Tecnológico de Monterrey

Ciencia Y Analítica De Datos (Gpo 10)

Trimestral Sep-Dic 2022

Profesor Jobish Vallikavungal Devassia

Reto Entrega 1 (16/11): Limpieza, análisis, visualización y K-means

Ramón Ariel Iván Muñoz Corona **A01330566**

Fecha: 15/11/2022

Ramon Muñoz A01330566

Datos de Calidad del Agua 2020

Veamos las primeras 5 filas del dataset

	CLAVE	SITIO	ORGANISMO_DE_CUENCA	ESTADO	MUNICIPIO	ACUIFERO	SUBTIPO	LONGITU
0	DLAGU6	POZO SAN GIL	LERMA SANTIAGO PACIFICO	AGUASCALIENTES	ASIENTOS	VALLE DE CHICALOTE	POZO	-102.022
1	DLAGU6516	POZO R013 CAÑADA HONDA	LERMA SANTIAGO PACIFICO	AGUASCALIENTES	AGUASCALIENTES	VALLE DE CHICALOTE	POZO	-102.200
2	DLAGU7	POZO COSIO	LERMA SANTIAGO PACIFICO	AGUASCALIENTES	COSIO	VALLE DE AGUASCALIENTES	POZO	-102.288
3	DLAGU9	POZO EL SALITRILLO	LERMA SANTIAGO PACIFICO	AGUASCALIENTES	RINCON DE ROMOS	VALLE DE AGUASCALIENTES	POZO	-102.294
4	DLBAJ107	RANCHO EL TECOLOTE	PENINSULA DE BAJA CALIFORNIA	BAJA CALIFORNIA SUR	LA PAZ	TODOS SANTOS	POZO	-110.244

Entendamos el Shape del Dataset

```
(1068, 57)
```

Son 1068 Filas y 57 columnas.

Ahora analicemos los tipos de datos de las columnas y la cantidad de datos no nulos que presenta

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1068 entries, 0 to 1067
Data columns (total 57 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   CLAVE                 1068 non-null   object 
 1   SITIO                 1068 non-null   object 
 2   ORGANISMO_DE_CUENCA   1068 non-null   object 
 3   ESTADO                1068 non-null   object 
 4   MUNICIPIO             1068 non-null   object 
 5   ACUIFERO              1068 non-null   object 
 6   SUBTIPO               1068 non-null   object 
 7   LONGITUD              1068 non-null   float64
 8   LATITUD               1068 non-null   float64
 9   PERIODO               1068 non-null   int64  
 10  ALC_mg/L              1064 non-null   float64
 11  CALIDAD_ALC           1064 non-null   object 
 12  CONDUCT_mS/cm         1062 non-null   float64
 13  CALIDAD_CONDUC        1062 non-null   object 
 14  SDT_mg/L              0 non-null      float64
 15  SDT_M_mg/L            1066 non-null   object 
 16  CALIDAD_SDT_ra        1066 non-null   object 
 17  CALIDAD_SDT_salin     1066 non-null   object 
 18  FLUORUROS_mg/L        1068 non-null   object 
```

