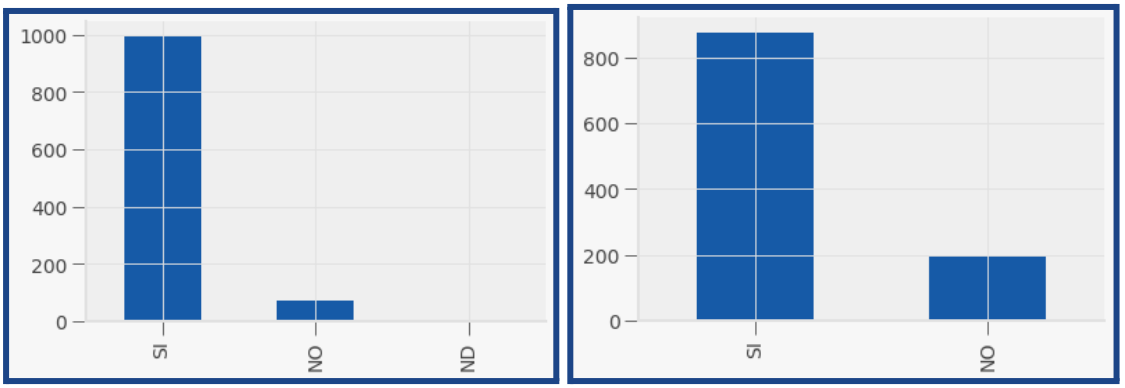
CUMPLE_CON_ALC: Al ser una variable categórica, asignamos missing values por la moda

CUMPLE_CON_COND: Al ser una variable categórica, asignamos missing values por la moda



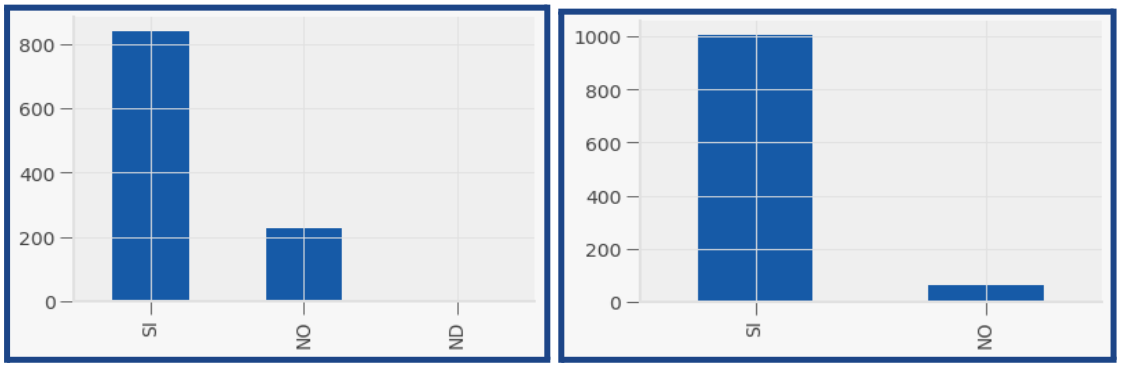
CUMPLE_CON_SDT_ra:Al ser una variable categórica, asignamos missing values por la moda.

CUMPLE_CON_SDT_salin:Al ser una variable categórica, asignamos missing values por la moda.



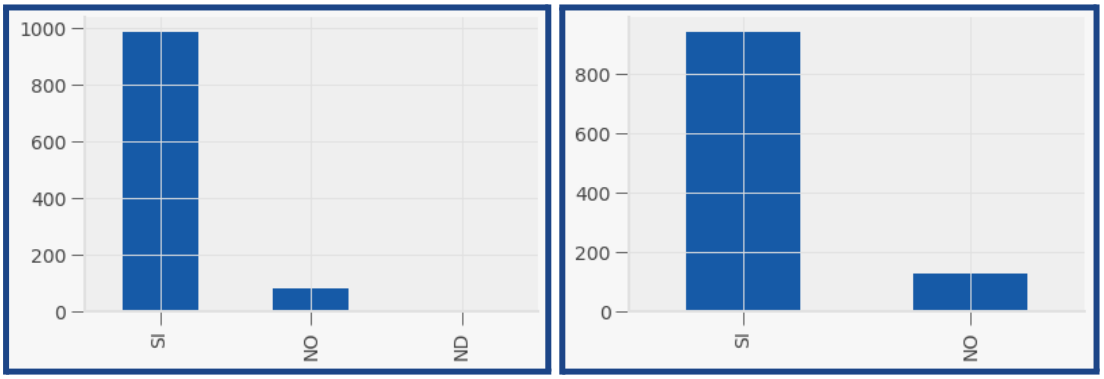
CUMPLE_CON_FLUO:Al ser una variable categórica, asignamos missing values por la moda.

CUMPLE_CON_DUR: Al ser una variable categórica, asignamos missing values por la moda.



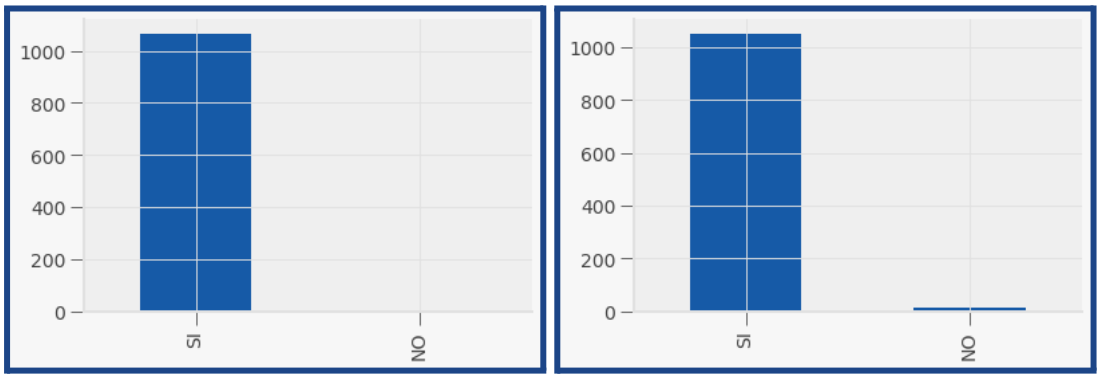
CUMPLE_CON_CF: Al ser una variable categórica, asignamos missing values por la moda.

CUMPLE_CON_NO3: Al ser una variable categórica, asignamos missing values por la moda



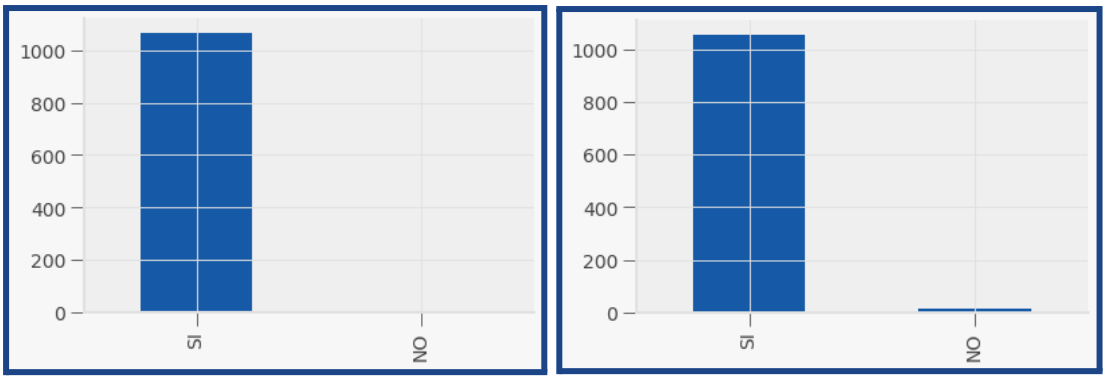
CUMPLE_CON_AS: Al ser una variable categórica, asignamos missing values por la moda

CUMPLE_CON_CD. Al ser una variable categórica, asignamos missing values por la moda



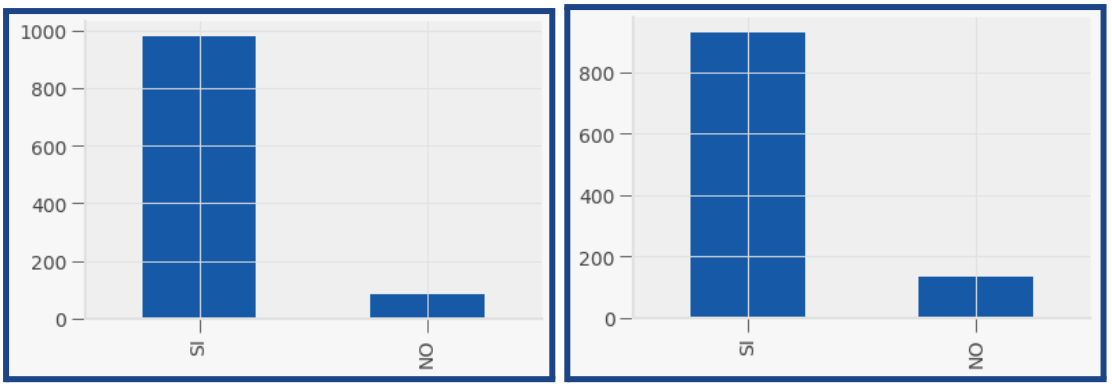
CUMPLE_CON_CR: Al ser una variable categórica, asignamos missing values por la moda

CUMPLE_CON_HG: Al ser una variable categórica, asignamos missing values por la moda



CUMPLE_CON_PB: Al ser una variable categórica, asignamos missing values por la moda