```
 19  CALIDAD_FLUO           1068 non-null   object
 20  DUR_mg/L               1067 non-null   object
 21  CALIDAD_DUR            1067 non-null   object
 22  COLI_FEC_NMP/100_mL    1068 non-null   object
 23  CALIDAD_COLI_FEC       1068 non-null   object
 24  N_NO3_mg/L             1067 non-null   object
 25  CALIDAD_N_NO3          1067 non-null   object
 26  AS_TOT_mg/L            1068 non-null   object
 27  CALIDAD_AS             1068 non-null   object
 28  CD_TOT_mg/L            1068 non-null   object
 29  CALIDAD_CD             1068 non-null   object
 30  CR_TOT_mg/L            1068 non-null   object
 31  CALIDAD_CR             1068 non-null   object
 32  HG_TOT_mg/L            1068 non-null   object
 33  CALIDAD_HG             1068 non-null   object
 34  PB_TOT_mg/L            1068 non-null   object
 35  CALIDAD_PB             1068 non-null   object
 36  MN_TOT_mg/L            1068 non-null   object
 37  CALIDAD_MN             1068 non-null   object
 38  FE_TOT_mg/L            1068 non-null   object
 39  CALIDAD_FE             1068 non-null   object
 40  SEMAFORO               1068 non-null   object
 41  CONTAMINANTES          634 non-null    object
 42  CUMPLE_CON_ALC         1068 non-null   object
 43  CUMPLE_CON_COND        1068 non-null   object
 44  CUMPLE_CON_SDT_ra      1068 non-null   object
 45  CUMPLE_CON_SDT_salin   1068 non-null   object
 46  CUMPLE_CON_FLUO        1068 non-null   object
 47  CUMPLE_CON_DUR         1068 non-null   object
 48  CUMPLE_CON_CF          1068 non-null   object
 49  CUMPLE_CON_NO3         1068 non-null   object
 50  CUMPLE_CON_AS          1068 non-null   object
 51  CUMPLE_CON_CD          1068 non-null   object
 52  CUMPLE_CON_CR          1068 non-null   object
 53  CUMPLE_CON_HG          1068 non-null   object
 54  CUMPLE_CON_PB          1068 non-null   object
 55  CUMPLE_CON_MN          1068 non-null   object
 56  CUMPLE_CON_FE          1068 non-null   object
dtypes: float64(5), int64(1), object(51)
memory usage: 475.7+ KB
```

De aquí podemos ver que hay una columna vacía; "SDT_mg/L". Ya que no existe nada de información de ella, vamos a eliminarla.

Podemos observar que existen distintos valores faltantes en varias columnas y que a todas se les asigno un data type de objeto cuando realmente varian de tipo de dato.

Primero dividiremos las 57 columnas en categoricas, numericas, ordinales y binarias, para posteriormente poder hacer un analisis un poco mas profundo de los valores de las variables.

De lo que se alcanza a ver, dentro de las columnas numericas hay valores en formato de texto especificando que son menores a un valor. Esto hace que nuestras variables no puedan ser tomadas como numericas, por lo tanto, se sustituiran por el valor minimo que se establecen.