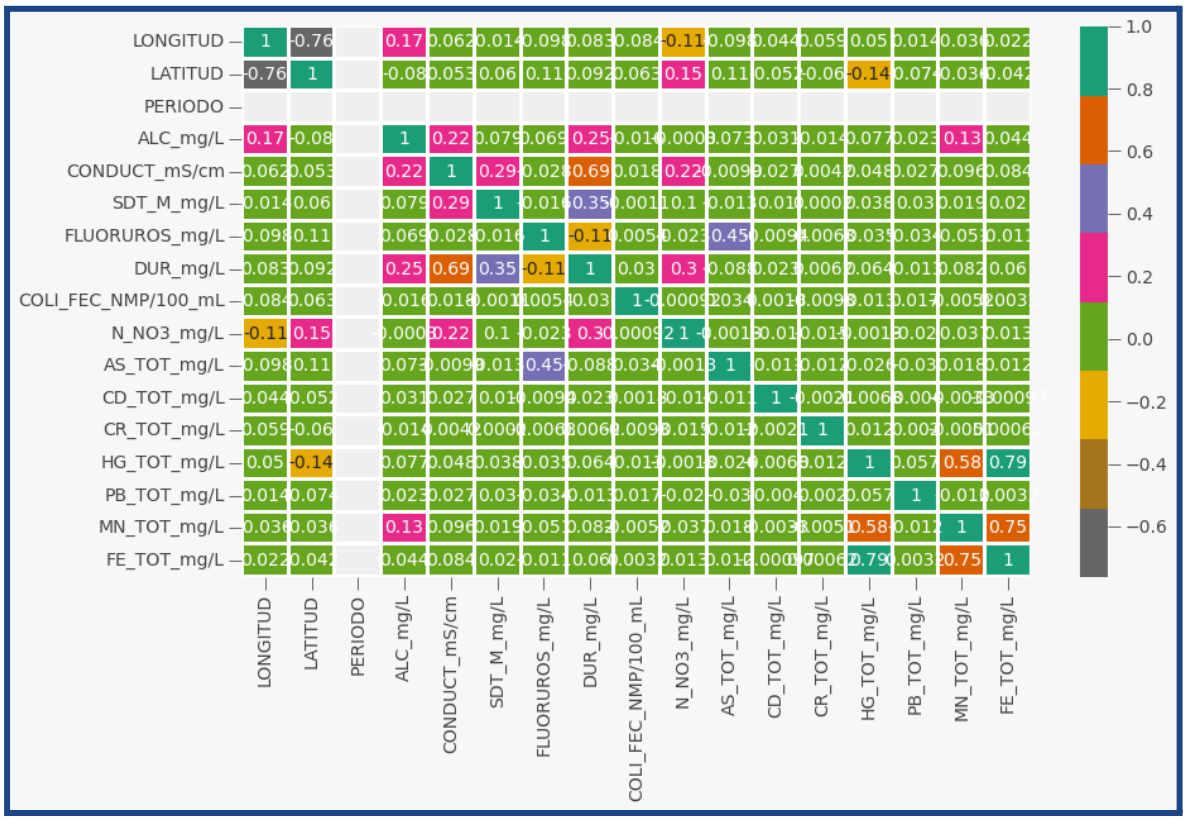
CUMPLE_CON_MN: Al ser una variable categórica, asignamos missing values por la moda



CUMPLE_CON_FE:Al ser una variable categórica, asignamos missing values por la moda

Matriz de correlaciones

Sabemos que la matriz de correlaciones aplica para variables numéricas. La correlación es un tipo de asociación entre dos variables numéricas, específicamente evalúa la tendencia creciente o decreciente que tienen los datos. Por lo que podemos decir, que dos variables están asociadas cuando una variable nos da información acerca de la otra.



Se observa una correlación alta entre la concentración de MN (manganeso) comparado con HG (mercurio), es decir la cantidad de mercurio que se encuentra en el agua es alta por lo que mide que a mayor contaminación es más probable que se encuentren esos elementos presentes en las aguas. También se observa una correlación de 0.45 entre AS(arsénico) y el flúor presente en el agua. Por lo que también la cantidad de sólidos permite conocer la dureza

del agua, pues son los minerales que concentra. Además de que a más dureza también hay una relación con la conductividad del agua. Por lo tanto, cuanto más dura es el agua de un lugar, mayor es su capacidad de conducir la corriente porque contiene más sales (“4. Dureza y conductividad del agua”).

Comprobación del Data cleaning

Observamos el DF

```
df.head()
```

	CLAVE	SITIO	ORGANISMO_DE_CUENCA	ESTADO	MUNICIPIO	ACUIFERO	SUBTIPO	LONGITUD	LATITUD	PERIODO	...	CUMPLE_CON_DUR	CUM
0	DLAGU6	POZO SAN GIL	LERMA SANTIAGO PACIFICO	AGUASCALIENTES	ASIENTOS	VALLE DE CHICALOTE	POZO	-102.02210	22.20887	2020	...	SI	
1	DLAGU6516	POZO R013 CAÑADA HONDA	LERMA SANTIAGO PACIFICO	AGUASCALIENTES	AGUASCALIENTES	VALLE DE CHICALOTE	POZO	-102.20075	21.99958	2020	...	SI	
2	DLAGU7	POZO COSIO	LERMA SANTIAGO PACIFICO	AGUASCALIENTES	COSIO	VALLE DE AGUASCALIENTES	POZO	-102.28801	22.36685	2020	...	SI	
3	DLAGU9	POZO EL SALITRILLO	LERMA SANTIAGO PACIFICO	AGUASCALIENTES	RINCON DE ROMOS	VALLE DE AGUASCALIENTES	POZO	-102.29449	22.18435	2020	...	SI	
4	DLBAJ107	RANCHO EL TECOLOTE	PENINSULA DE BAJA CALIFORNIA	BAJA CALIFORNIA SUR	LA PAZ	TODOS SANTOS	POZO	-110.24480	23.45138	2020	...	SI	

5 rows × 56 columns

Mostramos que ya no existen missing values

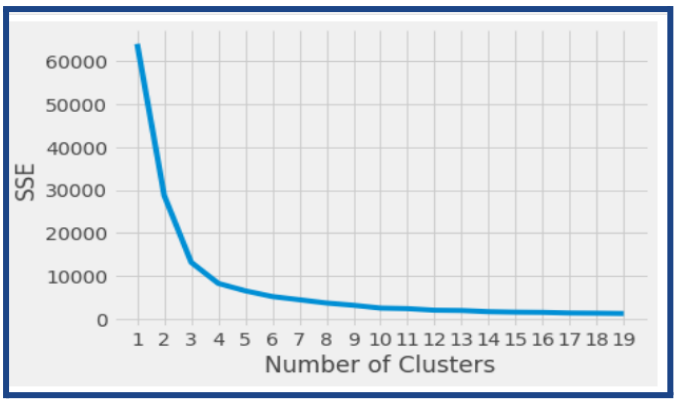
Se observa que ya no existen valores nulos lo que garantiza que es una base que cumple con:

-*Exactitud*: todos los datos deben ser precisos.
-*Coherencia:* la coherencia de los datos te permite saber si la información es la misma en diferentes bases de datos
-*Validez:* todos los datos deben cumplir con reglas o restricciones definidas.
-*Uniformidad:* es importante que todos los datos dentro de tus bases tengan los mismos valores o unidades. Este es un elemento realmente indispensable a la hora de hacer data cleaning

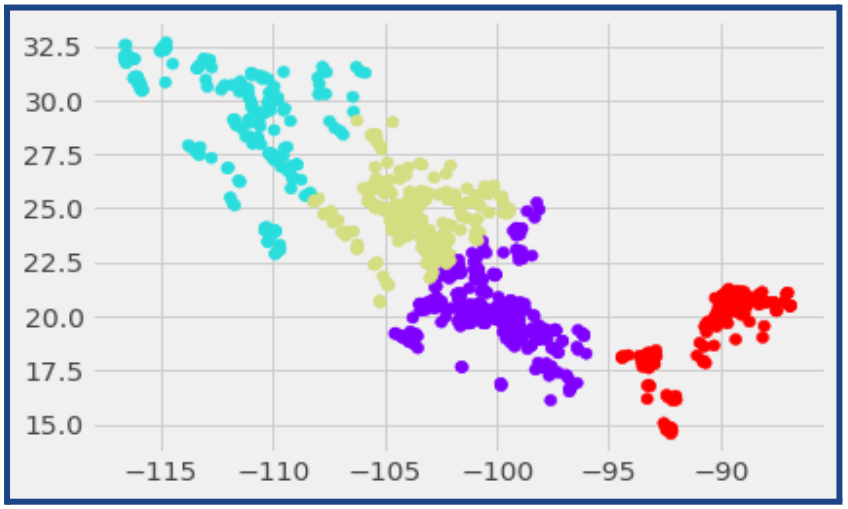
K- means para relación de calidad del agua y su ubicación

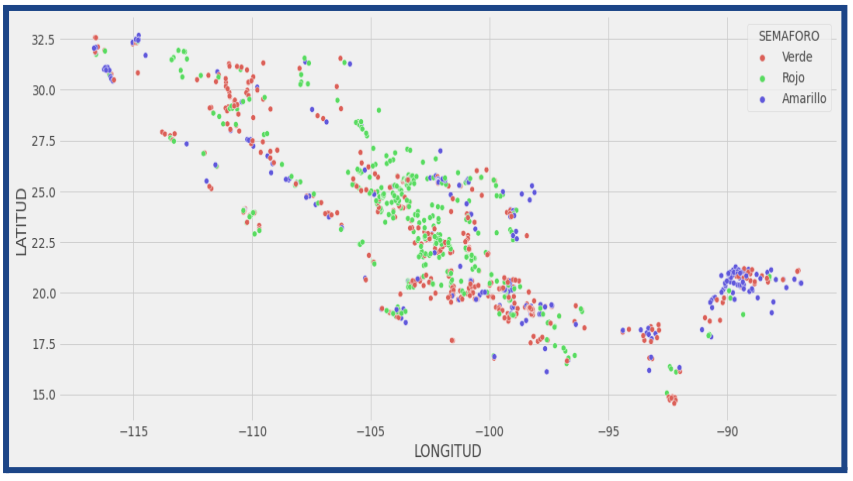
La importancia de aplicar este método,se sustenta en que el mapa de vulnerabilidad obtenido con el algoritmo K-medias en otros estudios relacionados ha presentado mayor precisión al lograr un coeficiente de correlación de Pearson igual a 0,72 entre la concentración de nitrato en las aguas subterráneas y las clases de vulnerabilidad definidas(“Classification of aquifer vulnerability using K-means cluster analysis”).