Una vez convertidas todos los valores a numericos, procederemos a convertir las columnas a float.

Hecho esto, podemos empezar con un analisis estadistico de las columnas numericas

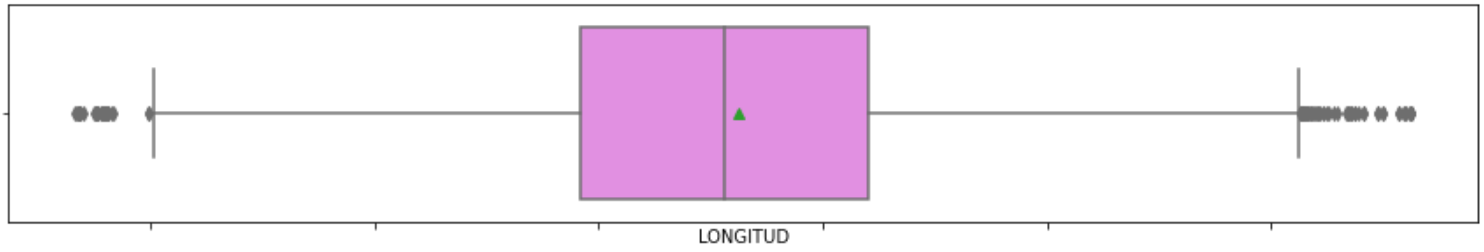
	count	mean	std	min	25%	50%	75%	max
LONGITUD	1068.0	-101.891007	6.703263	-116.66425	-105.388865	-102.174180	-98.974716	-86.864120
LATITUD	1068.0	23.163618	3.887670	14.56115	20.212055	22.617190	25.510285	32.677713
PERIODO	1068.0	2020.000000	0.000000	2020.00000	2020.000000	2020.000000	2020.000000	2020.000000
ALC_mg/L	1064.0	235.633759	116.874291	26.64000	164.000000	215.527500	292.710000	1650.000000
CONDUCT_mS/cm	1062.0	1138.953013	1245.563674	50.40000	501.750000	815.000000	1322.750000	18577.000000
SDT_M_mg/L	1066.0	896.101567	2751.530590	25.00000	337.500000	550.400000	916.100000	82170.000000
FLUORUROS_mg/L	1068.0	1.075600	1.924278	0.20000	0.267175	0.503500	1.139850	34.803300
DUR_mg/L	1067.0	347.938073	359.669452	20.00000	121.194800	245.335800	453.930000	3810.692200
COLI_FEC_NMP/100_mL	1068.0	355.490356	2052.457014	1.10000	1.100000	1.100000	13.250000	24196.000000
N_NO3_mg/L	1067.0	4.319759	8.345134	0.02000	0.650294	2.080932	5.201698	121.007813
AS_TOT_mg/L	1068.0	0.019618	0.035209	0.01000	0.010000	0.010000	0.010000	0.452200
CD_TOT_mg/L	1068.0	0.003030	0.000894	0.00300	0.003000	0.003000	0.003000	0.032110
CR_TOT_mg/L	1068.0	0.013276	0.154391	0.00500	0.005000	0.005000	0.005000	5.003200
HG_TOT_mg/L	1068.0	0.000557	0.000467	0.00050	0.000500	0.000500	0.000500	0.014150
PB_TOT_mg/L	1068.0	0.005282	0.003254	0.00500	0.005000	0.005000	0.005000	0.080900
MN_TOT_mg/L	1068.0	0.072478	0.376512	0.00150	0.001500	0.001500	0.009947	8.982000
FE_TOT_mg/L	1068.0	0.410387	5.537974	0.02500	0.025000	0.046960	0.173380	178.615000

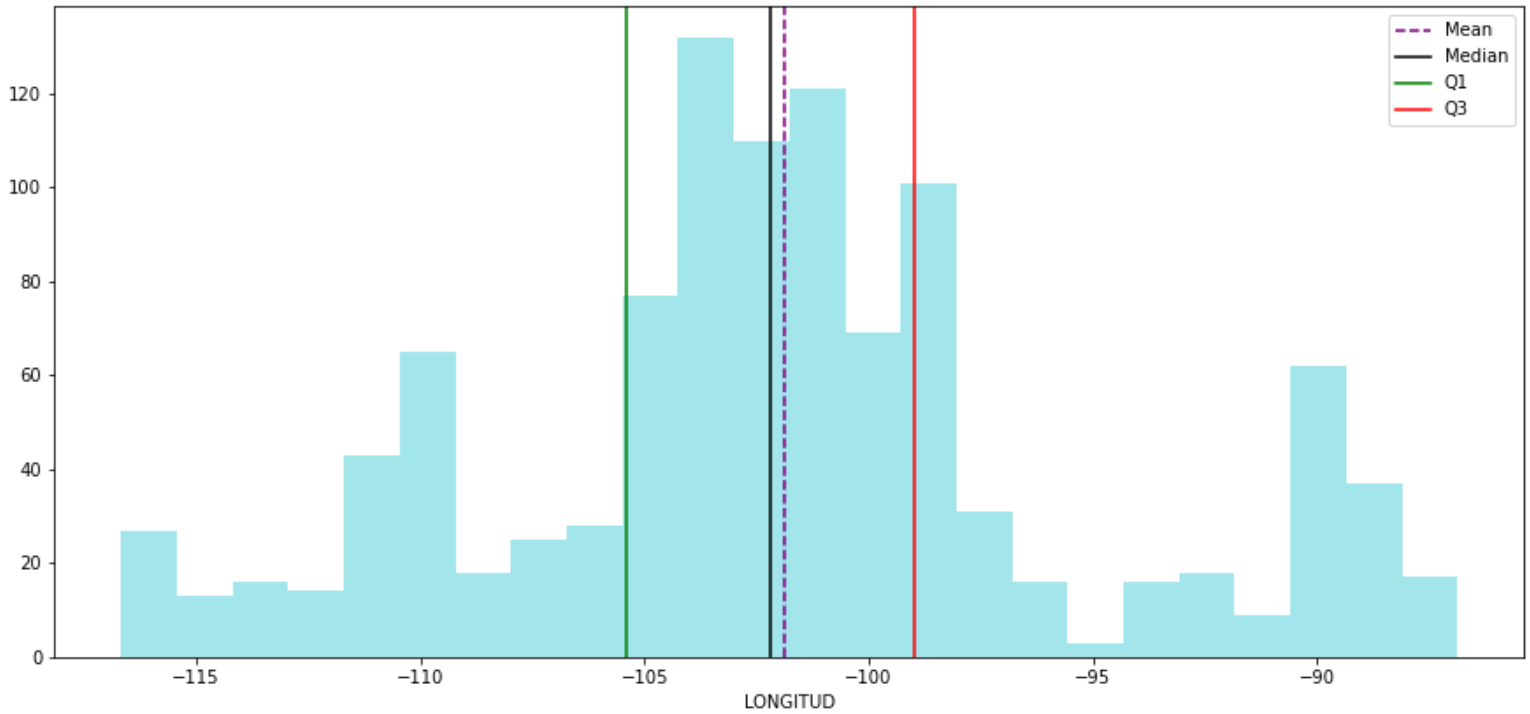
Podemos observar que el valor del Periodo es el mismo en todos los casos, asi que podemos eliminar esa columna.

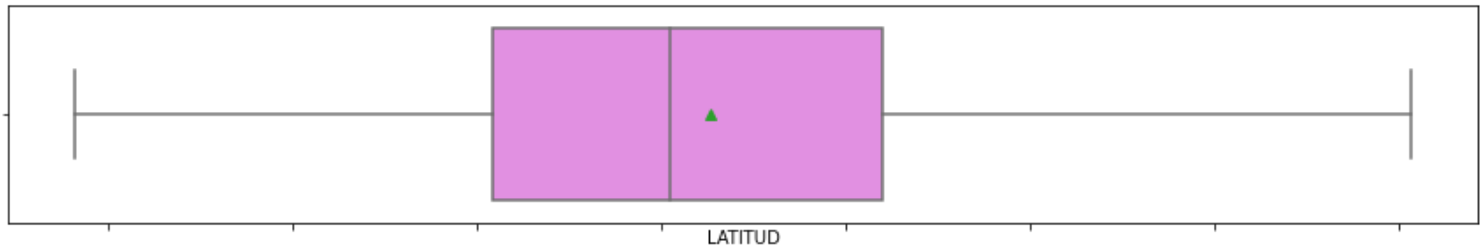
```
'PERIODO'
```

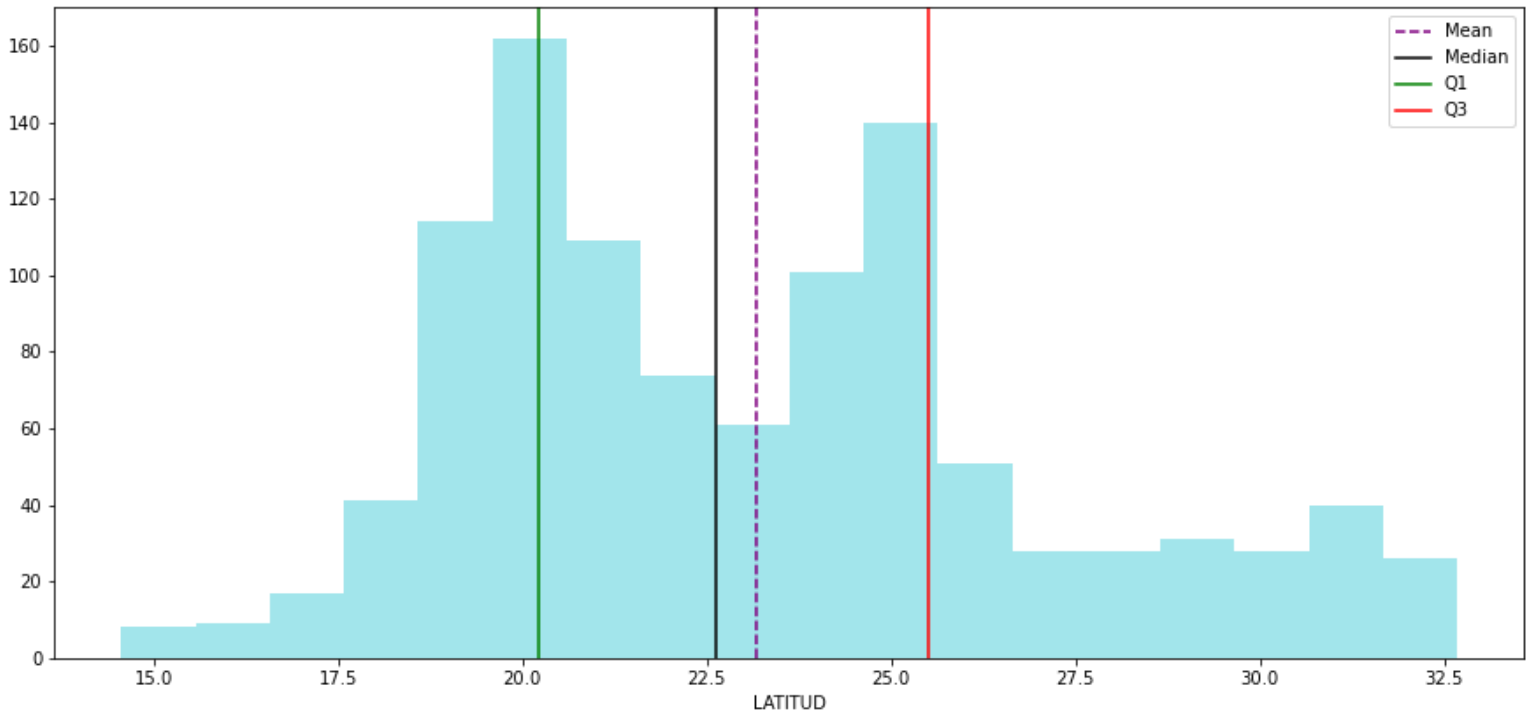
Además, podemos ver que todo se encuentra ene un rango normal, por lo que solo deberemos preocuparnos por los valores faltantes de cada columna.

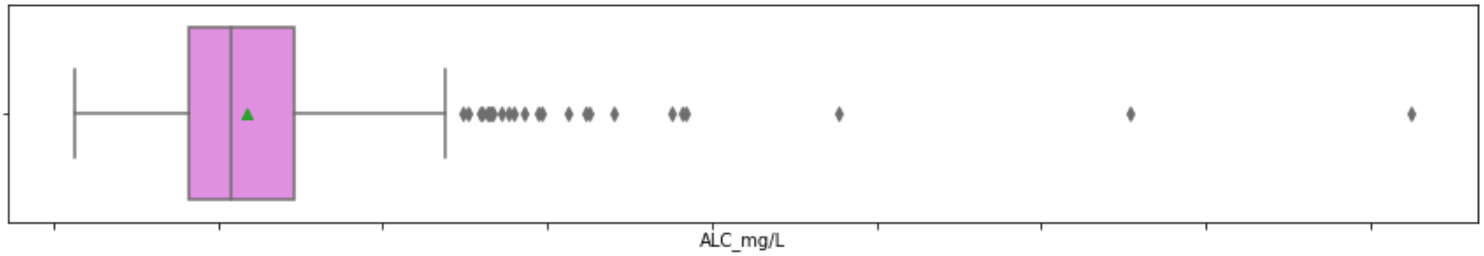
En las siguientes graficas mostraremos tanto los boxplot para mostrar los outliers, como la distribucion de cada columna

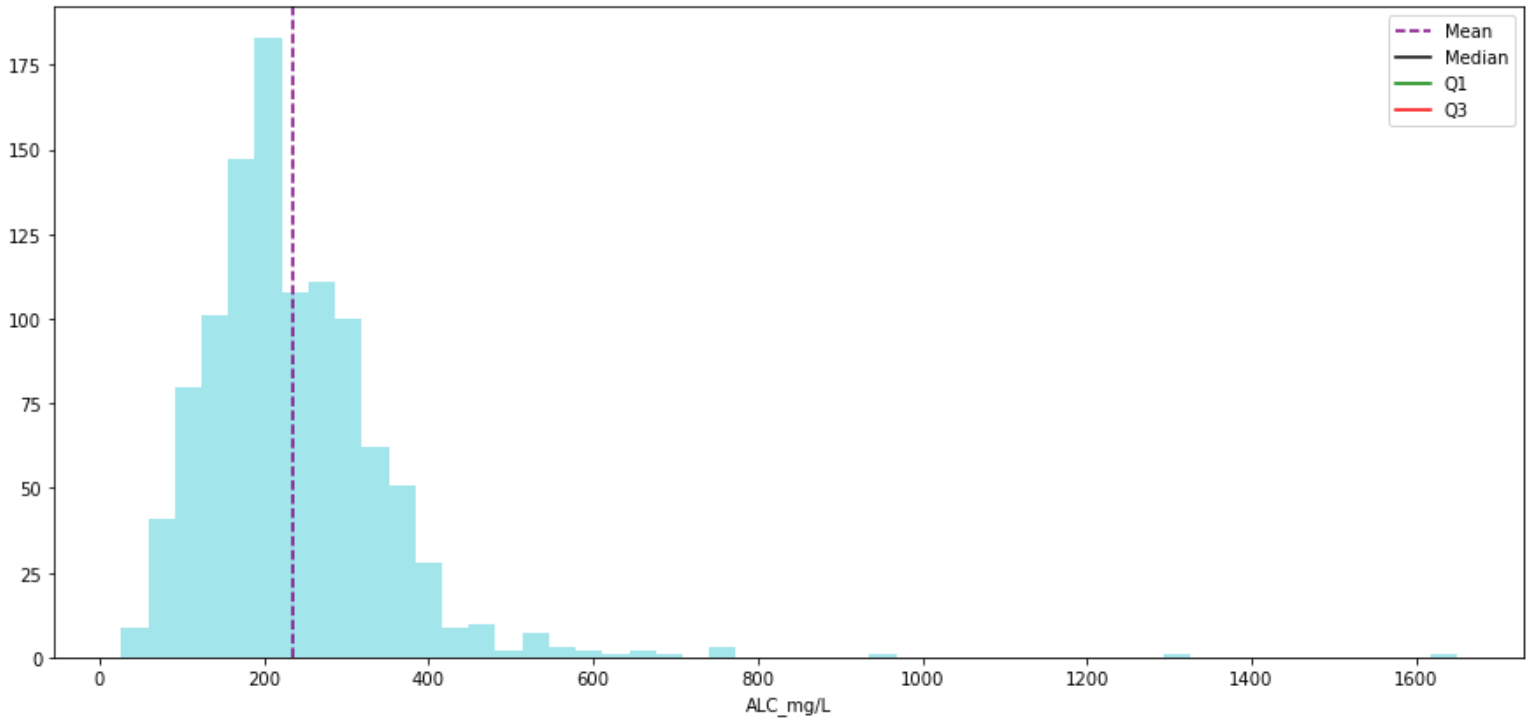

LONGITUD

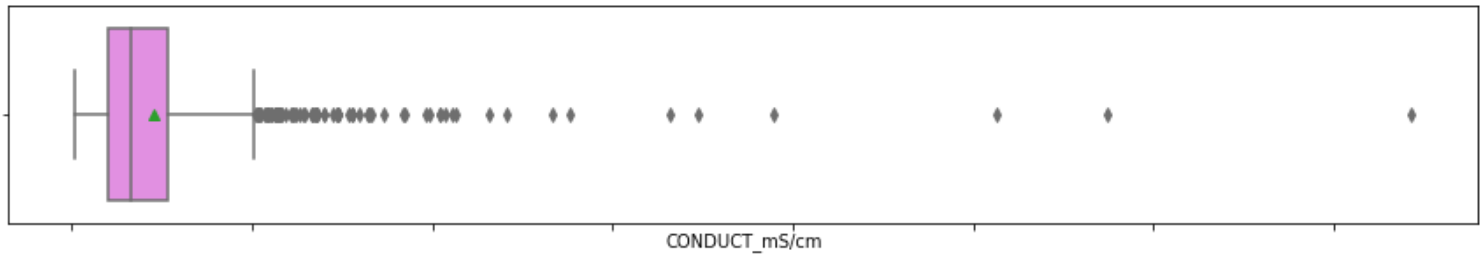

Mean
Median
Q1
Q3
120
100
80
60
40
20
0
−115
−110
−105
−100
−95
−90
LONGITUD

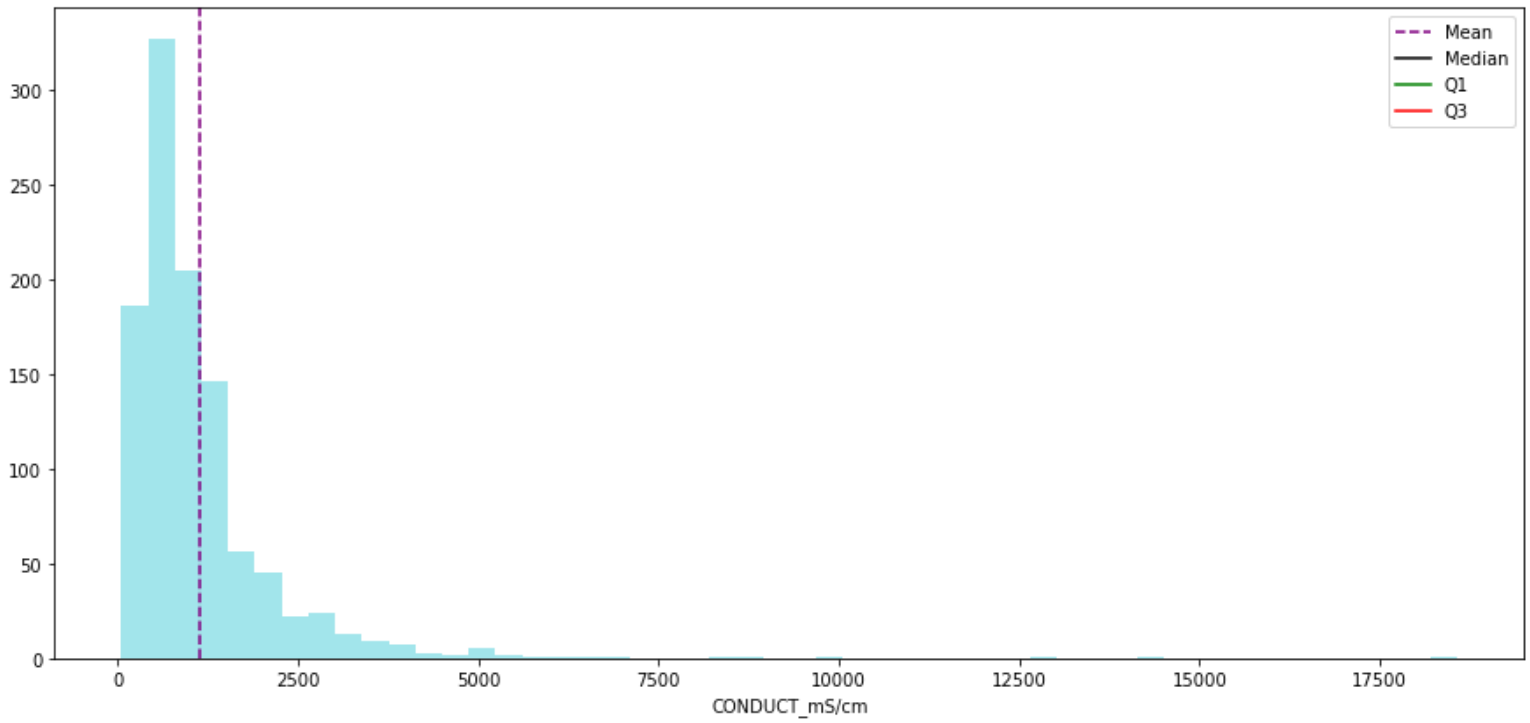

LATITUD

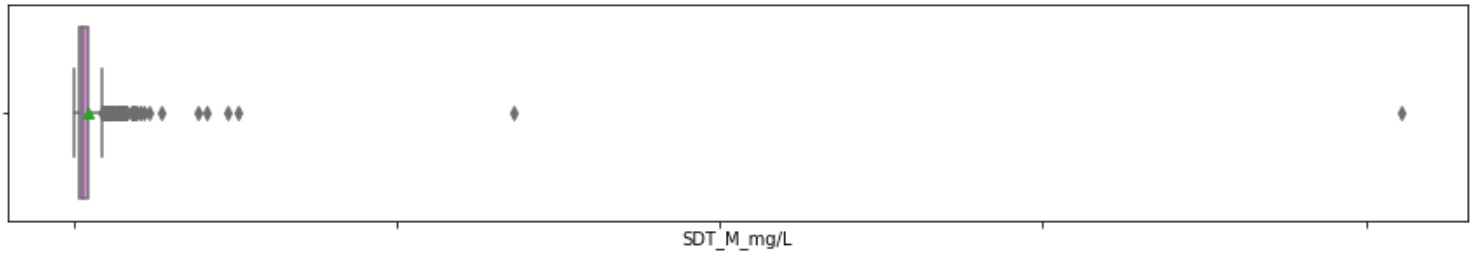

Mean
Median
Q1
Q3
160
140
120
100
80
60
40
20
0
15.0
17.5
20.0
22.5
25.0
27.5
30.0
32.5
LATITUD

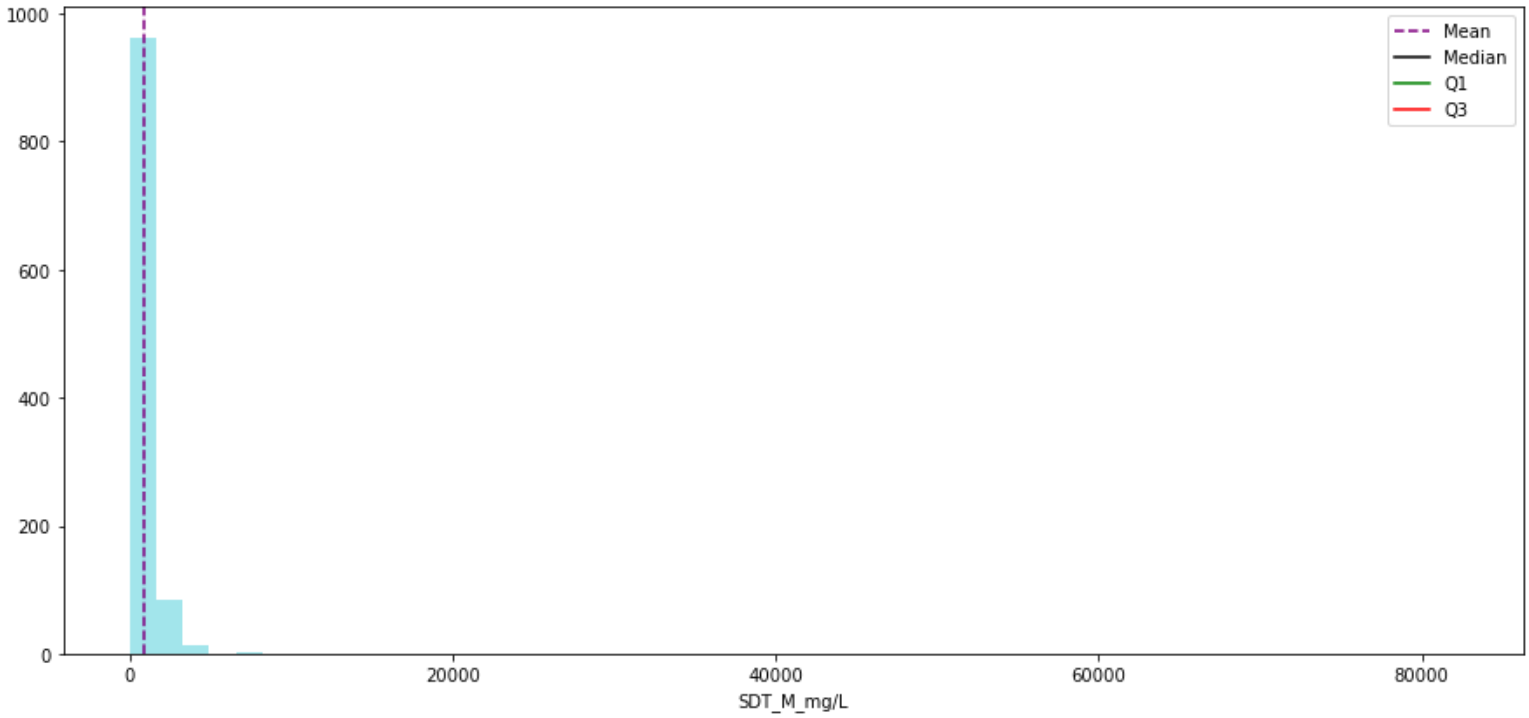

ALC_mg/L

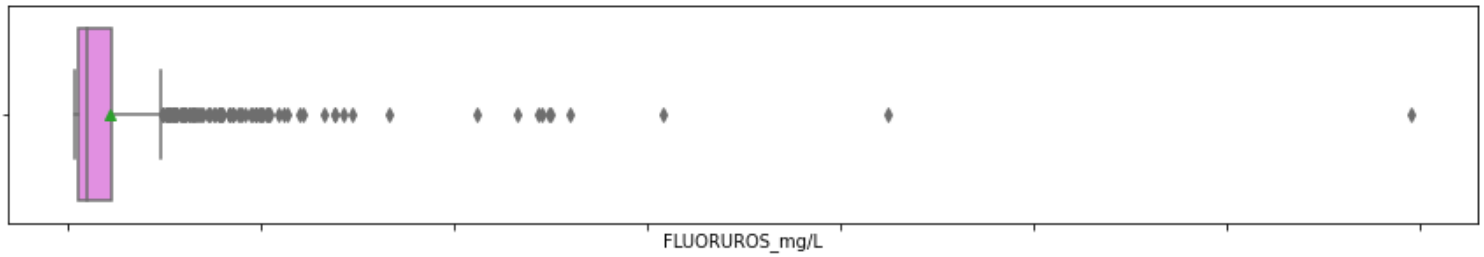

Mean
Median
Q1
Q3
0
25
50
75
100
125
150
175
0
200
400
600
800
1000