El codo de jambu (Elbow Method) permite conocer el número óptimo de clusters, de acuerdo a la distancia promedia del centroide a todos los puntos del clúster, en este caso se integraron con la variable de semáforo para medir la calidad del agua. Finalmente, dónde se muestra la disminución del error entre las diferencias de las personas dentro de un cluster conforme aumenta el número de clusters. Por lo que, se decide que el número óptimo de clusters será de 4.



La distribución de los clusters óptima es por zona geográfica es decir la región de ubicación de los acuíferos ya que eso permite que se optimice por las condiciones en las que llegó el agua a esos acuíferos.





La distribución de los registros de acuerdo al semáforo muestra que hay distintos tipos de acuíferos de acuerdo a las características que se mencionan sobre el tipo de suelo y la temperatura del clima.

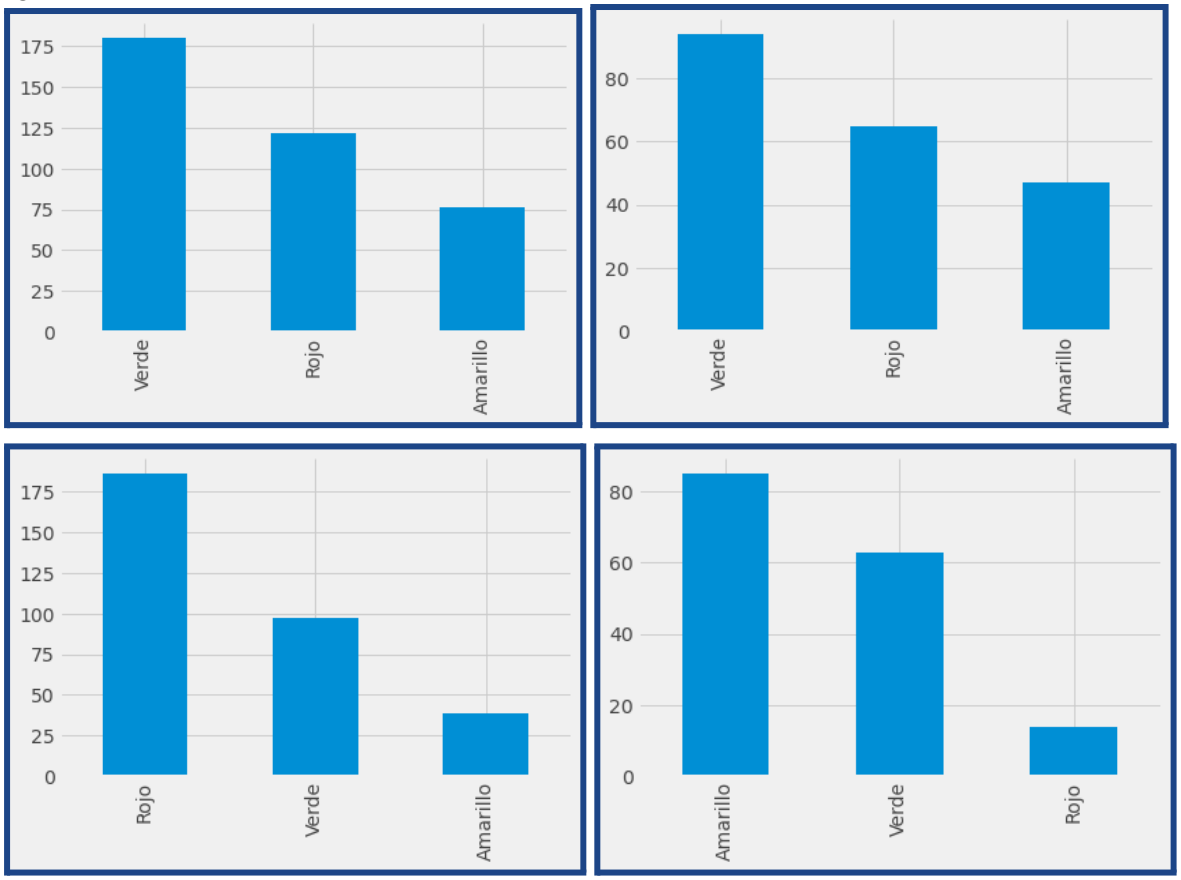
La clasificación no supervisada permite un mayor poder resolutivo para cartografiar la vulnerabilidad natural a la contaminación de los acuíferos (“Aplicación de la minería de datos

a la evaluación de la vulnerabilidad de acuíferos."), al brindar un modelo con cuatro clases de susceptibilidad a la degradación del agua subterránea.

Relación de cluster con calidad del agua

Con el análisis de agrupación de registros por k means, observamos que si existe una relación de los grupos formados con la calidad del agua.

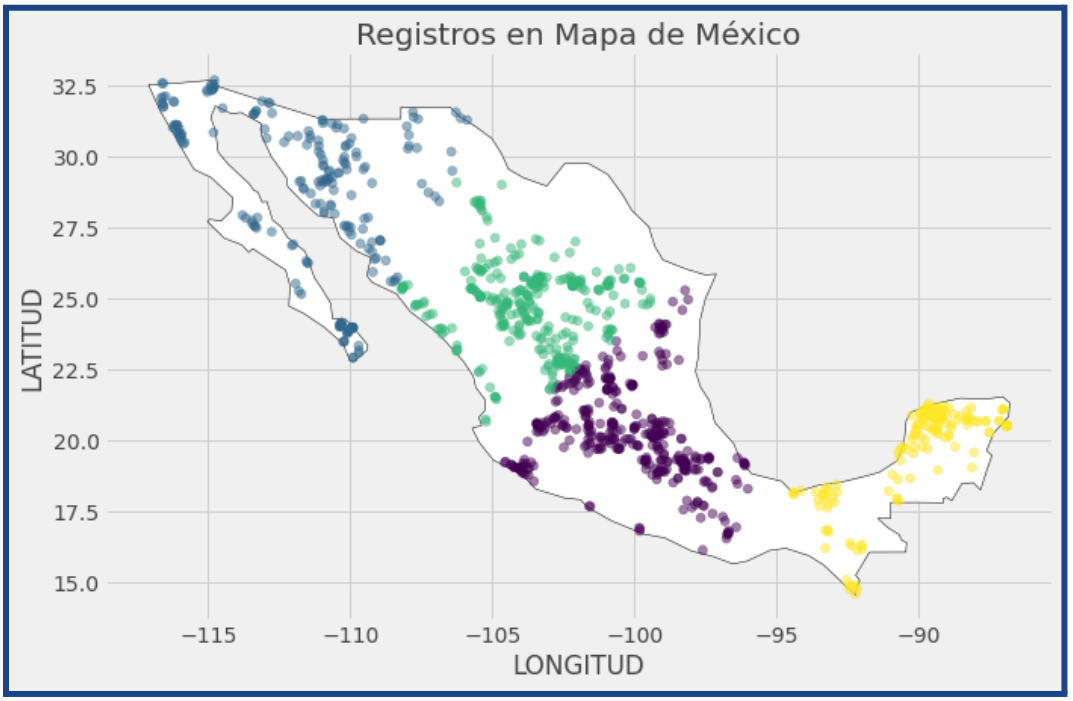
Cluster 1: Se observa que dentro del primer cluster, existen más registros con calidad de agua VERDE.
Cluster 2: Se observa que dentro del segundo cluster, existen más registros con calidad de agua VERDE.
Cluster 3: Se observa que dentro del tercer cluster, existen más registros con calidad de agua ROJO.
Cluster 4: Se observa que dentro del cuarto cluster, existen más registros con calidad de agua AMARILLO.



Estado	Amarillo	Rojo	Verde
DURANGO	4%	66%	30%
CHIHUAHUA	14%	63%	23%
ZACATECAS	4%	63%	33%
NAYARIT	0%	63%	38%
AGUASCALIENTES	7%	50%	43%
COLIMA	19%	50%	31%
COAHUILA DE ZARAGOZA	31%	49%	20%
JALISCO	12%	45%	42%
SAN LUIS POTOSI	13%	45%	43%
GUANAJUATO	15%	41%	44%
BAJA CALIFORNIA SUR	14%	41%	45%
NUEVO LEON	7%	40%	53%
OAXACA	15%	40%	45%
QUERETARO ARTEAGA	0%	33%	67%
PUEBLA	22%	30%	48%
SONORA	15%	30%	55%
SINALOA	38%	28%	34%
MORELOS	9%	27%	64%
HIDALGO	38%	24%	38%
TAMAULIPAS	52%	24%	24%
CAMPECHE	56%	20%	24%
GUERRERO	20%	20%	60%
CHIAPAS	24%	19%	57%
VERACRUZ DE IGNACIO DE LA	31%	19%	50%
MEXICO	13%	17%	71%
BAJA CALIFORNIA	55%	13%	32%
TLAXCALA	21%	8%	71%
MICHOACAN DE OCAMPO	26%	7%	67%
YUCATAN	59%	6%	35%
DISTRITO FEDERAL	50%	0%	50%
QUINTANA ROO	67%	0%	33%
TABASCO	38%	0%	62%

Observamos que los estados que se agrupan en rojo pertenecen a la misma región, con desiertos y grandes temperaturas cálidas. Sin embargo, se observa tambien una proporcion mixta en la que hay estados con gran cantidad de cenotes que tienen aguas del segmento amarillo y verde.