1200
1400
1600
ALC_mg/L

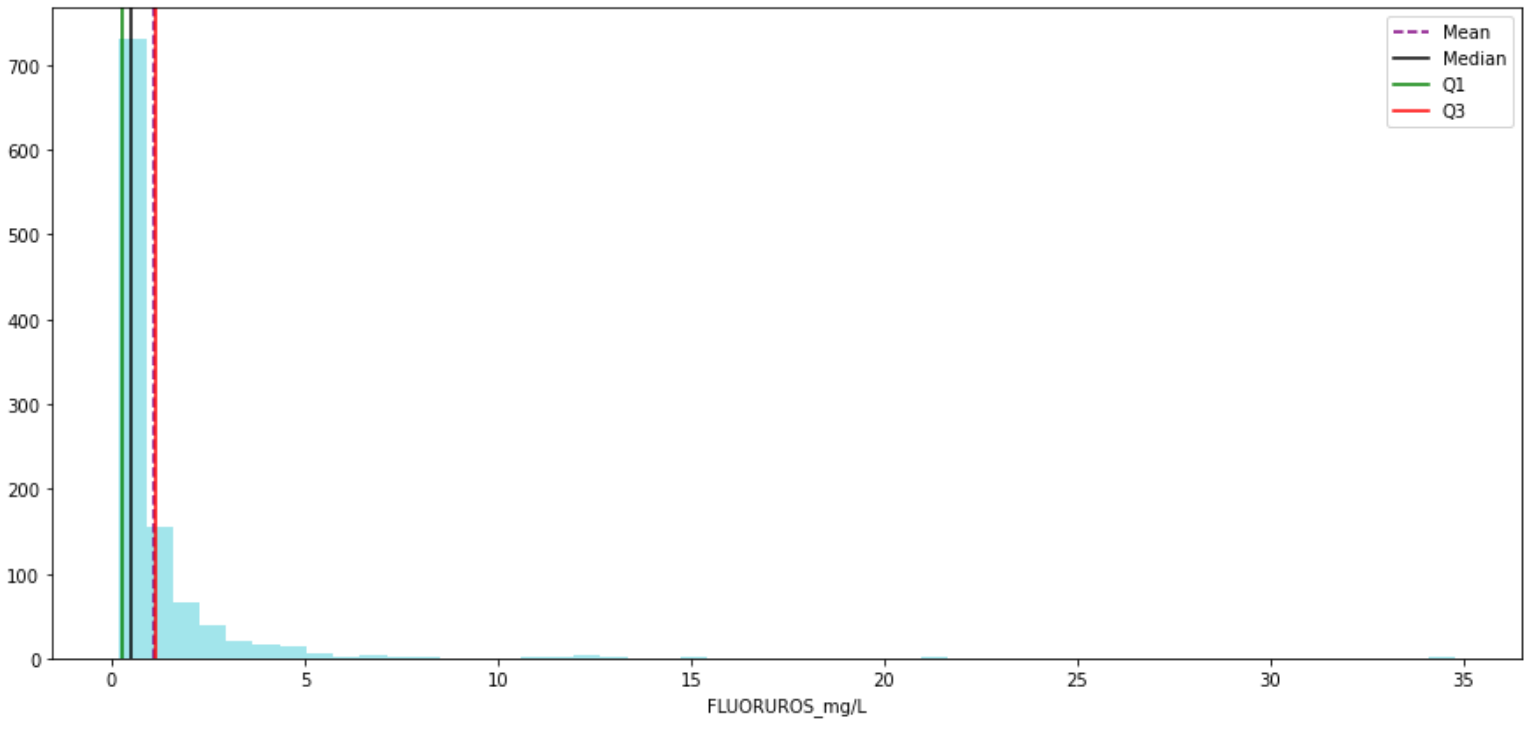

CONDUCT_mS/cm

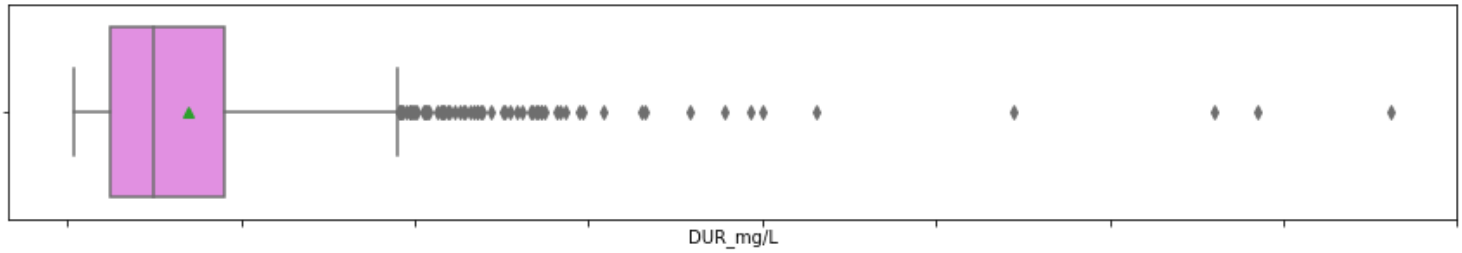

Mean
Median
Q1
Q3
0
50
100
150
200
250
300
0
2500
5000
7500
10000
12500
15000
17500
CONDUCT_mS/cm

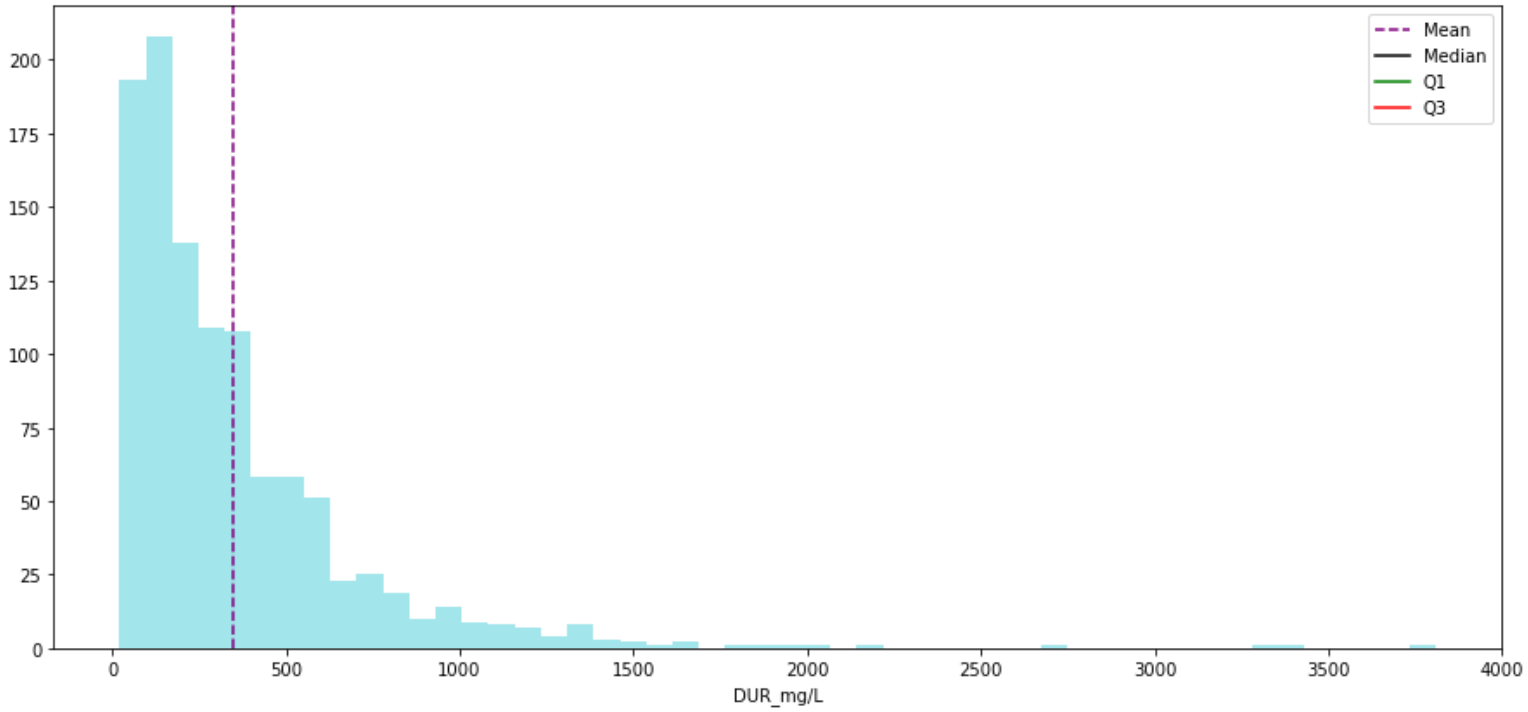

SDT_M_mg/L

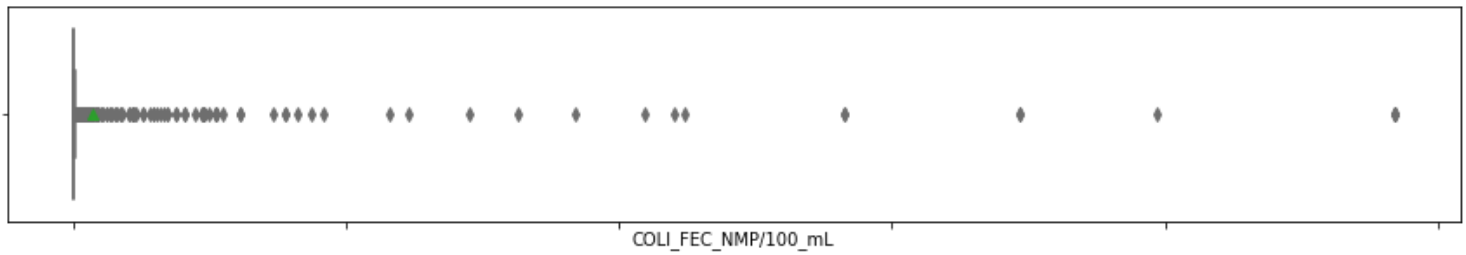

Mean
Median
Q1
Q3
1000
800
600
400
200
0
0
20000
40000
60000
80000
SDT_M_mg/L

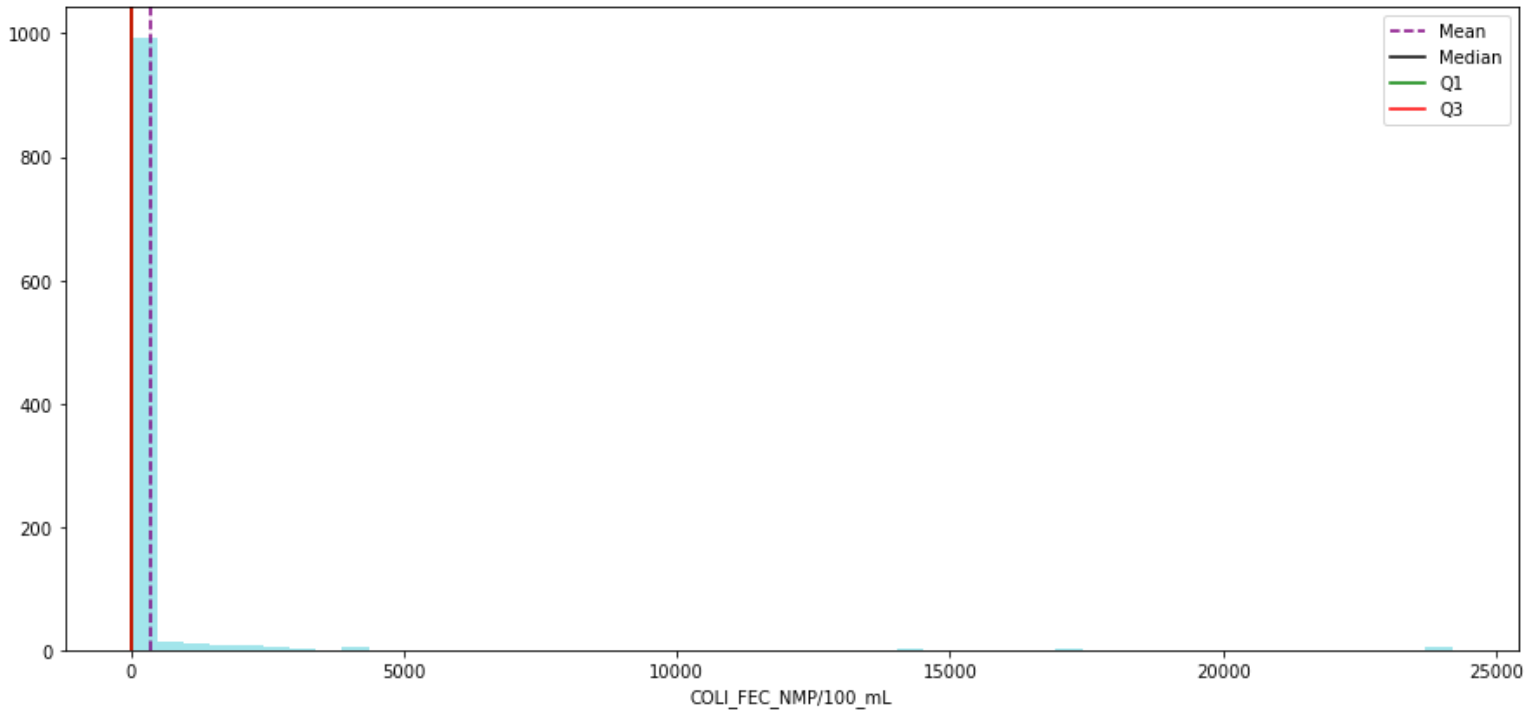

FLUORUROS_mg/L

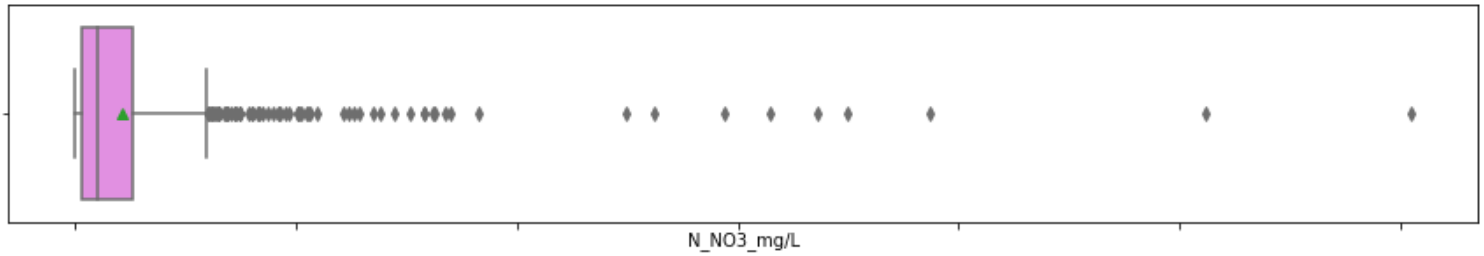

Mean
Median
Q1
Q3
700
600
500
400
300
200
100
0
0
5
10
15
20
25
30
35
FLUORUROS_mg/L

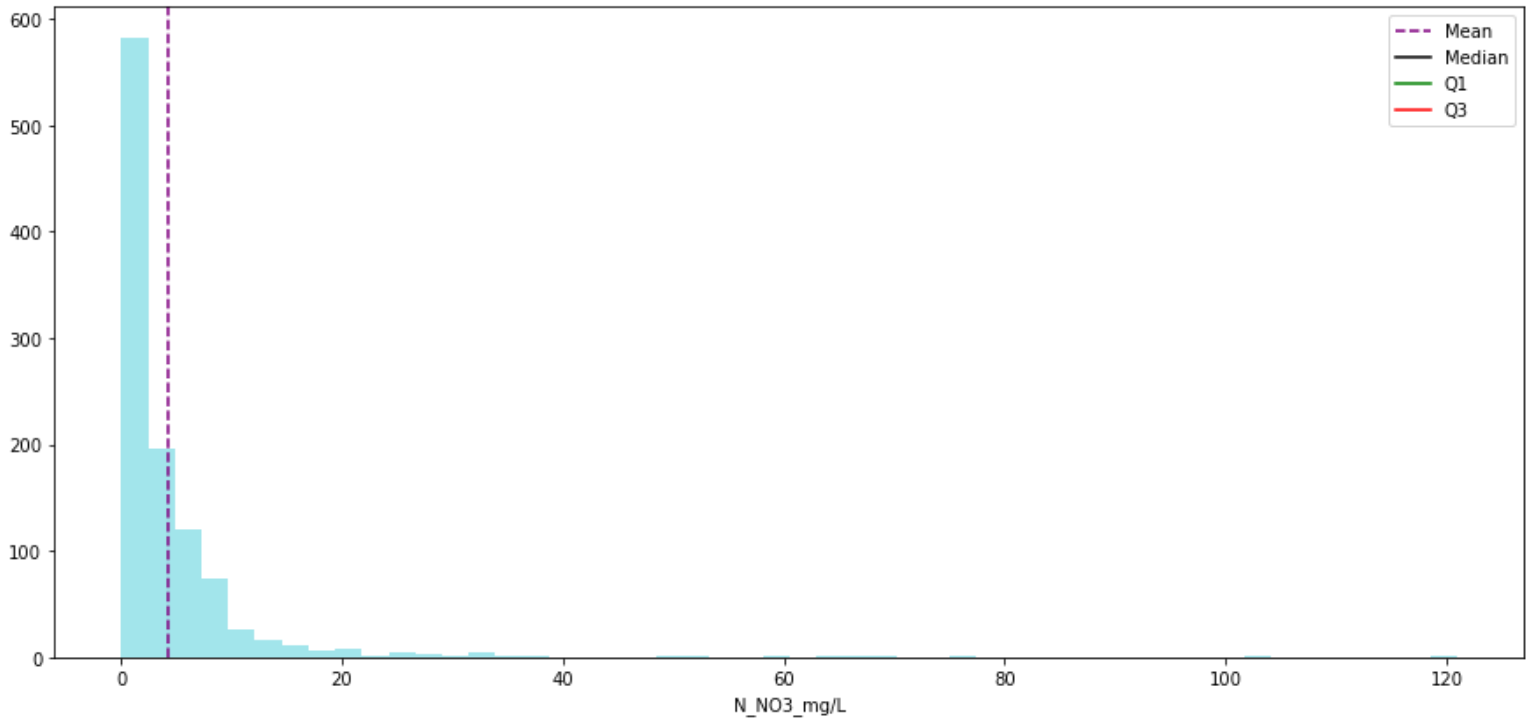

DUR_mg/L

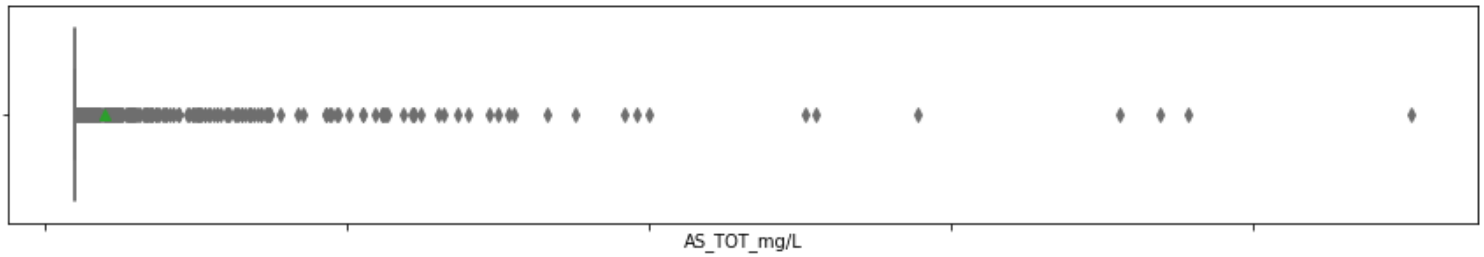

Mean
Median
Q1
Q3
0
25
50
75
100
125
150
175
200
0
500
1000
1500
2000
2500
3000
3500
4000
DUR_mg/L

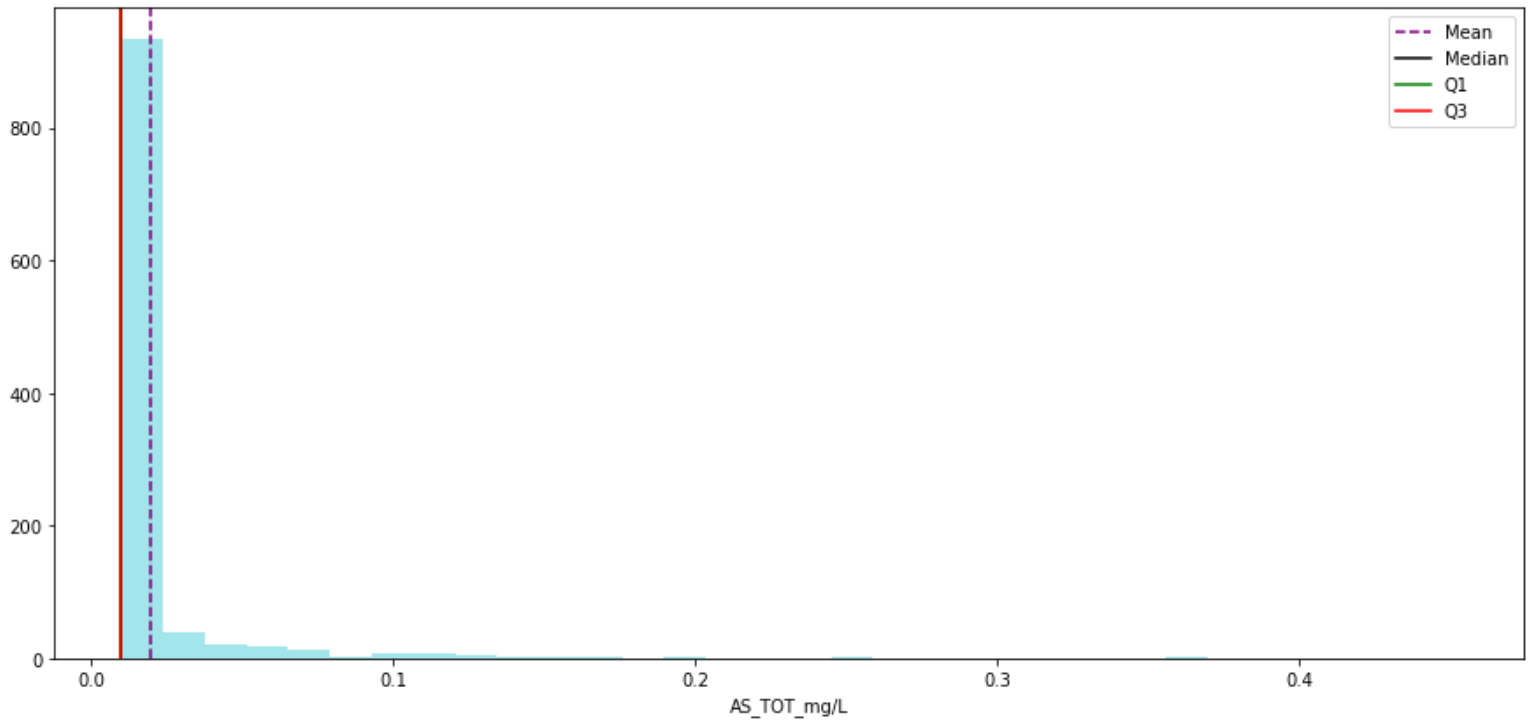

COLI_FEC_NMP/100_mL

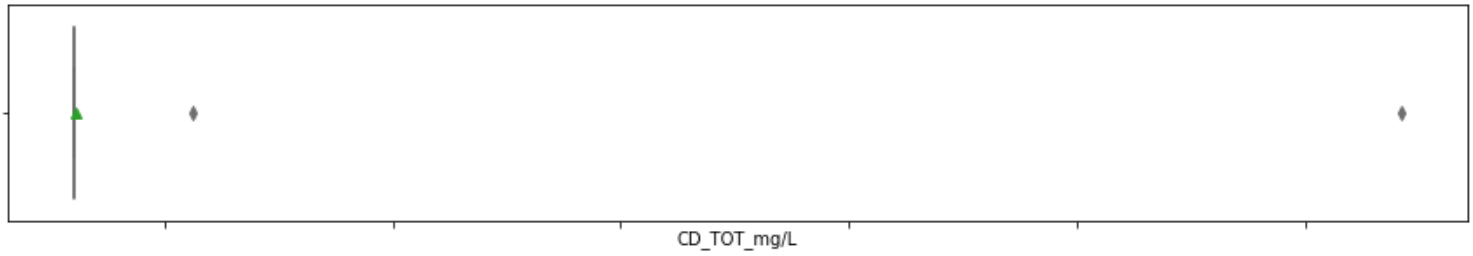

Mean
Median
Q1
Q3
0
200
400
600
800
1000
0
5000
10000
15000
20000
25000
COLI_FEC_NMP/100_mL

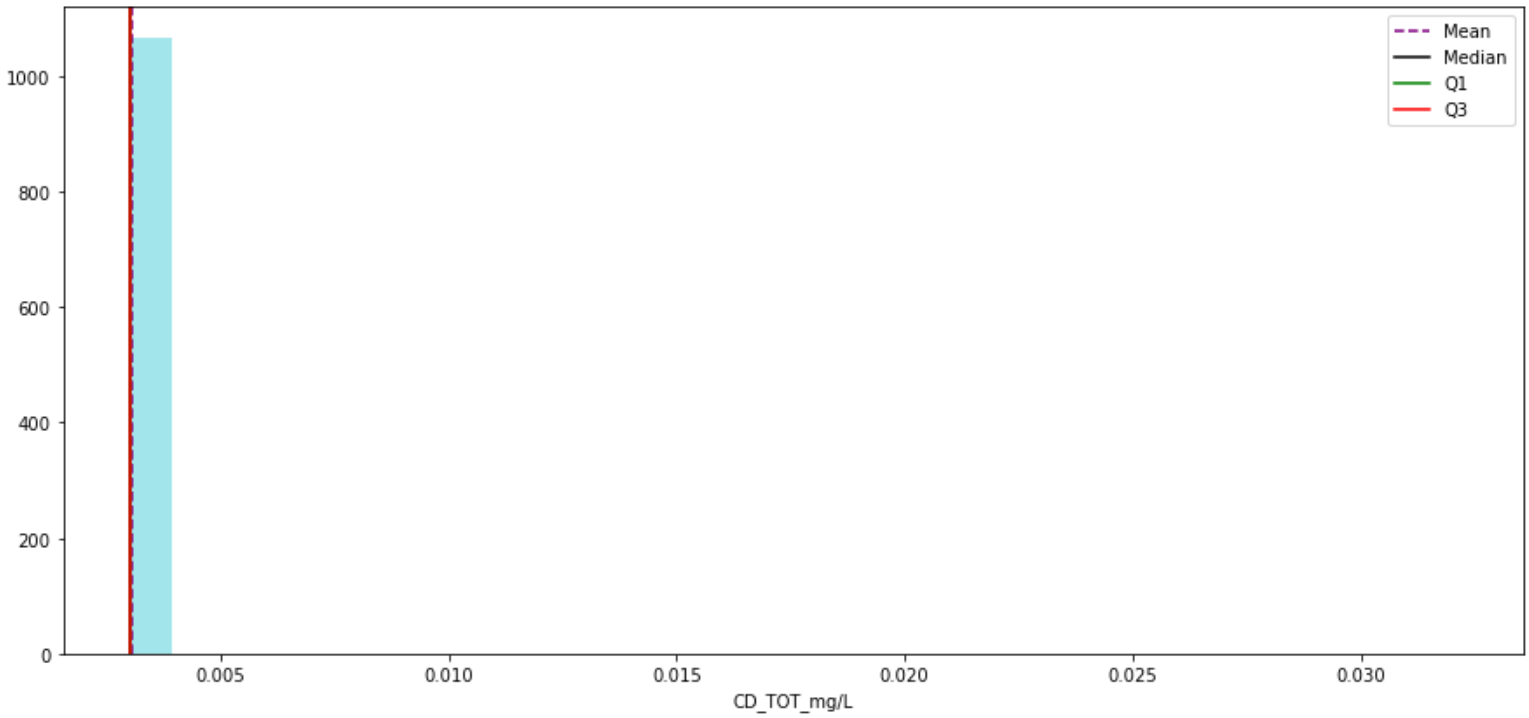

N_NO3_mg/L

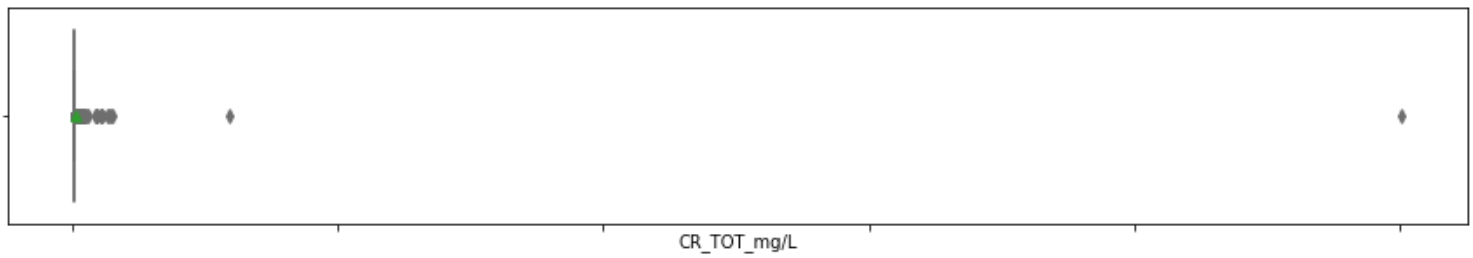

Mean
Median
Q1
Q3
600
500
400
300
200
100
0
0
20
40
60
80
100
120
N_NO3_mg/L

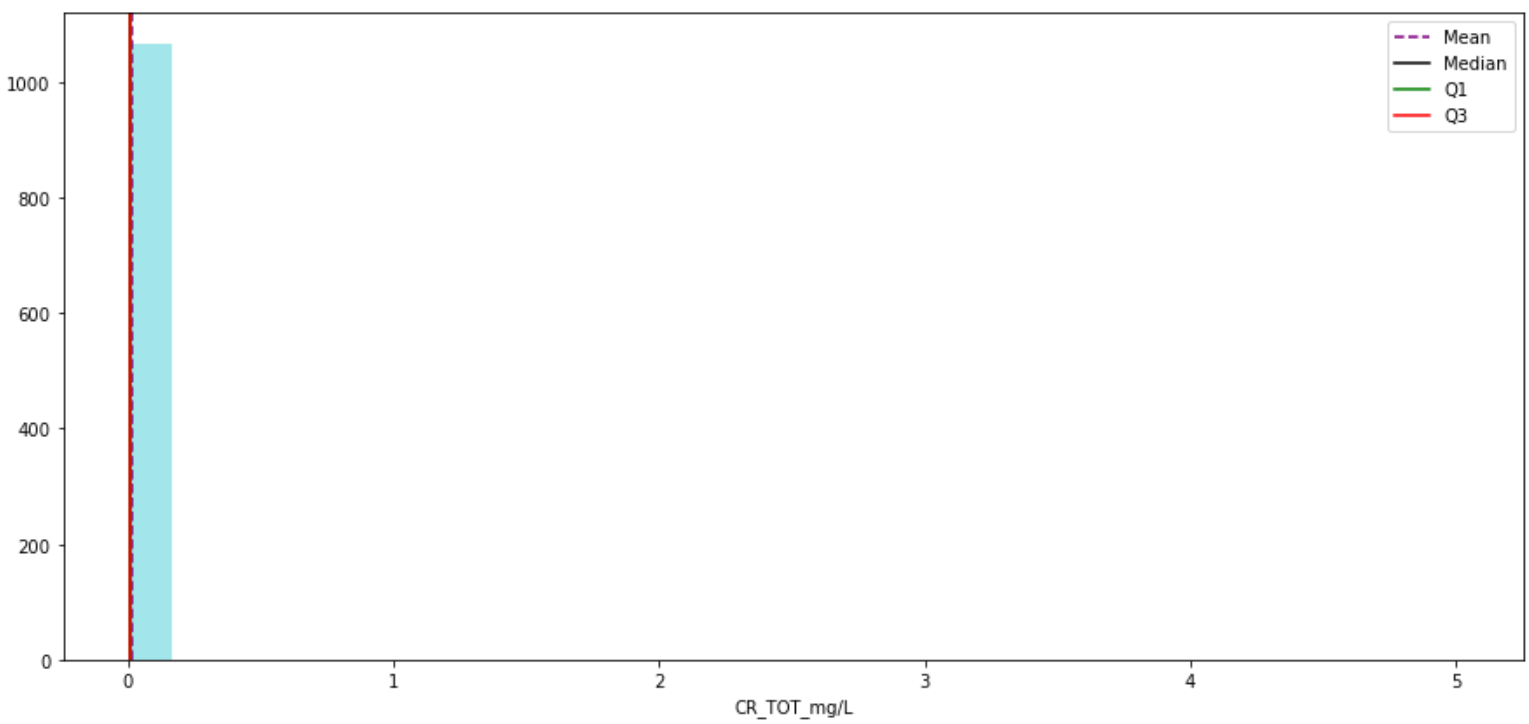

AS_TOT_mg/L

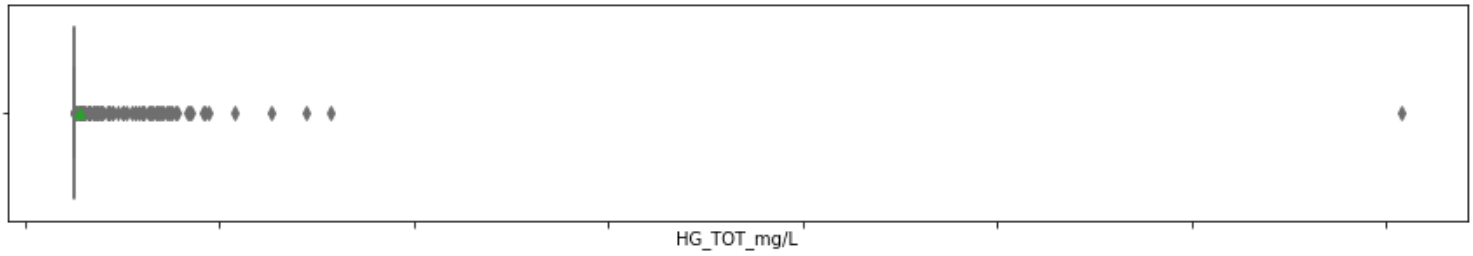

Mean
Median
Q1
Q3
800
600
400
200
0
0.0
0.1
0.2
0.3
0.4
AS_TOT_mg/L

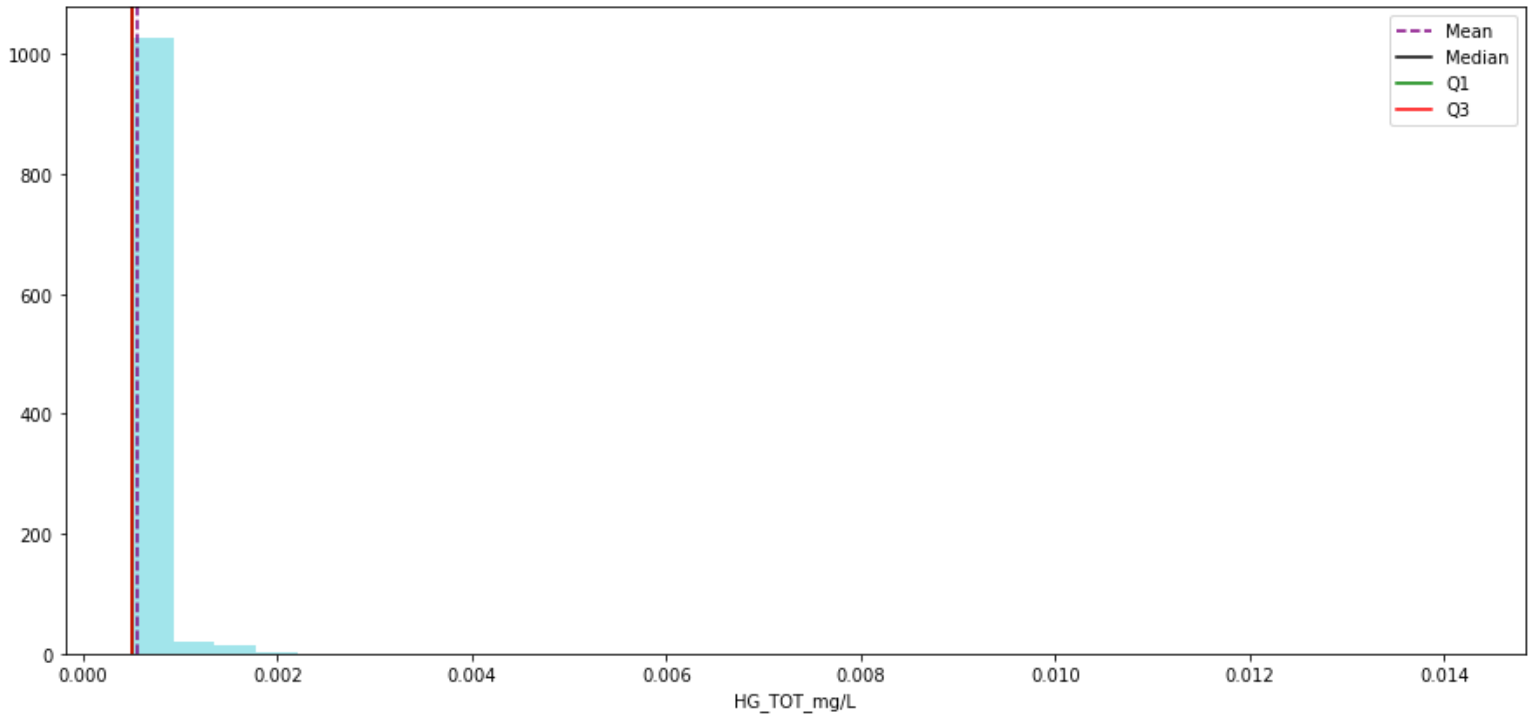

CD_TOT_mg/L

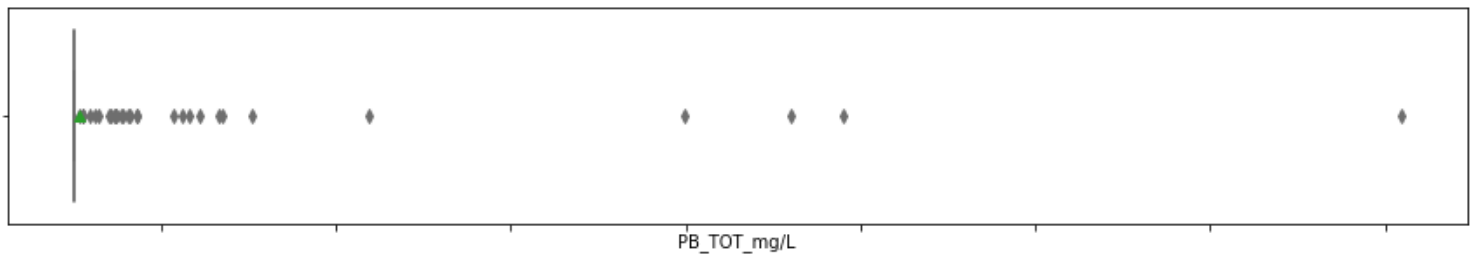

Mean
Median
Q1
Q3
1000
800
600
400
200
0
0.005
0.010
0.015
0.020
0.025
0.030
CD_TOT_mg/L

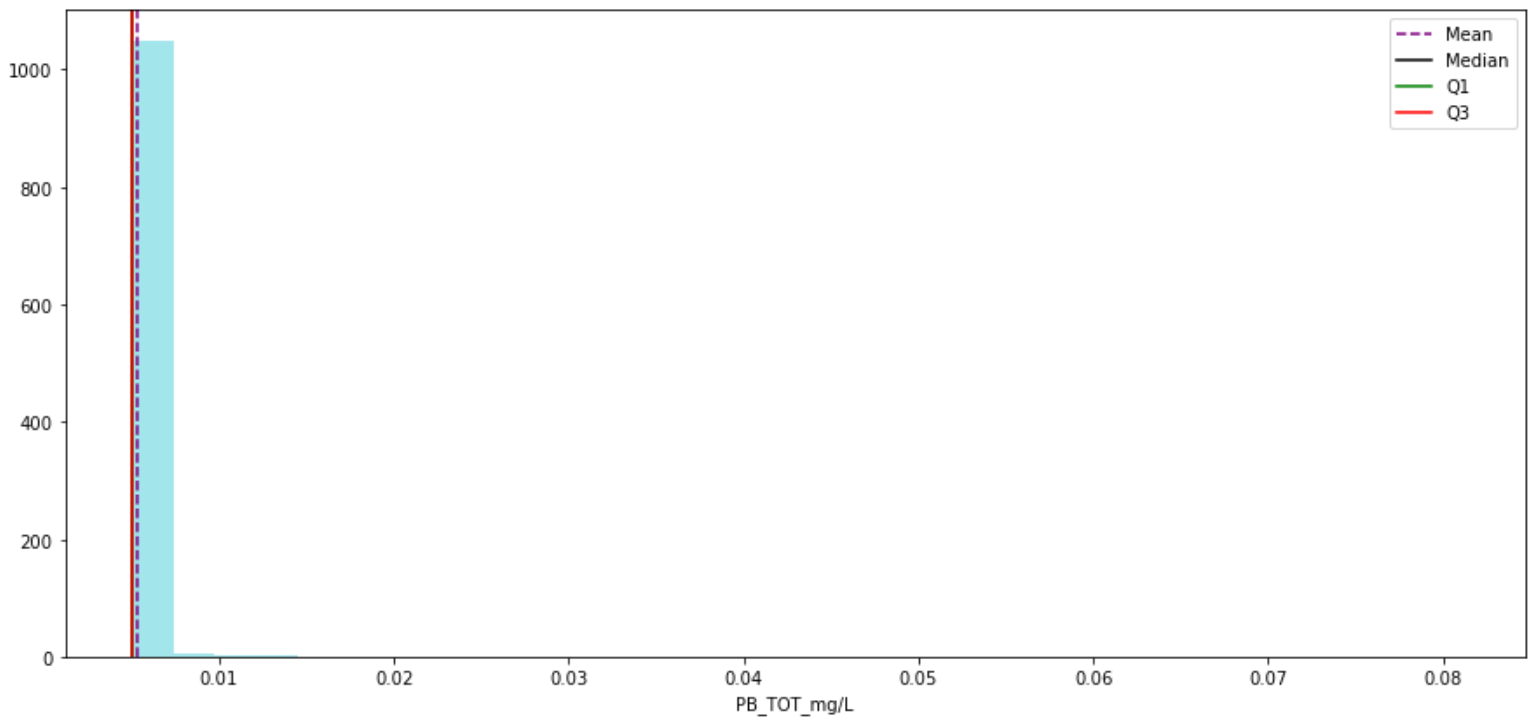

CR_TOT_mg/L

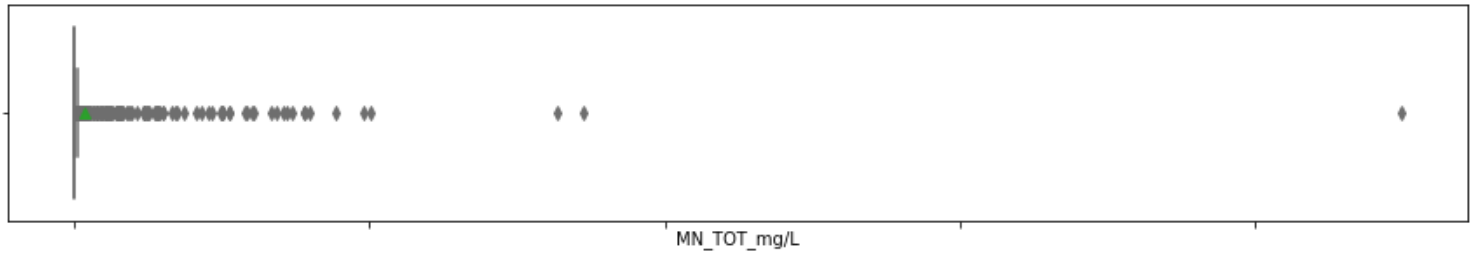

Mean
Median
Q1
Q3
1000
800
600
400
200
0
0
1
2
3
4
5
CR_TOT_mg/L

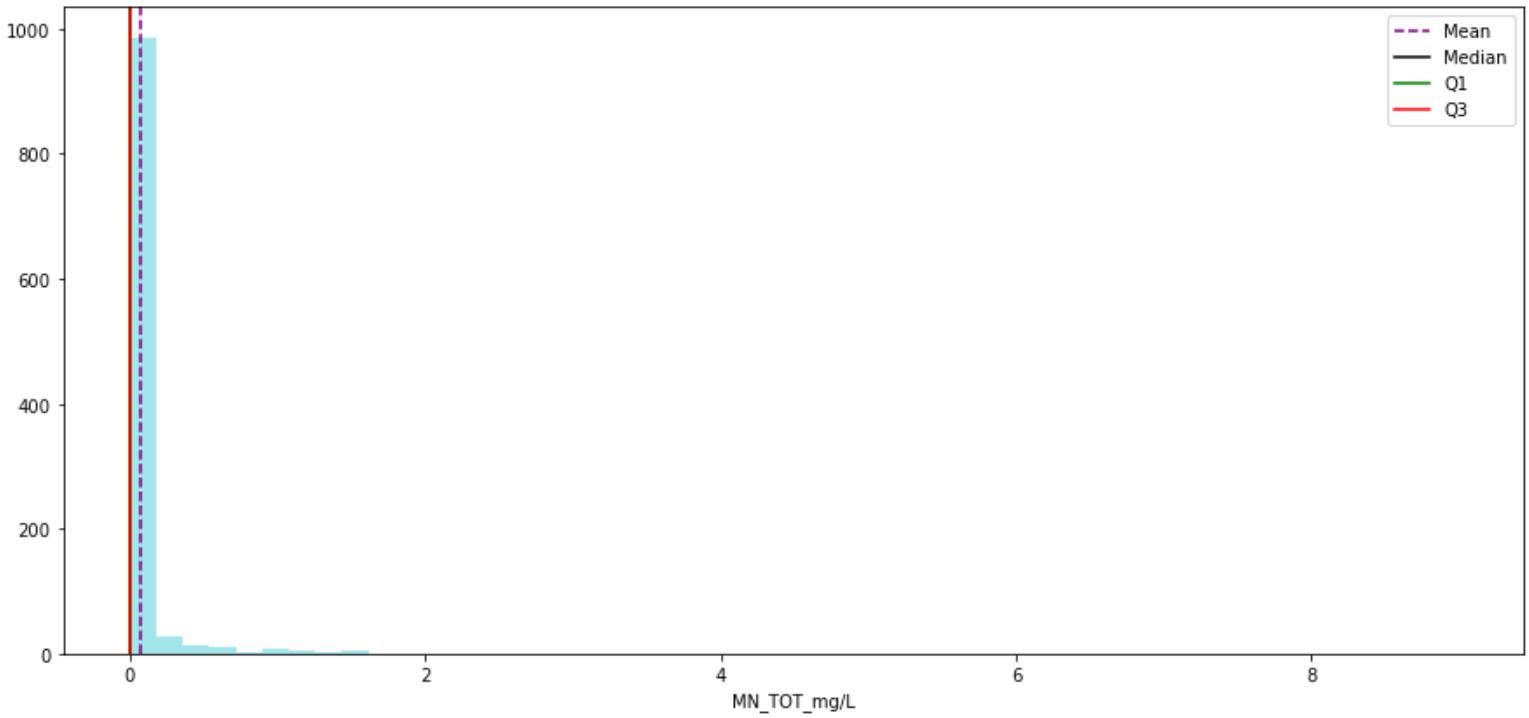

HG_TOT_mg/L

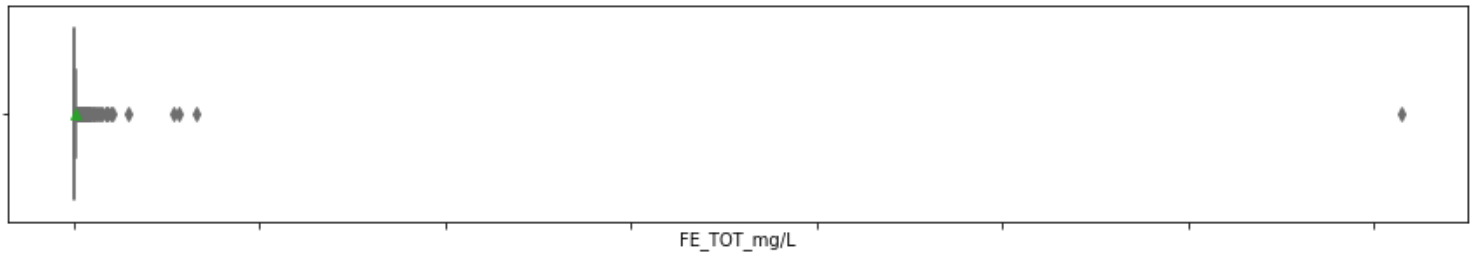

Mean
Median
Q1
Q3
1000
800
600
400
200
0
0.000
0.002
0.004
0.006
0.008
0.010
0.012
0.014
HG_TOT_mg/L

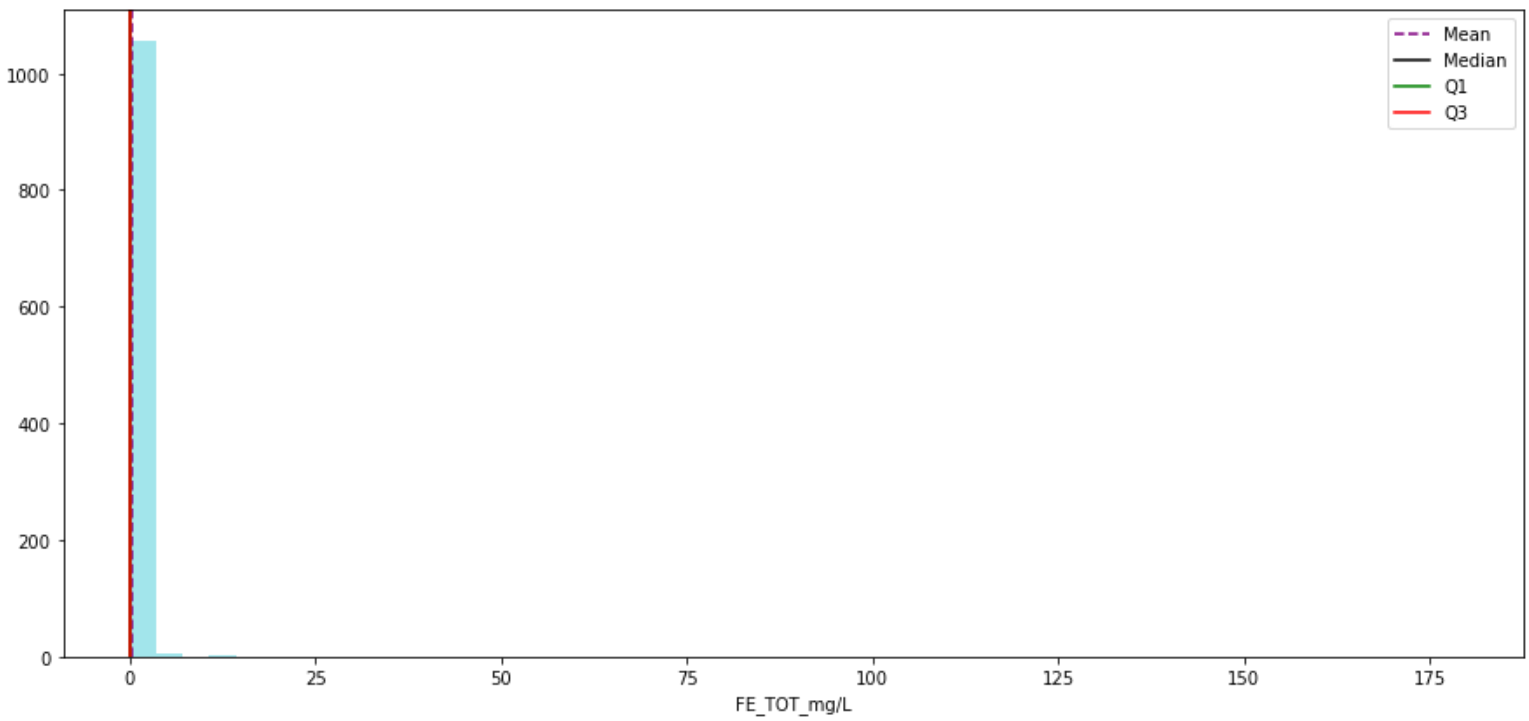

PB_TOT_mg/L


Mean
Median
Q1
Q3
1000
800
600
400
200
0
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
PB_TOT_mg/L


MN_TOT_mg/L

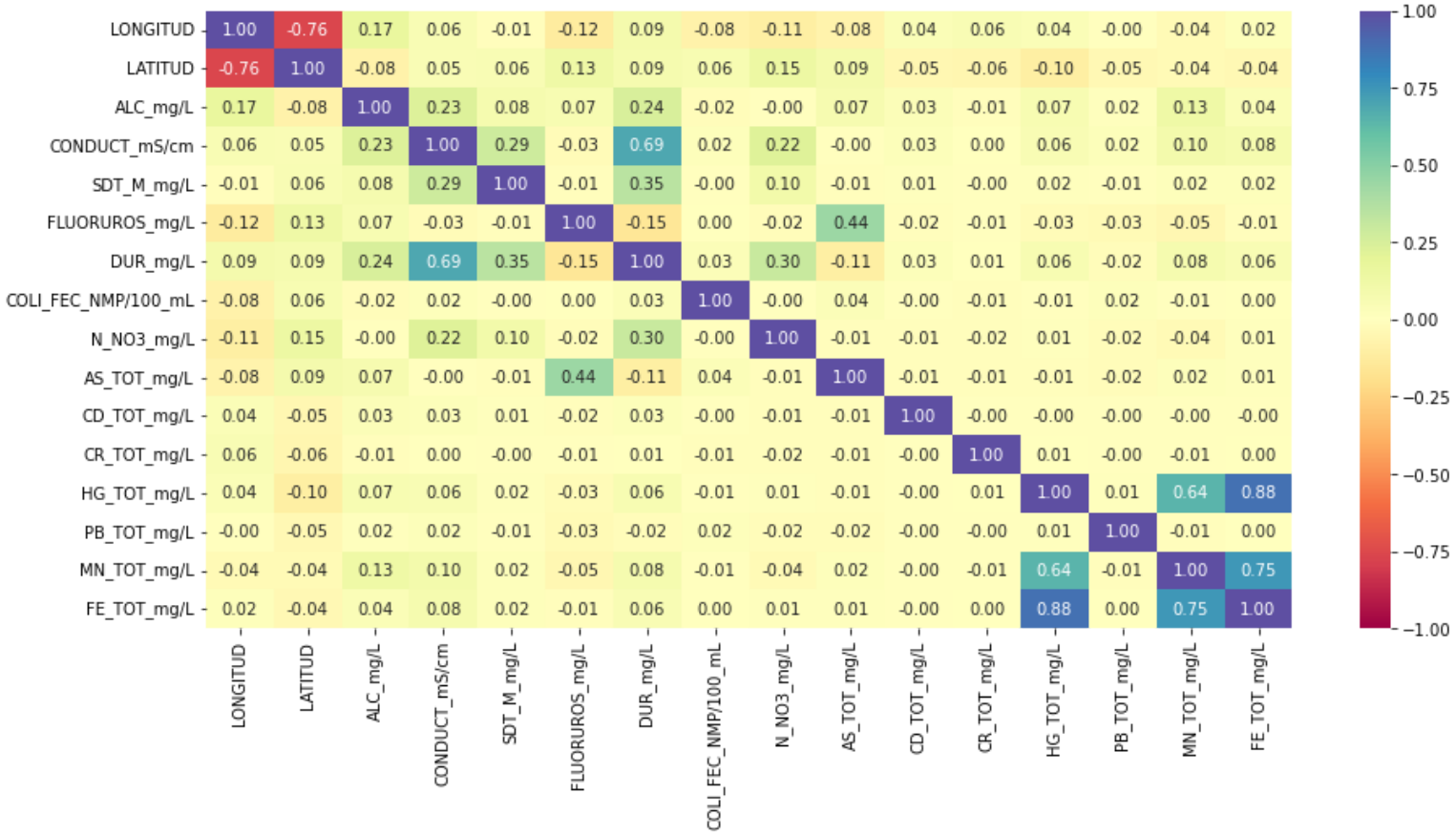

Mean
Median
Q1
Q3
1000
800
600
400
200
0
0
2
4
6
8
MN_TOT_mg/L


FE_TOT_mg/L


Mean
Median
Q1
Q3
1000
800
600
400
200
0
0
25
50
75
100
125
150
175
FE_TOT_mg/L

Podemos ver que hay un gran numero de outliers en todas las medidas por litro. Probablemente, no sea por ruido sino por casos muy aislados que se deben considerar al momento de hacer el analisis.

Dicho esto, procederemos a revisar la correlacion entre las variables numericas.



Podemos ver que existe una correlacion relativamente grande entre los contaminantes por Litro. Por ejemplo:

- Dur_mg/L vs. CONDUCT_mS/cm
- HG_TOT_mg/L vs. FE_TOT_mg/L
- MN_TOT_mg/L vs. FE_TOT_mg/L
- HG_TOT_mg/L vs. MN_TOT_mg/L

Ahora analizaremos las variables categoricas