Mostrar resultados de agrupamiento de latitudes y longitudes con K means en el mapa de México



El mapa de clusters por región y semáforo nos otorga gran visibilidad de las condiciones de los acuíferos en el país. México tiene delimitados 653 acuíferos, de los cuales 195 (30%) están sin disponibilidad por concesión. Del total, 106 acuíferos (16%) se encuentran en condición de sobreexplotación, 31 (5%) con presencia de suelos salinos y agua salobre y 15 (2%) con intrusión marina (“Acuíferos de México: Aguas Subterráneas en México - Mapa”).

Conclusiones

Medir la vulnerabilidad del agua dulce a través de las aguas subterráneas nos acerca más a conocer sobre el mejor uso y recomendación de los recursos hídricos, pues constituyen gran parte de la reserva accesible para consumo potable. Ya que pueden ser administradas de mejor manera en el país, así como también considerar que en las cosechas el uso de fertilizantes contamina el uso como agua potable. México al ser país megadiverso cuenta con distintos tipos de suelo que favorece tener distintos tipos de acuíferos, y con ellos el tipo de cuidado del suelo. Parte importante son las microcuencas que juegan un rol importante del ciclo del agua y nos ayudan a mejorar nuestra vida.

La calidad del agua a través de los datos que se tienen, nos muestra que son datos que podemos considerar para aplicar una investigación y realizar algoritmos de predicción de la calidad del agua de acuerdo a los componentes y elementos químicos provenientes de los acuíferos. De hecho, se observó que la mayoría de las variables eran bastante consistentes, además de que se tiene buena distribución de los valores por lo que los resultados de la investigación pueden ser confiables, sobre todo por la distribución de los acuíferos que es en todo el país lo que brinda una mayor diversidad de condiciones de estudio. Al no tener el tipo de variable categórica directamente el método de k medias nos permite tener grupos de clasificación mediante los cuales ahora podemos asociar regiones con características de tipo de agua.

Los principales indicadores fisicoquímicos y microbiológicos como lo son fluoruros, coliformes fecales, nitrógeno de nitratos, arsénico total, cadmio, cromo, mercurio, plomo total, alcalinidad, conductividad, dureza del agua, sólidos disueltos totales, manganeso y Hierro. La calidad del agua se determinó para cada indicador obteniéndose en el caso de los metales, un cumplimiento mínimo de 54.1% para el caso de arsénico, y un máximo de 100.0% para mercurio. lo que garantiza que las variables a analizar se resumieron a binarias que facilitan la segmentación de las observaciones.

Al ser el acceso al agua saneada un derecho humano, se genera un aparte fundamental el estudio de la calidad del agua subterránea, además de que ayuda a prevenir el funcionamiento adecuado del ciclo del agua que genera estabilidad de los principales recursos naturales, que de acuerdo con reportes nacionales tan sólo en 2020 el 36.2% de los pozos de agua subterránea se catalogo como contaminada, parte importante de lo que observamos en los datos, pues la cantidad de segmentos rojos es en durango estado que concentra gran cantidad de pozos.