```
===== Binary Columns =====
SI    1005
NO      59
ND       4
Name: CUMPLE_CON_ALC, dtype: int64
SI    939
NO    123
ND      6
Name: CUMPLE_CON_COND, dtype: int64
SI    995
NO     71
ND      2
Name: CUMPLE_CON_SDT_ra, dtype: int64
SI    995
NO     71
ND      2
Name: CUMPLE_CON_SDT_salin, dtype: int64
SI    876
NO    192
```

```
Name: CUMPLE_CON_FLUO, dtype: int64
SI    841
NO    226
ND      1
Name: CUMPLE_CON_DUR, dtype: int64
SI    1007
NO      61
Name: CUMPLE_CON_CF, dtype: int64
SI    985
NO     82
ND      1
Name: CUMPLE_CON_NO3, dtype: int64
SI    941
NO    127
Name: CUMPLE_CON_AS, dtype: int64
SI    1066
NO       2
Name: CUMPLE_CON_CD, dtype: int64
SI    1053
NO      15
Name: CUMPLE_CON_CR, dtype: int64
SI    1067
NO       1
Name: CUMPLE_CON_HG, dtype: int64
SI    1056
NO      12
Name: CUMPLE_CON_PB, dtype: int64
SI    982
NO     86
Name: CUMPLE_CON_MN, dtype: int64
SI    932
NO    136
Name: CUMPLE_CON_FE, dtype: int64
===== Categorical Columns =====
DLAGU6         1
OCGCE3209      1
OCFSU2993      1
OCFSU2994      1
OCFSU3048      1
              ..
DLHID6458      1
DLHID6461      1
DLHID6463      1
DLHID6467      1
OCRBR5109M1    1
Name: CLAVE, Length: 1068, dtype: int64
EL FUERTE                            2
POZO VILLA UNION                     2
POZO BERRIOZABAL                     1
RANCHO GIUSEPPE CONSTANZO            1
QUINTA 2 POTRILLOS                   1
                                   ..
POZO SAN FRANCISCO BOJAY COLONIA     1
POZO SANTA ANA AHUEHUEPAN            1
POZO SANTA MARIA DAXTHO              1
POZO PEDRO MARIA ANAYA               1
COMUNIDAD LA REFORMA                 1
Name: SITIO, Length: 1066, dtype: int64
CUENCAS CENTRALES DEL NORTE     232
LERMA SANTIAGO PACIFICO         170
PENINSULA DE YUCATAN            125
NOROESTE                         94
PENINSULA DE BAJA CALIFORNIA     89
BALSAS                           69
RIO BRAVO                        65
PACIFICO NORTE                   62
```