Referencias

- Dobre, C. (2019, octubre 7). *Crisp-DM: los 6 pasos del proceso de Data Mining*. Blog Smartup. https://blog.smartup.es/crisp-dm-6-pasos-proceso-data-mining/
- Sridharan, M. (2018, septiembre 25). *CRISP-DM - A framework for data mining & analysis*. Think Insights. https://thinkinsights.net/data-literacy/crisp-dm/
- *IBM Documentation*. (2021, agosto 17). Ibm.com. https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview
- Corporativa, I. (2021, abril 22). *¿Qué son las aguas subterráneas y por qué preocupa su nivel de contaminación?* Iberdrola. https://www.iberdrola.com/sostenibilidad/aguas-subterraneas
- del Agua, I. M. de T. (s/f). *Aguas subterráneas*. gob.mx. Recuperado el 15 de noviembre de 2022, de https://www.gob.mx/imta/articulos/aguas-subterraneas
- *infografía proceso de aguas subterráneas*. (s/f). Iberdrola.com. Recuperado el 15 de noviembre de 2022, de https://www.iberdrola.com/documents/20125/508732/Infografia_Aguas_Subterraneas.pdf/6f856aab-1a2d-b401-8571-cb95792357b3?t=1636354047987
- *J. Jaime Gómez: "La investigación en aguas subterráneas adolece de la invisibilidad del recurso"*. (2020, septiembre 8). iAgua. https://www.iagua.es/noticias/redaccion-iagua/j-jaime-gomez-investigacion-aguas-subterraneas-adolece-invisibilidad
- (S/f). Igme.es. Recuperado el 15 de noviembre de 2022, de https://www.igme.es/ZonaInfantil/MateDivul/guia_didactica/pdf_carteles/cartel4/CARTEL%204_4-2.pdf
- ACTUALIZACIÓN DE LA DISPONIBILIDAD MEDIA ANUAL DE AGUA EN EL ACUÍFERO LA PAZ (0324), ESTADO DE BAJA CALIFORNIA SUR,. Gob.mx. Recuperado el 15 de noviembre de 2022, de https://sigagis.conagua.gob.mx/gas1/Edos_Acuiferos_18/BajaCaliforniaSur/DR_0324.pdf
- *DOF - Diario Oficial de la Federación*. (s/f). Gob.mx. Recuperado el 16 de noviembre de 2022, de https://www.dof.gob.mx/nota_detalle_popup.php?codigo=5312870
- Capítulo 9-> Unsupervised Learning Techniques -> K means
- https://github.com/ageron/handson-ml3/blob/main/09_unsupervised_learning.ipynb
- https://realpython.com/k-means-clustering-python/
- *La Alcalinidad del Agua y su Efecto en los Sustratos*. (s/f). Intagri.com. Recuperado el 16 de noviembre de 2022, de https://www.intagri.com/articulos/agua-riego/la-alcalinidad-del-agua-y-su-efecto-en-los-sustratos
- "Ley de Darcy. Conductividad hidráulica." *Francisco Javier Sánchez San Román*, https://hidrologia.usal.es/temas/Ley_Darcy.pdf. Accessed 16 November 2022.
- ""Estudio de la hidrogeoquímica y calidad del agua subterránea en la zona urbana de Zamora, Michoacán" I.Q. Claudia Alejan." *Biblioteca Virtual UMSNH*, http://bibliotecavirtual.dgb.umich.mx:8083/xmlui/bitstream/handle/DGB_UMICH/4770/FIQ-M-2019-1783.pdf?sequence=1&isAllowed=y. Accessed 16 November 2022.
- Significado de los sólidos disueltos totales en agua (TDS)." *Carbotecnia*, 13 October 2021, https://www.carbotecnia.info/aprendizaje/quimica-del-agua/solidos-disueltos-totales-tds/. Accessed 16 November 2022.
- "ANÁLISIS DE LA CONCENTRACIÓN DE NITRATOS (N-NO3 (mg/L) EN POZOS JAPAY EN EL AGUA SUBTERRÁNEA, UBICADOS EN LA ZONA HIDROGEOLÓGICA DEL SEMICÍRCULO DE CENOTES EN EL ESTADO DE YUCATÁN. - Composiciones de ..." *ClubEnsayos.com*, https://www.clubensayos.com/Temas-Variados/AN%C3%81LISIS-DE-LA-CONCENTRACI%C3%93N-DE-NITRATOS-N-NO3-mgL/832976.html. Accessed 16 November 2022.

- Davis, S. N. and R. De Wiest. 1971. Hidrogeología. Ariel. Barcelona, España. ASIN: B00OF6KRTY.
- marzo. (2022, marzo 21). *Aguas subterráneas: nuestro recurso oculto más valioso*. The Nature Conservancy. https://www.nature.org/es-us/que-hacemos/nuestra-vision/perspectivas/aguas-subterraneas-nuestro-recurso-mas-valioso/
- "La dureza del agua." *Facsa*, 23 January 2017, https://www.facsa.com/la-dureza-del-agua/. Accessed 16 November 2022.
- "4. Dureza y conductividad del agua." *Continuemos estudiando*, 21 May 2021, https://continuemosestudiando.abc.gob.ar/contenido/recursos/4-dureza-y-conductividad-del-agua?u=60a7e6ce84c20b396d56d2a0. Accessed 16 November 2022.
- *Data Cleansing: ¿cómo hacer la limpieza de datos?* (s/f). https://www.crehana.com. Recuperado el 16 de noviembre de 2022, de https://www.crehana.com/blog/transformacion-digital/data-cleansing/
- Valcarce Ortega, Rosa María. "Aplicación de la minería de datos a la evaluación de la vulnerabilidad de acuíferos." *Redalyc*, https://www.redalyc.org/journal/3783/378367420001/html/. Accessed 16 November 2022.
- Javadi, S.; Hashemy, S. M.; Mohammadi, K.; Howard, K. W.; Neshat, A. Classification of aquifer vulnerability using K-means cluster analysis. Journal of Hydrology, 2017, (549): p. 27-37.
- Pulido, M. (2019, julio 10). *Análisis de datos mediante clustering*. SlashMobility | Soluciones mobile. https://slashmobility.com/blog/2019/07/clustering-como-obtener-agrupaciones-inherentes-en-los-datos/
- "Acuíferos de México: Aguas Subterráneas en México - Mapa." *Para todo México*, https://paratodomexico.com/geografia-de-mexico/hidrografia-de-mexico/acuiferos-de-mexico.html. Accessed 16 November 2022.
- Comisión Nacional del Agua. (s/f). *Calidad del agua en México*. gob.mx. Recuperado el 16 de noviembre de 2022, de https://www.gob.mx/conagua/articulos/calidad-del-agua