```
GOLFO NORTE                        53
AGUAS DEL VALLE DE MEXICO          38
FRONTERA SUR                       34
GOLFO CENTRO                       21
PACIFICO SUR                       16
Name: ORGANISMO_DE_CUENCA, dtype: int64
DURANGO                             121
SONORA                              103
YUCATAN                              85
ZACATECAS                            75
COAHUILA DE ZARAGOZA                 59
BAJA CALIFORNIA SUR                  49
SAN LUIS POTOSI                      47
GUANAJUATO                           41
HIDALGO                              37
CHIHUAHUA                            35
JALISCO                              33
SINALOA                              32
BAJA CALIFORNIA                      31
MICHOACAN DE OCAMPO                  27
COLIMA                               26
CAMPECHE                             25
TAMAULIPAS                           25
MEXICO                               24
TLAXCALA                             24
PUEBLA                               23
CHIAPAS                              21
OAXACA                               20
VERACRUZ DE IGNACIO DE LA LLAVE      16
NUEVO LEON                           15
QUINTANA ROO                         15
AGUASCALIENTES                       14
TABASCO                              13
MORELOS                              11
NAYARIT                               8
QUERETARO ARTEAGA                     6
GUERRERO                              5
DISTRITO FEDERAL                      2
Name: ESTADO, dtype: int64
LA PAZ                   27
ENSENADA                 26
PARRAS                   24
HERMOSILLO               17
MERIDA                   16
                         ..
CUAUTEPEC DE HINOJOSA     1
LAS ROSAS                 1
SOCOLTENANGO              1
COMITAN DE DOMINGUEZ      1
MELCHOR OCAMPO            1
Name: MUNICIPIO, Length: 452, dtype: int64
PENINSULA DE YUCATAN            119
PRINCIPAL-REGION LAGUNERA        28
ALTO ATOYAC                      19
TEPEHUANES-SANTIAGO              16
LA PAILA                         12
                               ...
CABRERA-OCAMPO                    1
TENANCINGO                        1
IRAPUATO-VALLE                    1
ROSARIO-TESOPACO-EL QUIRIEGO      1
TEPEJI DEL RIO                    1
Name: ACUIFERO, Length: 273, dtype: int64
POZO            1039
MANANTIAL         12
CENOTE             7
```

```
POZO NORIA          4
NORIA               3
DESCARGA            1
Pozo                1
BOMBEO CENOTE       1
Name: SUBTIPO, dtype: int64
Potable - Dulce         834
Ligeramente salobres    161
Salobres                 68
Salinas                   3
Name: CALIDAD_SDT_salin, dtype: int64
Potable - Dura                                        577
Muy dura e indeseable usos industrial y domestico     226
Potable - Moderadamente suave                         168
Potable - Suave                                        96
Name: CALIDAD_DUR, dtype: int64
Potable - Excelente        788
Potable - Buena calidad    197
No apta como FAAP           82
Name: CALIDAD_N_NO3, dtype: int64
Potable - Excelente    816
No apta como FAAP      127
Apta como FAAP         125
Name: CALIDAD_AS, dtype: int64
Potable - Excelente    1066
No apta como FAAP         2
Name: CALIDAD_CD, dtype: int64
Potable - Excelente    1053
No apta como FAAP        15
Name: CALIDAD_CR, dtype: int64
Potable - Excelente    1067
No apta como FAAP         1
Name: CALIDAD_HG, dtype: int64
Potable - Excelente    1056
No apta como FAAP        12
Name: CALIDAD_PB, dtype: int64
Potable - Excelente                                   982
Puede afectar la salud                                 50
Sin efectos en la salud - Puede dar color al agua      36
Name: CALIDAD_MN, dtype: int64
Potable - Excelente                                   932
Sin efectos en la salud - Puede dar color al agua     136
Name: CALIDAD_FE, dtype: int64
Verde       434
Rojo        387
Amarillo    247
Name: SEMAFORO, dtype: int64
FLUO,                                                 78
DT,                                                   65
FLUO,AS,                                              51
CF,                                                   31
AS,                                                   31
                                                      ..
ALC,CONDUC,SDT_ra,SDT_salin,DT,NO3,                    1
ALC,CONDUC,SDT_ra,SDT_salin,FLUO,DT,AS,MN,FE,          1
PB,MN,FE,                                              1
ALC,AS,FE,                                             1
ALC,DT,NO3,                                            1
Name: CONTAMINANTES, Length: 126, dtype: int64
===== Ordinal Columns =====
Alta                    794
Media                   187
Indeseable como FAAP     59
Baja                     24
Name: CALIDAD_ALC, dtype: int64
Permisible para riego    460
```

```
Buena para riego         434
Dudosa para riego         72
Indeseable para riego     51
Excelente para riego      45
Name: CALIDAD_CONDUC, dtype: int64
Excelente para riego            491
Cultivos sensibles              343
Cultivos con manejo especial    161
Cultivos tolerantes              64
Indeseable para riego             7
Name: CALIDAD_SDT_ra, dtype: int64
Baja                434
Potable - Optima    226
Media               216
Alta                192
Name: CALIDAD_FLUO, dtype: int64
Potable - Excelente        739
Buena calidad              208
Aceptable                   60
Contaminada                 49
Fuertemente contaminada     12
Name: CALIDAD_COLI_FEC, dtype: int64
```

De las variables categoricas podemos ver que hay mucho trabajo por hacer.

1. En las variables binarias, vemos que hay un typo donde escribieron ND en lugar de no. Por lo tanto, convertiremos estos valores a NO

1. Podemos ver que hay valores redundantes en cuanto a locación y hay otros que no nos sirven de nada para el análisis estadístico:

    a. No aportan información extra al modelo: CLAVE, SITIO
    b. Información Redundante: ESTADO, MUNICIPIO, ACUIFERO

1. La columna de contaminantes contiene multiples categorias, por lo que vamos a dividirlo en multiples columnas y convertirlos en columnas binarias. Como nota extra, podemos notar que los valores NA de esta columna, estan atribuidos a todos aquellos que no tienen contaminantes.

```
'CONTAMINANTES'
```

1. En el Subtipo de acuifero, hay valores que solo tienen un typo. Por lo que solo se requiere corregirlos. Y hay unos que son bi-clase, por lo que podemos juntarlos

```
POZO          1040
MANANTIAL       12
CENOTE           8
POZO/NORIA       7
DESCARGA         1
Name: SUBTIPO, dtype: int64
```

Hecho esto, lo primero que haremos será hacerle un One Hot Encoding a las variables categóricas.

Posteriormente, le aplicaremos el método de KNN Imputer para encontrar los valores faltantes

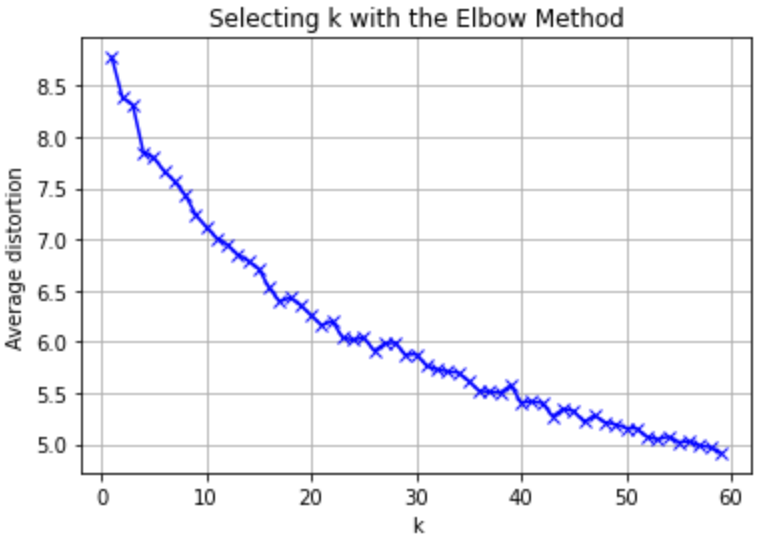
Y finalmente le haremos un escalamiento a los valores numéricos

	LONGITUD	LATITUD	ALC_mg/L	CONDUCT_mS/cm	SDT_M_mg/L	FLUORUROS_mg/L	DUR_mg/L	COLI_FEC_NMP/100_mL
0	-0.019566	-0.245699	-0.046414	-0.161939	-0.106261	-0.051472	-0.373084	-0.172747
1	-0.046229	-0.299558	-0.029289	-0.427832	-0.163836	-0.075804	-0.452894	-0.172747
2	-0.059253	-0.205043	-0.261083	-0.488699	-0.201467	0.378969	-0.631914	-0.172747
3	-0.060220	-0.252009	0.784261	-0.365363	-0.151753	0.024592	-0.411633	-0.172747
4	-1.246812	0.074054	0.637709	0.559657	0.103150	-0.437408	0.359484	-0.031436

Ahora que ya tenemos preparados los datos podemos empezar con el modelo inicial de K-means para escoger el número de clusters óptimo.

Para esto usaremos el método del codo.

```
Text(0.5, 1.0, 'Selecting k with the Elbow Method')
```



Ya que no es muy claro, usaremos el average silhouette coefficient para seleccionar el k óptimo

```
For K= 2 Silhouette Score ----> 0.394
For K= 3 Silhouette Score ----> 0.373
For K= 4 Silhouette Score ----> 0.141
For K= 5 Silhouette Score ----> 0.126
For K= 6 Silhouette Score ----> 0.145
For K= 7 Silhouette Score ----> 0.112
For K= 8 Silhouette Score ----> 0.118
For K= 9 Silhouette Score ----> 0.101
For K= 10 Silhouette Score ----> 0.096
For K= 11 Silhouette Score ----> 0.086
For K= 12 Silhouette Score ----> 0.104
For K= 13 Silhouette Score ----> 0.098
For K= 14 Silhouette Score ----> 0.117
For K= 15 Silhouette Score ----> 0.112
For K= 16 Silhouette Score ----> 0.112
For K= 17 Silhouette Score ----> 0.116
For K= 18 Silhouette Score ----> 0.125
For K= 19 Silhouette Score ----> 0.119
For K= 20 Silhouette Score ----> 0.109
For K= 21 Silhouette Score ----> 0.121
```

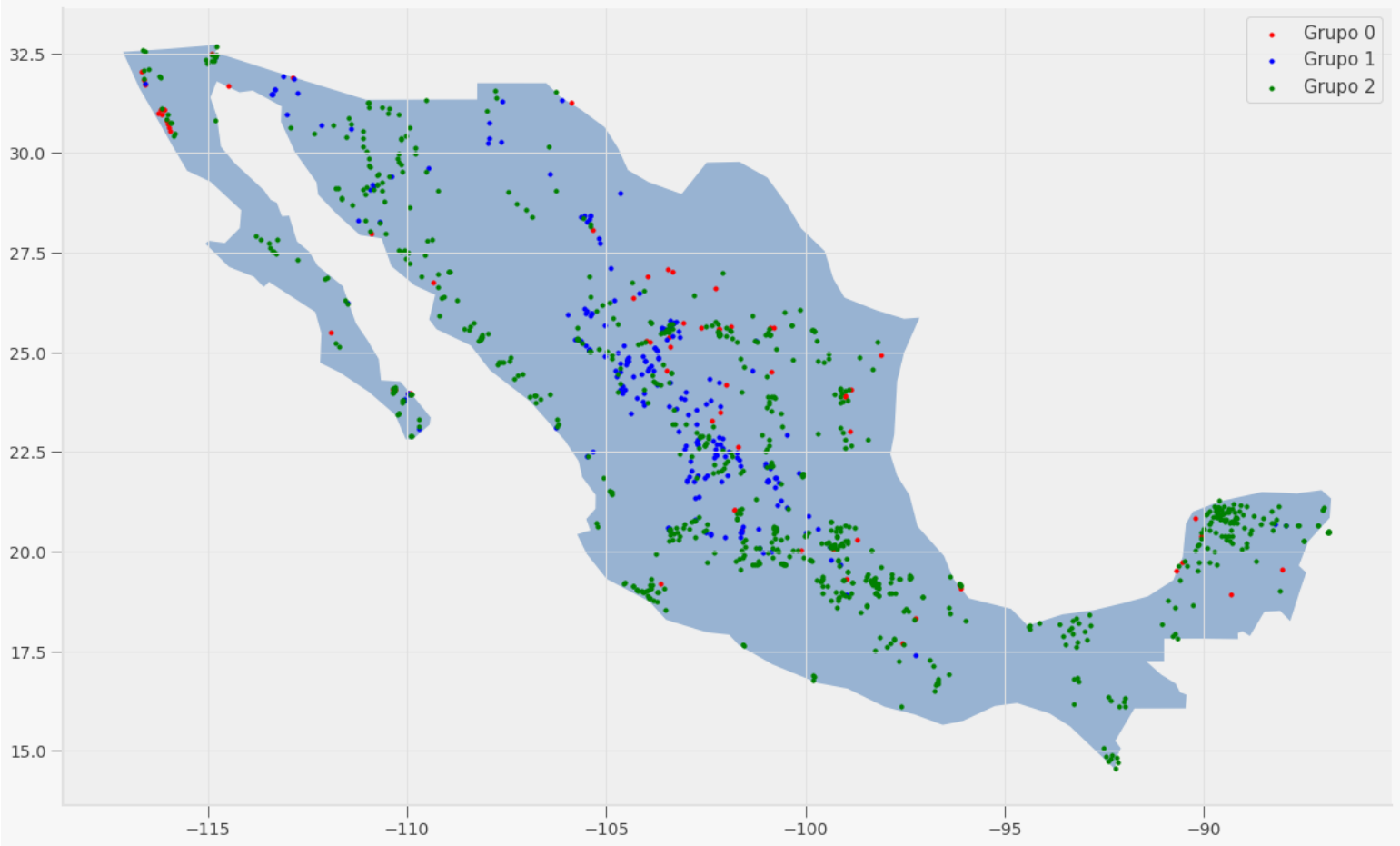
```
 For K= 22 Silhouette Score ----> 0.121
 For K= 23 Silhouette Score ----> 0.134
 For K= 24 Silhouette Score ----> 0.129
 For K= 25 Silhouette Score ----> 0.117
 For K= 26 Silhouette Score ----> 0.123
 For K= 27 Silhouette Score ----> 0.145
 For K= 28 Silhouette Score ----> 0.135
 For K= 29 Silhouette Score ----> 0.131
 For K= 30 Silhouette Score ----> 0.138
 For K= 31 Silhouette Score ----> 0.124
 For K= 32 Silhouette Score ----> 0.14
 For K= 33 Silhouette Score ----> 0.144
 For K= 34 Silhouette Score ----> 0.143
 For K= 35 Silhouette Score ----> 0.136
 For K= 36 Silhouette Score ----> 0.147
 For K= 37 Silhouette Score ----> 0.141
 For K= 38 Silhouette Score ----> 0.142
 For K= 39 Silhouette Score ----> 0.143
 For K= 40 Silhouette Score ----> 0.157
 For K= 41 Silhouette Score ----> 0.151
 For K= 42 Silhouette Score ----> 0.155
 For K= 43 Silhouette Score ----> 0.152
 For K= 44 Silhouette Score ----> 0.148
 For K= 45 Silhouette Score ----> 0.144
 For K= 46 Silhouette Score ----> 0.15
 For K= 47 Silhouette Score ----> 0.144
 For K= 48 Silhouette Score ----> 0.156
 For K= 49 Silhouette Score ----> 0.155
 For K= 50 Silhouette Score ----> 0.146
 For K= 51 Silhouette Score ----> 0.143
 For K= 52 Silhouette Score ----> 0.154
 For K= 53 Silhouette Score ----> 0.165
 For K= 54 Silhouette Score ----> 0.162
 For K= 55 Silhouette Score ----> 0.148
 For K= 56 Silhouette Score ----> 0.155
 For K= 57 Silhouette Score ----> 0.157
 For K= 58 Silhouette Score ----> 0.159
 For K= 59 Silhouette Score ----> 0.165
```

Usaremos k = 3 ya que es un buen número de clusters y también obtuvo un buen silhouette score

```
0    782
2    215
1     71
Name: GRUPO, dtype: int64
```

Una vez hechos los clusters, podemos revisar la relación entre los clusters y el resto de sus variables.

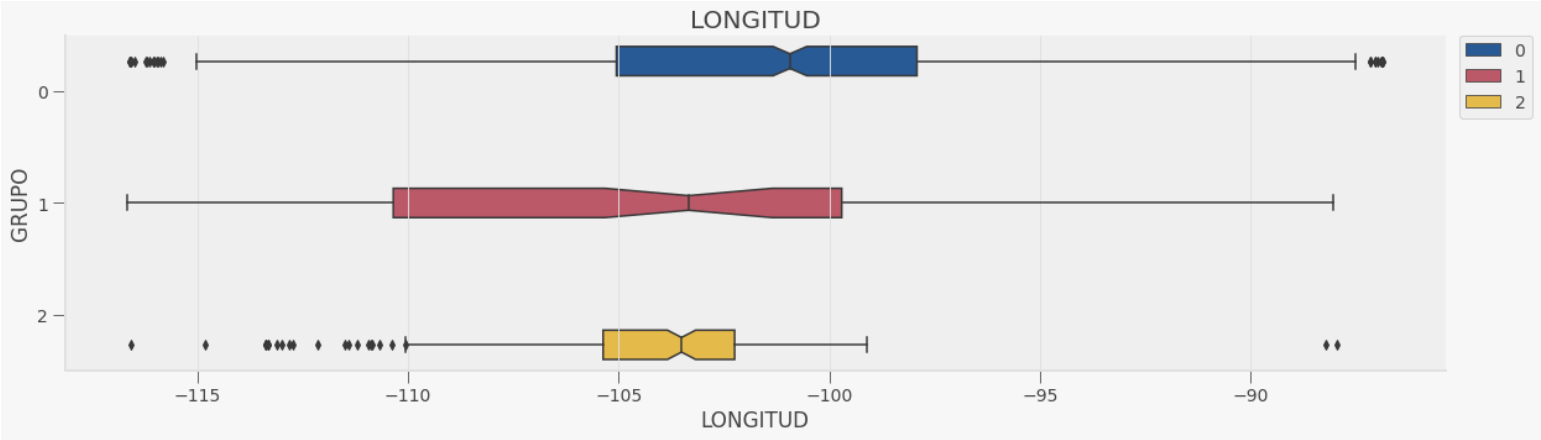
Cluster vs. Ubicación

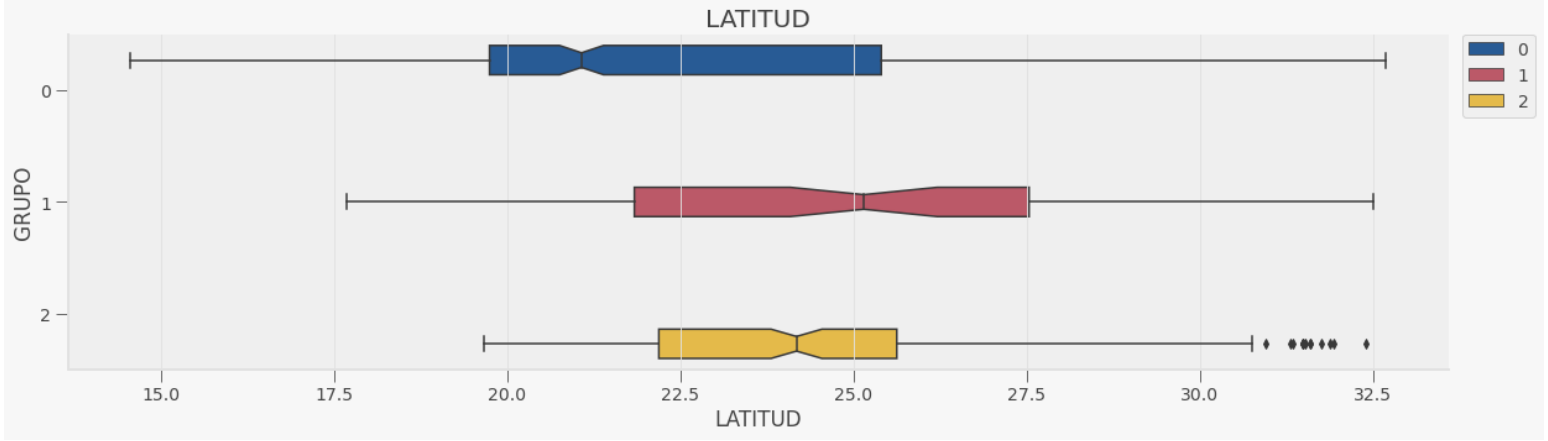


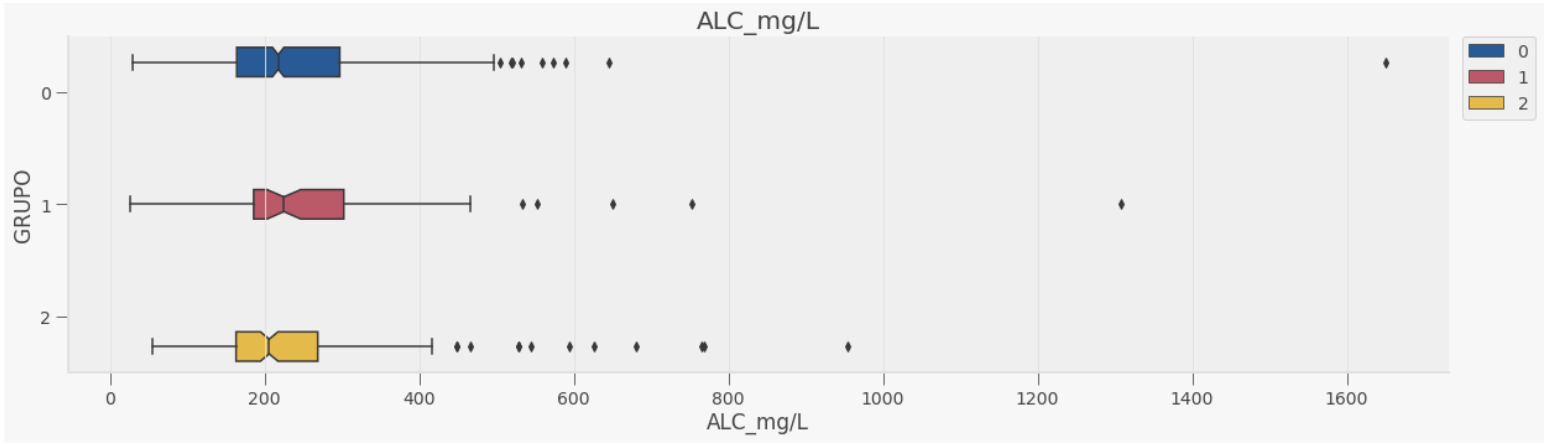
Podemos observar que los clusters no tienen tanto peso en cuanto a la localiazación. El grupo 0 esta mas localizado en el centro, el grupo 1 & 2 en centro y costas.

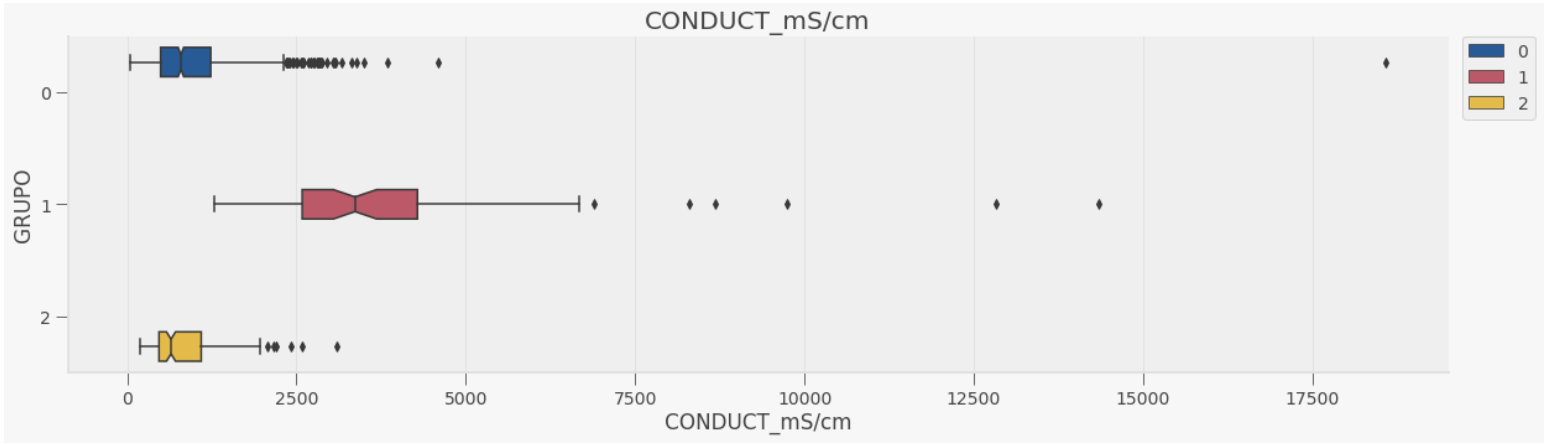
Dicho esto, vamos a analizar las variables numéricas. Para esto, obtendremos el promedio por grupo.

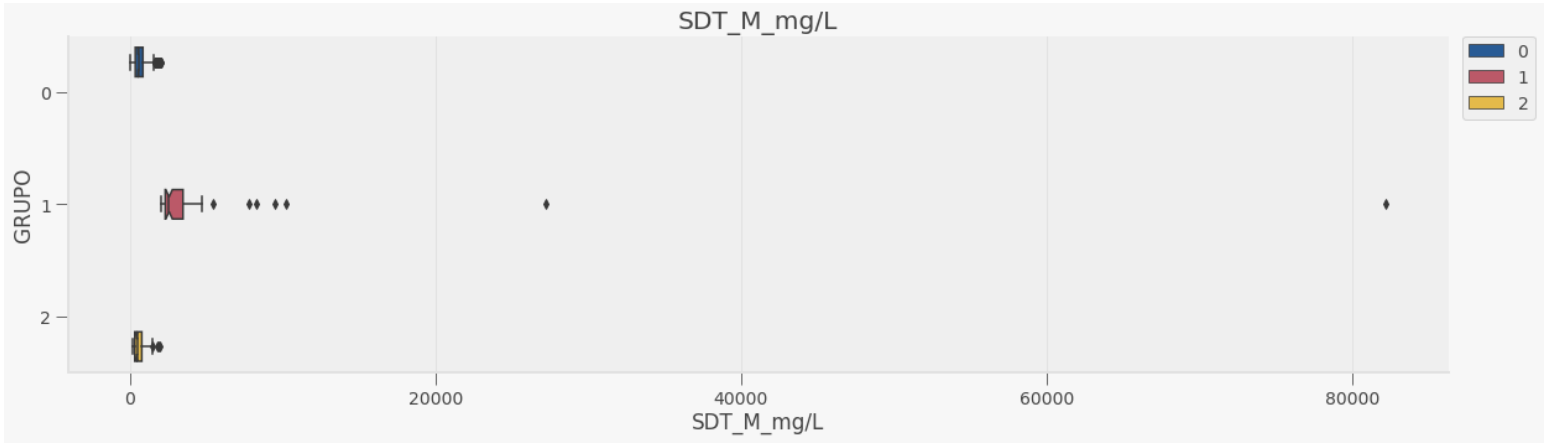
GRUPO	LONGITUD	LATITUD	ALC_mg/L	CONDUCT_mS/cm	SDT_M_mg/L	FLUORUROS_mg/L	DUR_mg/L	COLI_FEC_NMP/1
0	-101.017409	22.621437	232.922356	978.150834	646.389064	0.526215	324.601256	363.
1	-104.578911	25.143258	265.740423	3890.508571	4607.944600	1.123744	1183.714220	593.
2	-104.180833	24.481902	235.515117	822.784038	576.263876	3.057928	156.709920	248.

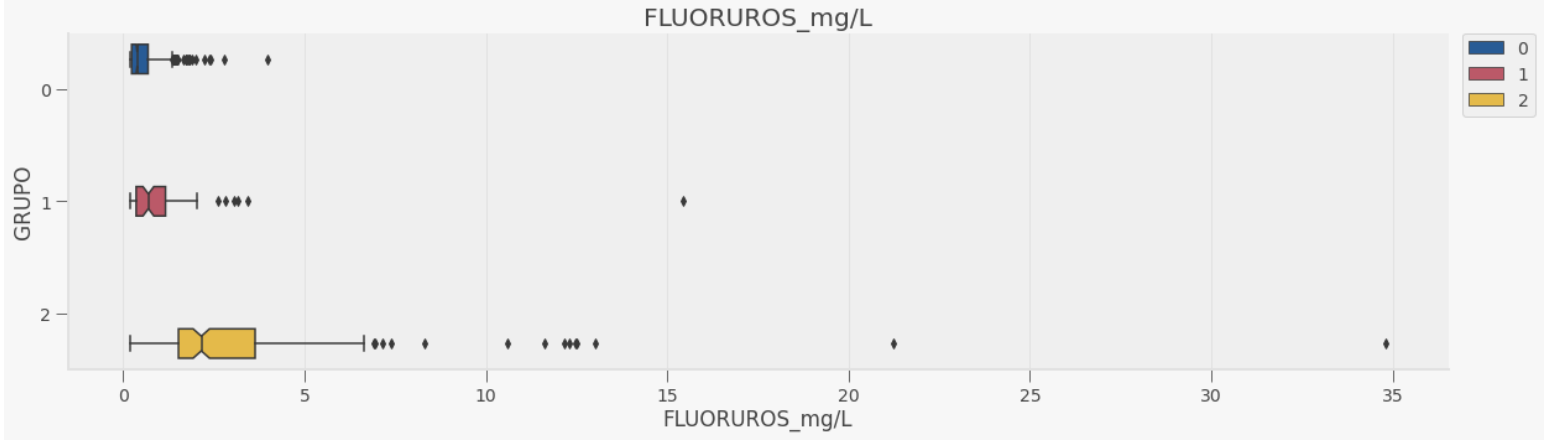


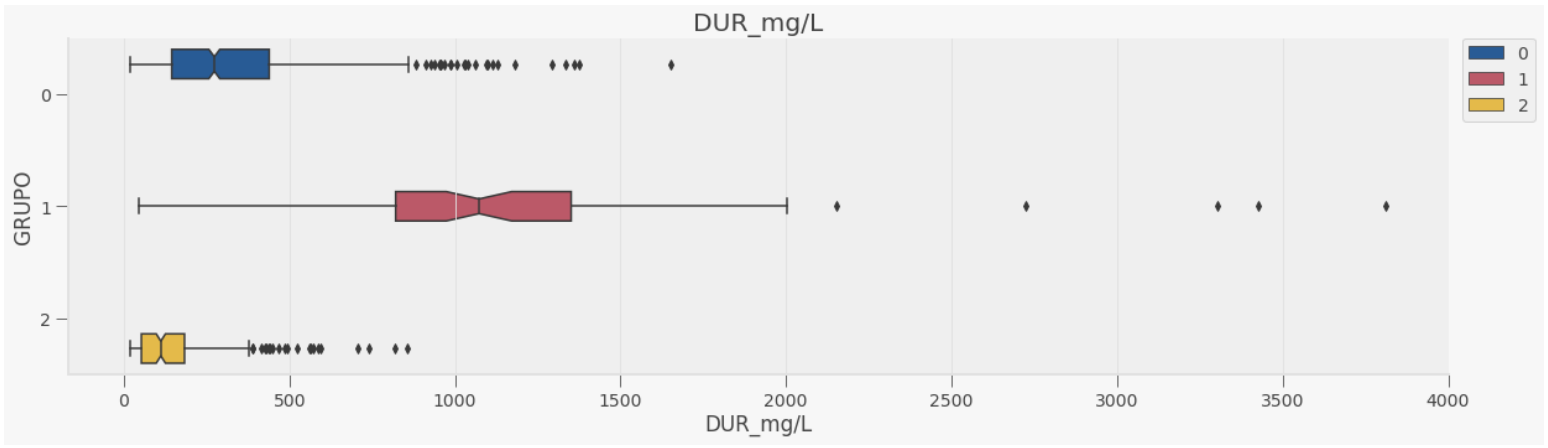

LATITUD
GRUPO
0
1
2
15.0
17.5
20.0
22.5
25.0
27.5
30.0
32.5
LATITUD

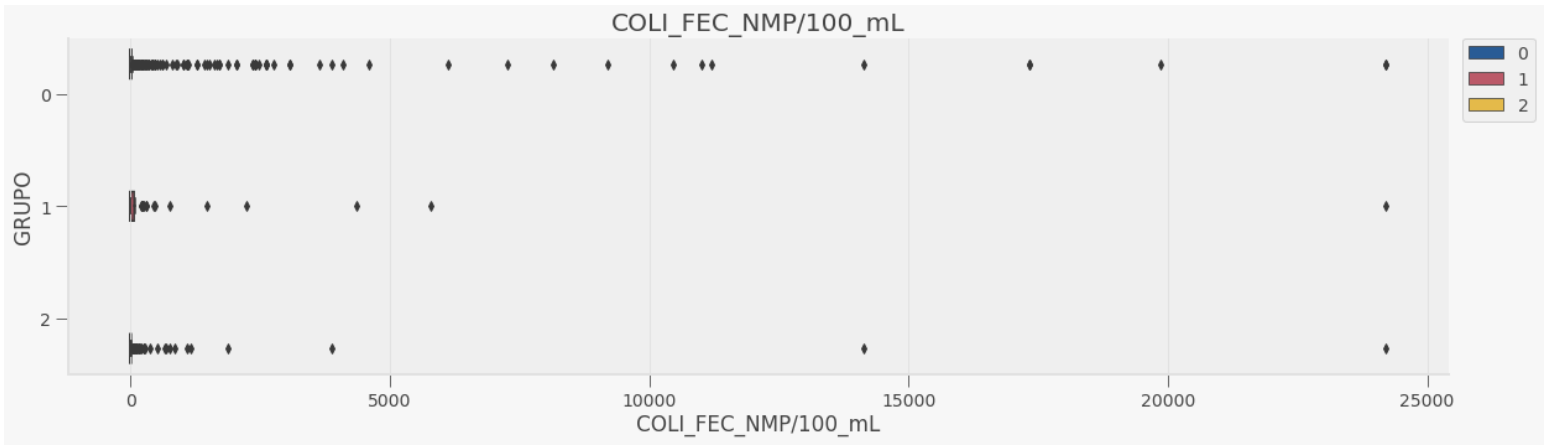

ALC_mg/L
GRUPO
0
1
2
0
200
400
600
800
1000
1200
1400
1600
ALC_mg/L

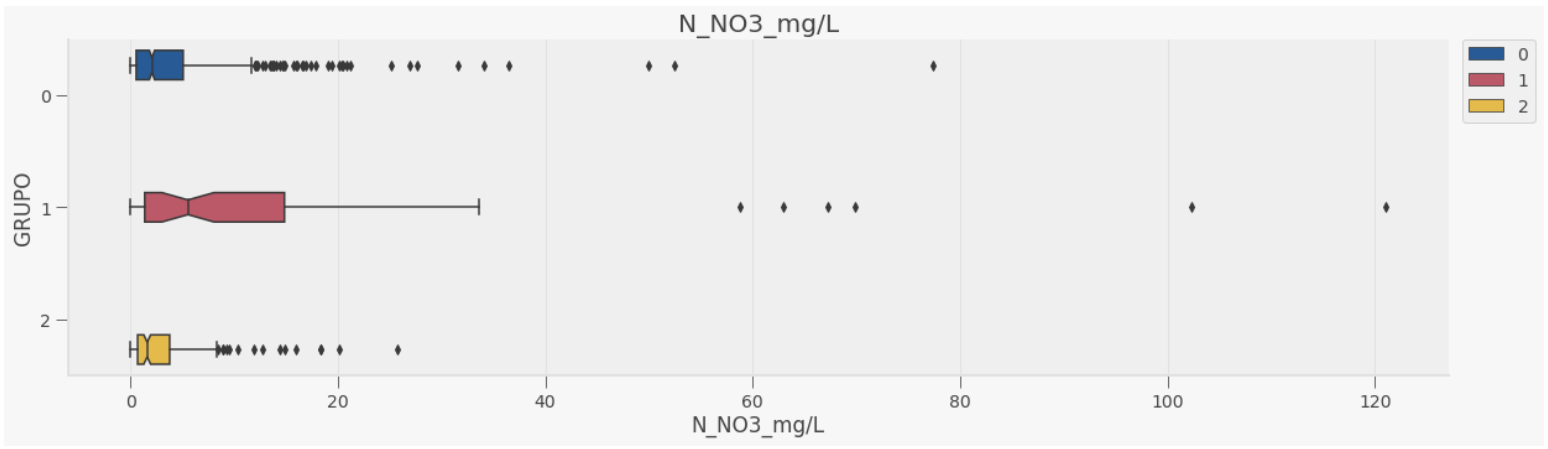

CONDUCT_mS/cm
GRUPO
0
1
2
0
2500
5000
7500
10000
12500
15000
17500
CONDUCT_mS/cm

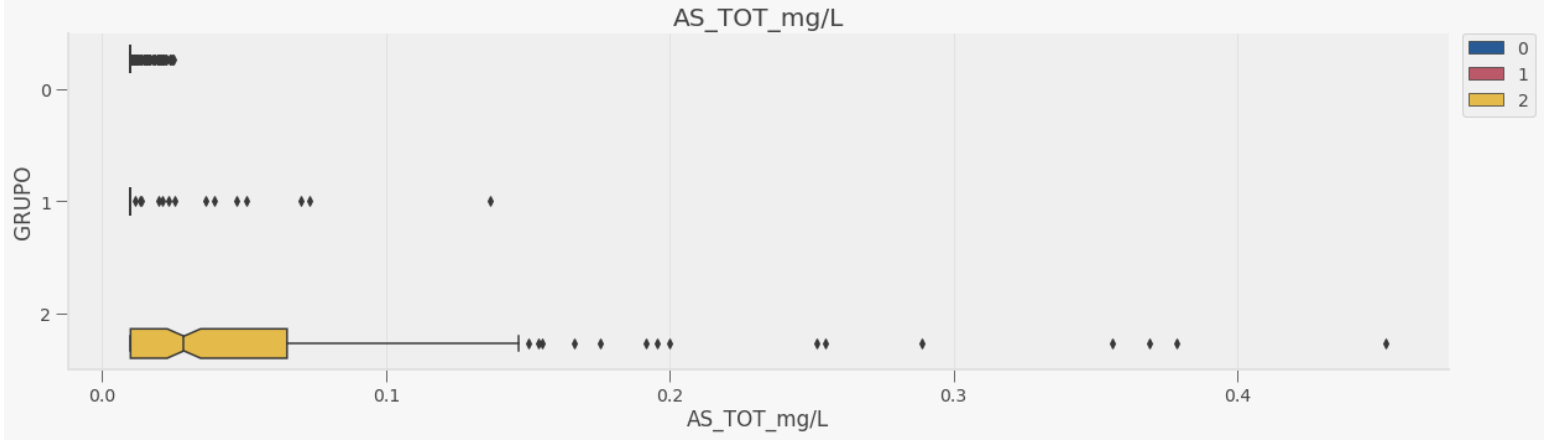

SDT_M_mg/L
GRUPO
0
1
2
0
20000
40000
60000
80000
SDT_M_mg/L

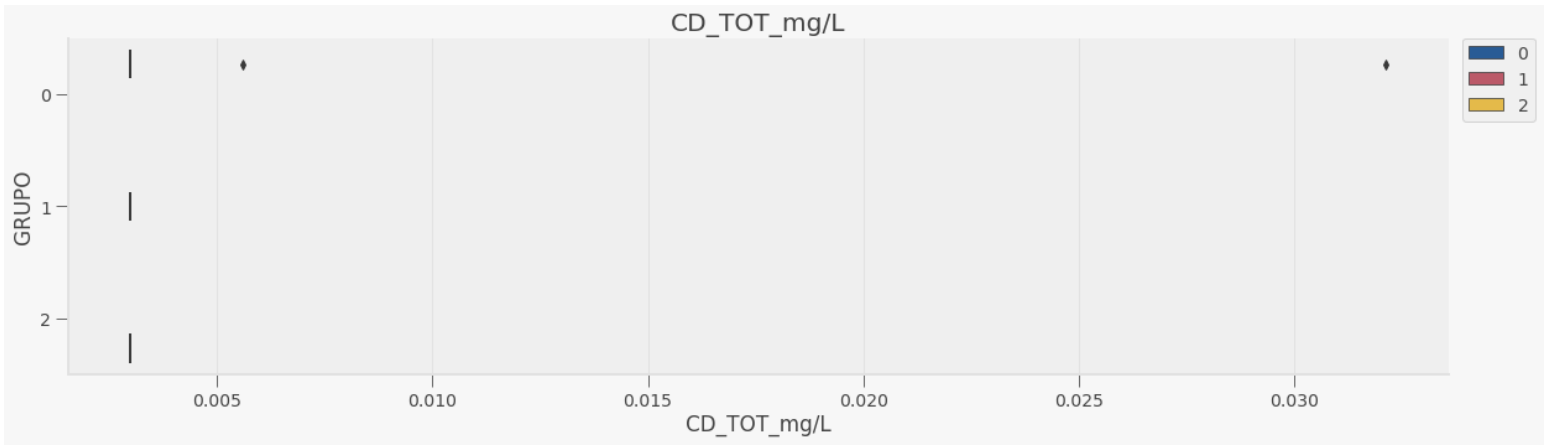

FLUORUROS_mg/L
GRUPO
0
1
2
0
5
10
15
20
25
30
35
FLUORUROS_mg/L

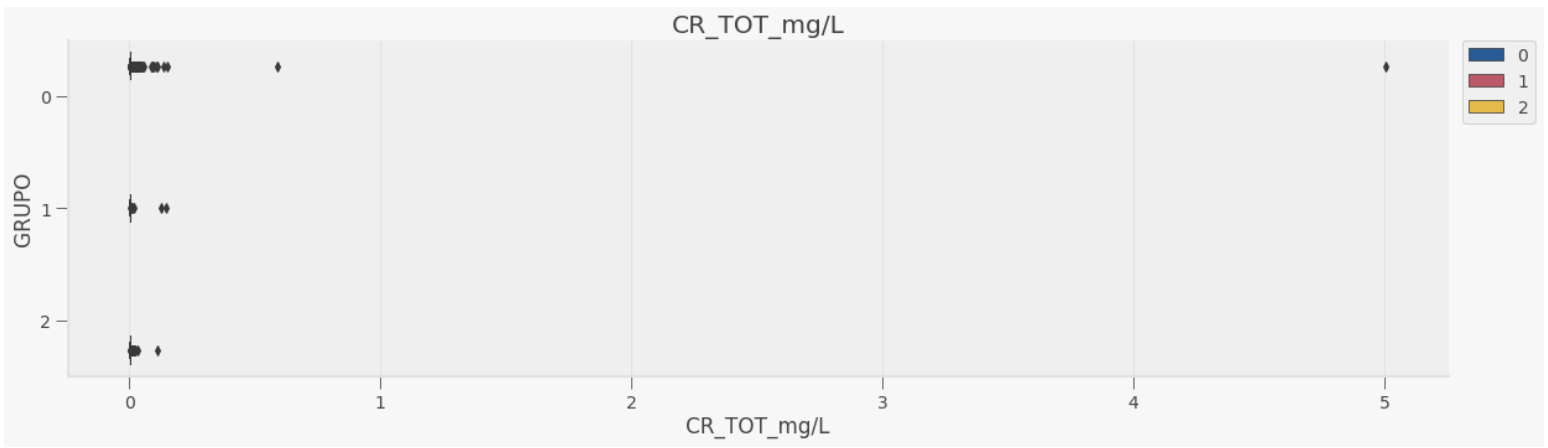

DUR_mg/L
GRUPO
0
1
2
0
500
1000
1500
2000
2500
3000
3500
4000
DUR_mg/L

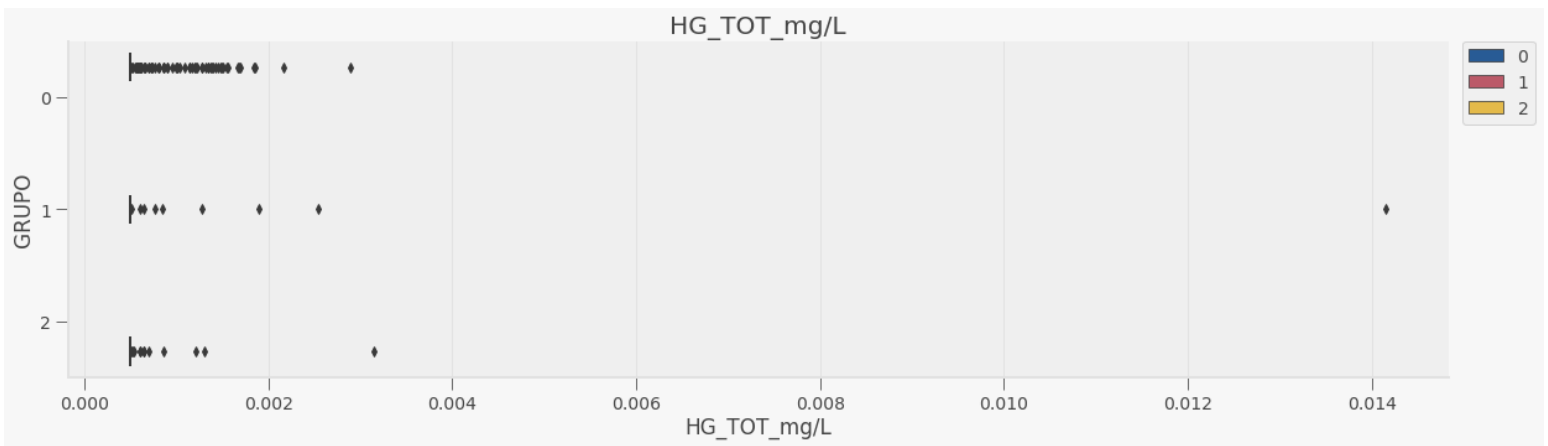

COLI_FEC_NMP/100_mL
GRUPO
0
1
2
0
5000
10000
15000
20000
25000
COLI_FEC_NMP/100_mL

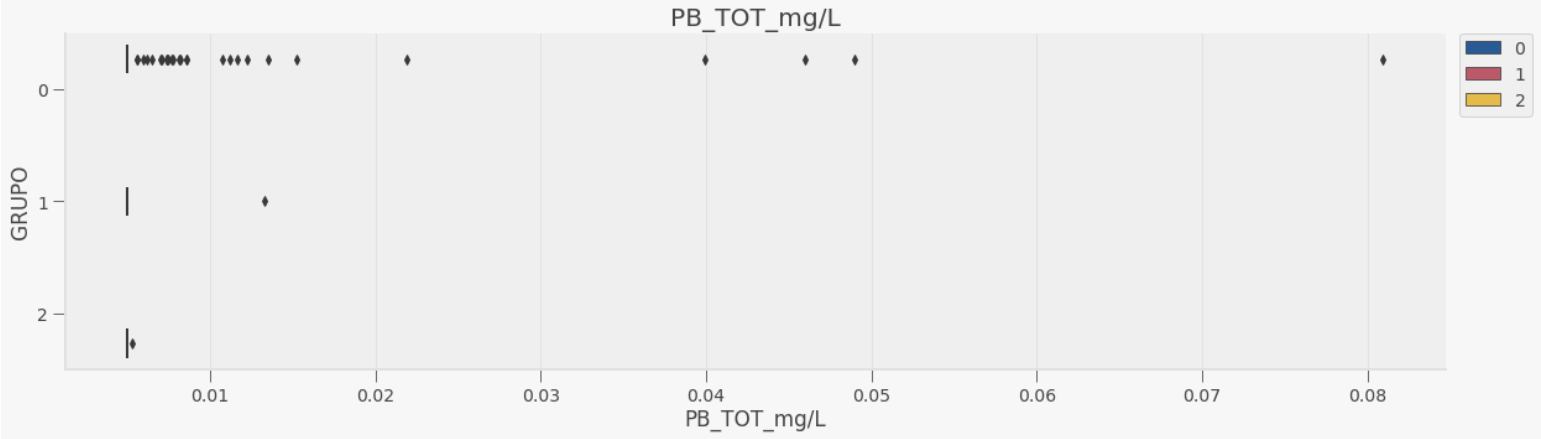

N_NO3_mg/L
GRUPO
0
1
2
0
20
40
60
80
100
120
N_NO3_mg/L

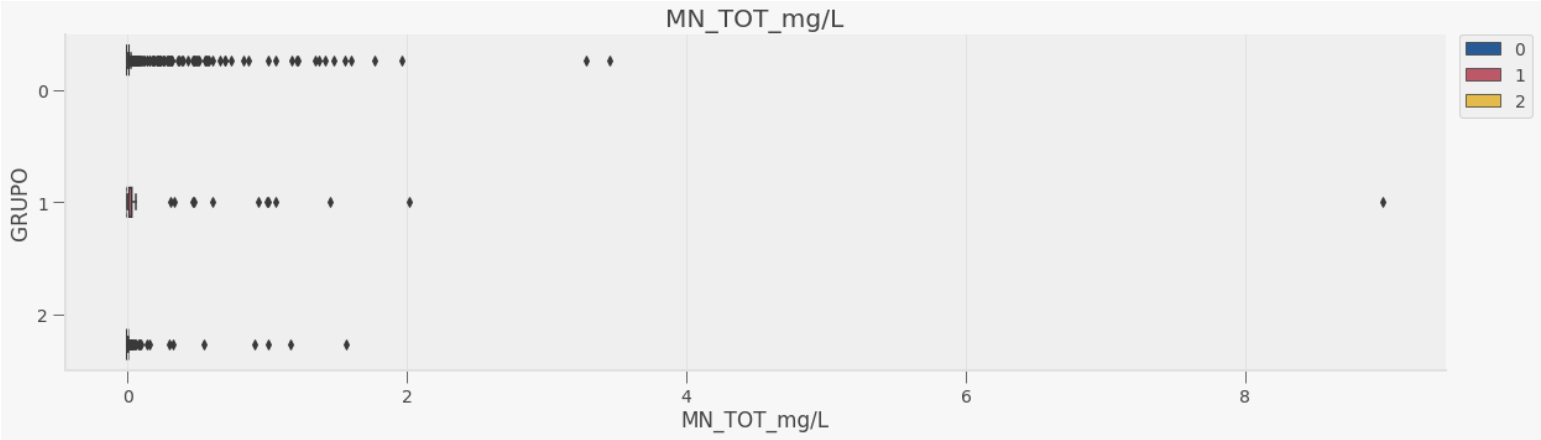

AS_TOT_mg/L
GRUPO
0
1
2
0.0
0.1
0.2
0.3
0.4
AS_TOT_mg/L

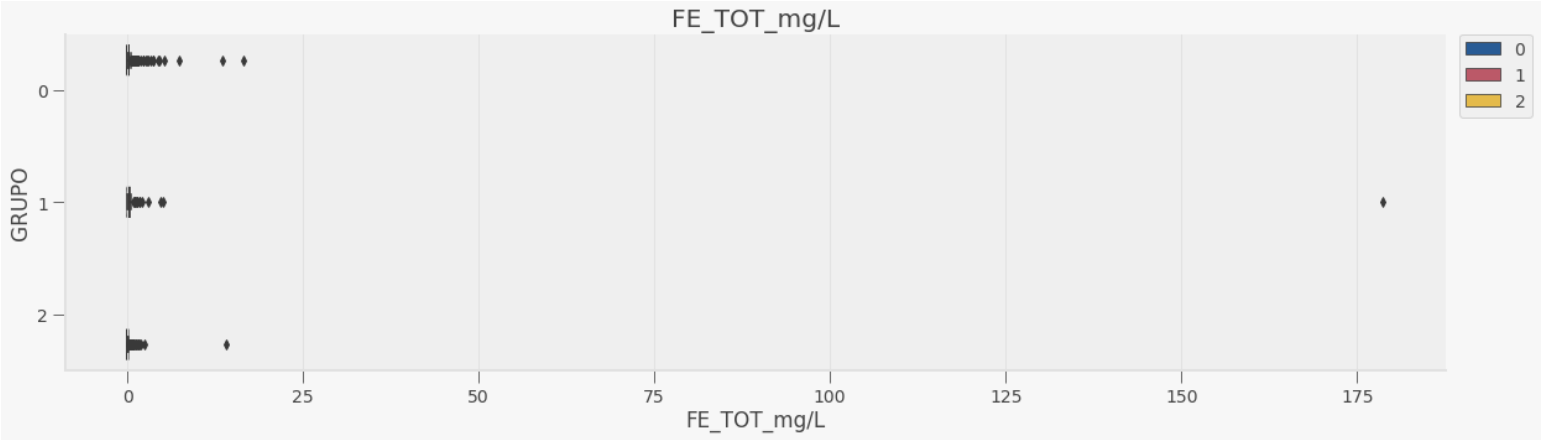

CD_TOT_mg/L
GRUPO
0
1
2
0.005
0.010
0.015
0.020
0.025
0.030
CD_TOT_mg/L


CR_TOT_mg/L
GRUPO
0
1
2
0
1
2
3
4
5
CR_TOT_mg/L


HG_TOT_mg/L
GRUPO
0
1
2
0.000
0.002
0.004
0.006
0.008
0.010
0.012
0.014
HG_TOT_mg/L



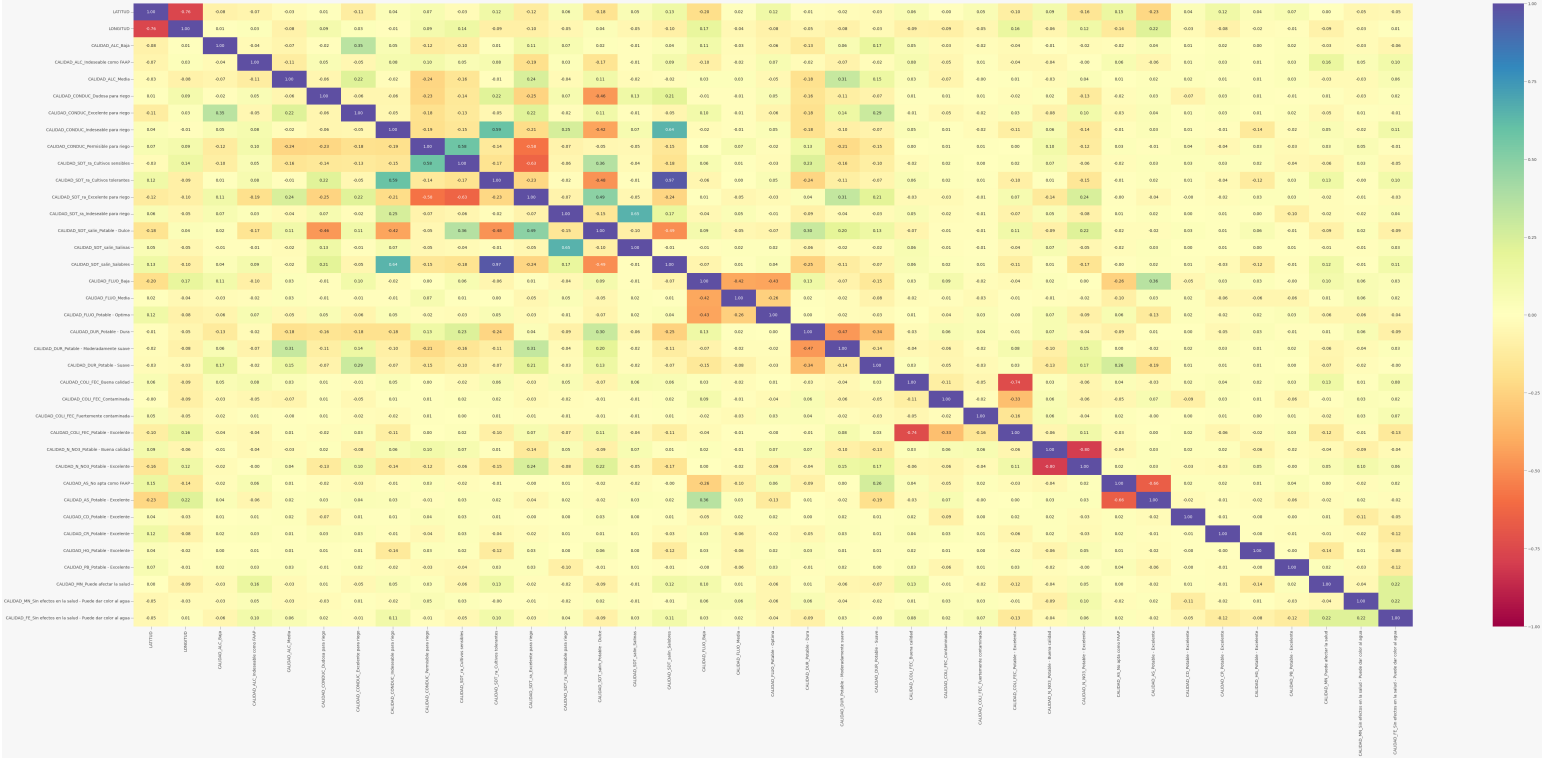




Podemos observar lo siguiente tanto de las gráficas, como de la tabla de agrupación:

1. La latitud y la longitud es relativamente similar dado que, como mencionamos en los puntos anteriores, los 3 grupos tienen puntos esparcidos a lo largo de la república, haciendo que el centroide sea mas o menos el mismo
2. El grupo 1 es el que cuenta, en promedio, con un mayor nivel de contaminantes

Hecho esto, haremos una tabla de correlación donde encontraremos la correlación entre la latitud y longitud vs. las columnas de calidad para entender si existe alguna correlación.



Con esta gráfica, podemos ver que como tal no existe una correlación fuerte entre la latitud y longitud, con respecto a la calidad del agua.

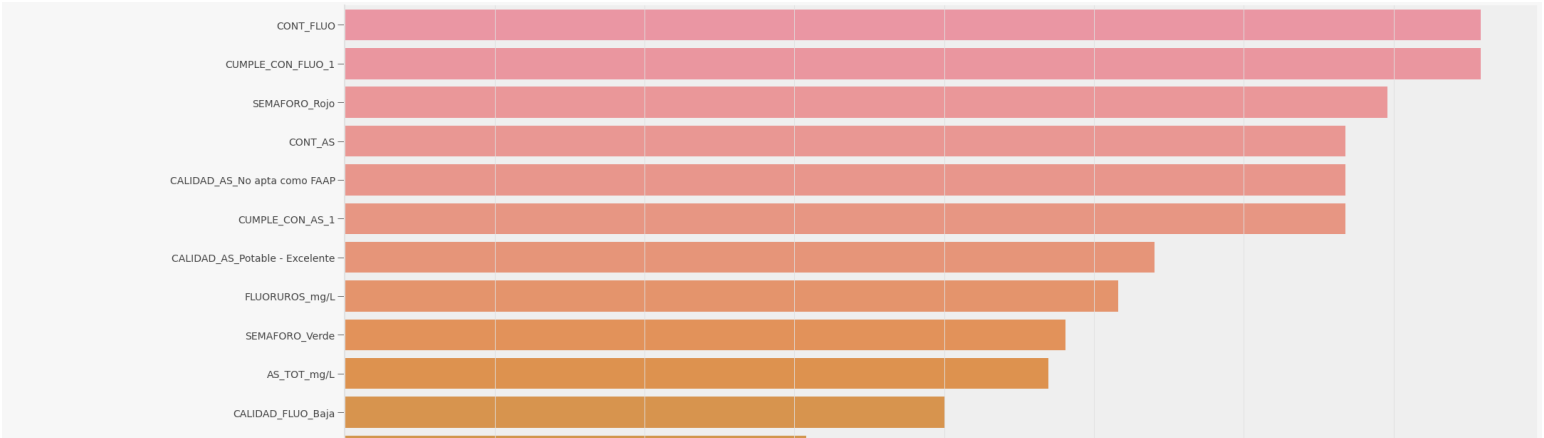
Finalmente, haremos una comparativa entre los tipos de mantos acuiferos vs. el grupo para ver si existe alguna correlación entre ellos

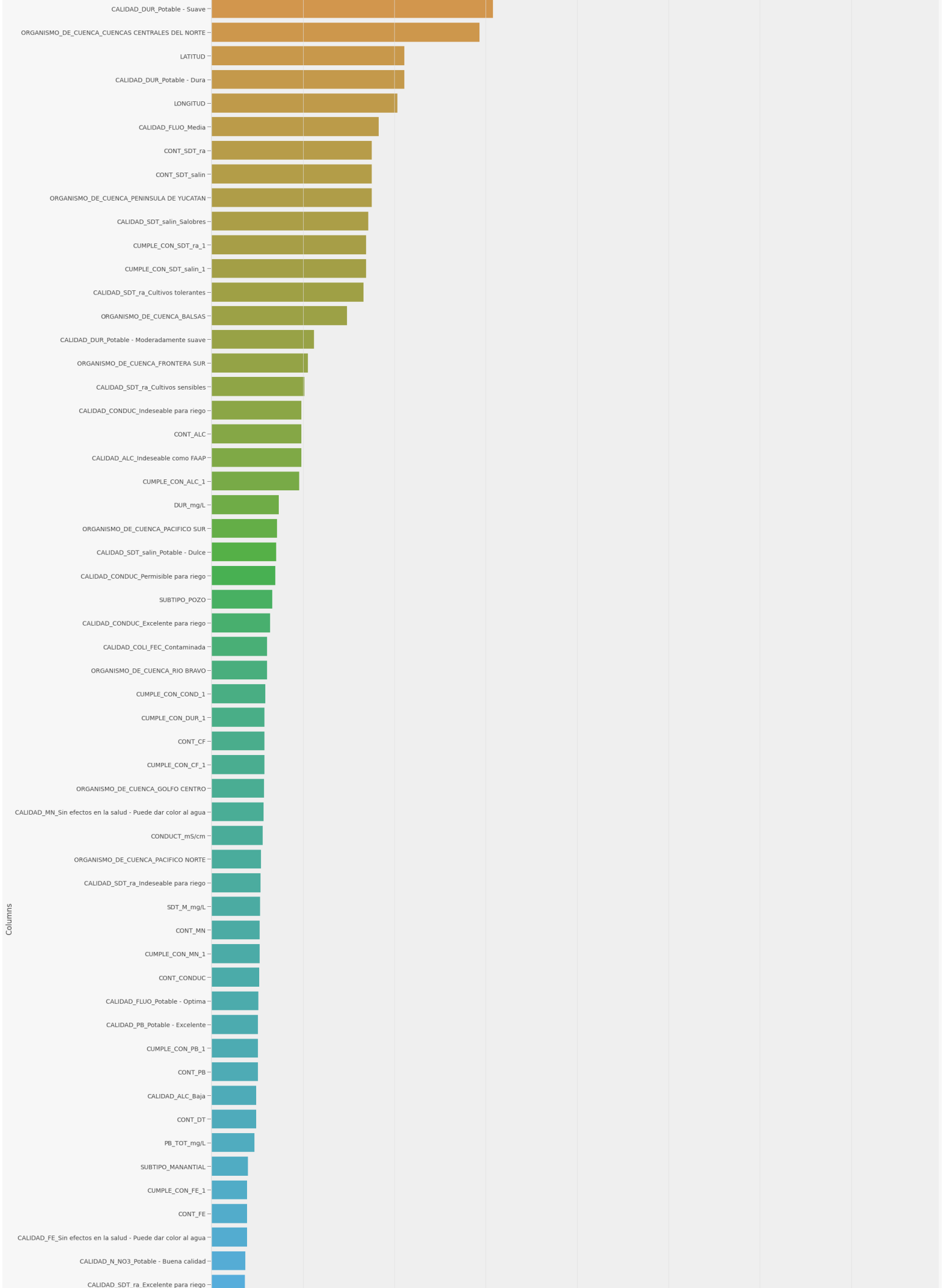
GRUPO	SUBTIPO_DESCARGA	SUBTIPO_MANANTIAL	SUBTIPO_POZO	SUBTIPO_POZO/NORIA
0	1.0	10.0	758.0	6.0
1	0.0	2.0	67.0	1.0
2	0.0	0.0	215.0	0.0

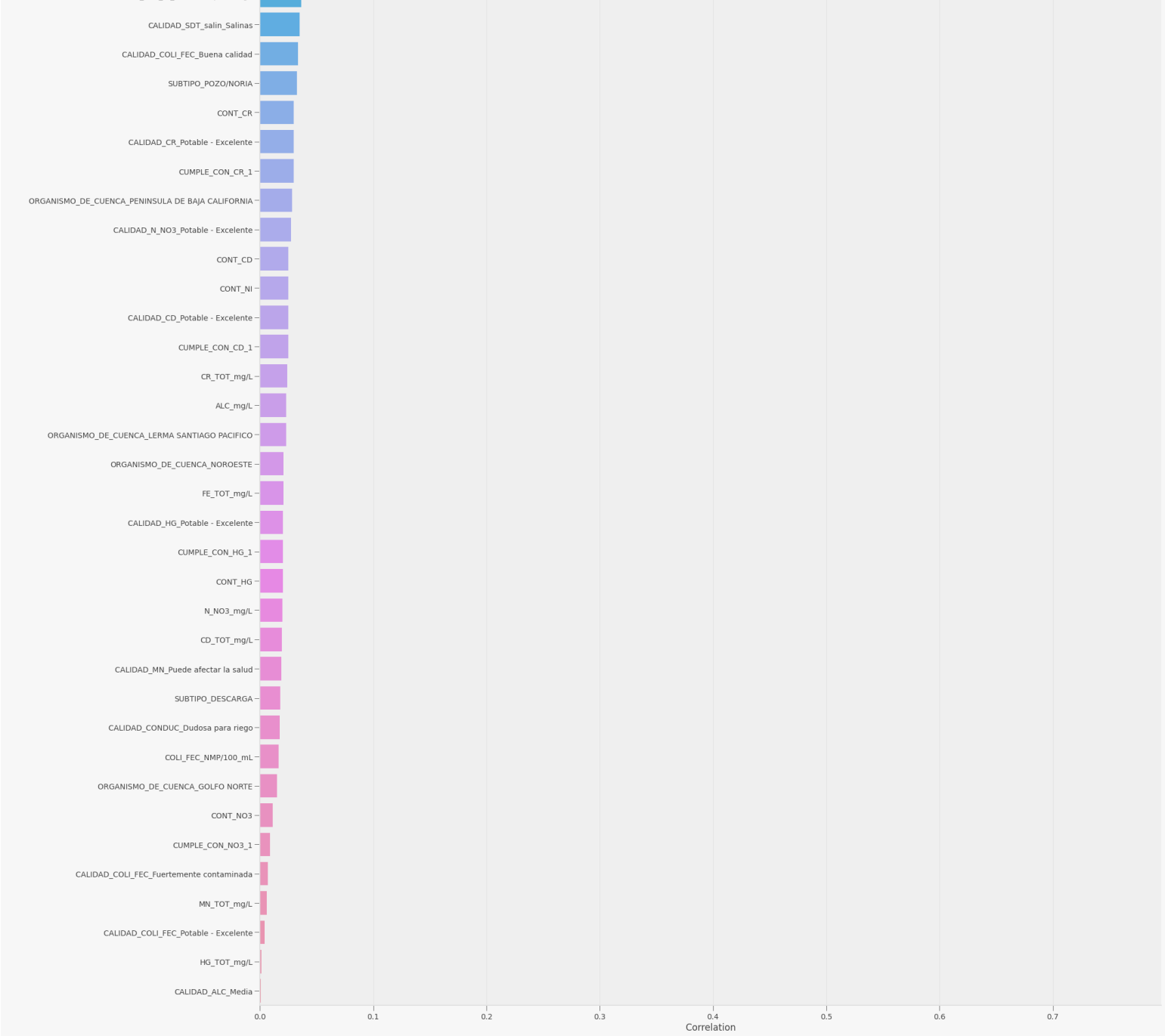
Podemos ver que estan organizados de esta manera:

- Grupo 1: Contiene 758 pozos, 10 manantiales y 6 norias
- Grupo 2: Contiene 67 pozos, 2 manantiales y 6 norias
- Grupo 3: Contiene 215 pozos

Por último, haremos una comparativa para ver si existe alguna correlación fuerte entre el grupo y cualquier otra:




Columns
CALIDAD_DUR_Potable - Suave
ORGANISMO_DE_CUENCA_CUENCAS CENTRALES DEL NORTE
LATITUD
CALIDAD_DUR_Potable - Dura
LONGITUD
CALIDAD_FLUO_Media
CONT_SDT_ra
CONT_SDT_salin
ORGANISMO_DE_CUENCA_PENINSULA DE YUCATAN
CALIDAD_SDT_salin_Salobres
CUMPLE_CON_SDT_ra_1
CUMPLE_CON_SDT_salin_1
CALIDAD_SDT_ra_Cultivos tolerantes
ORGANISMO_DE_CUENCA_BALSAS
CALIDAD_DUR_Potable - Moderadamente suave
ORGANISMO_DE_CUENCA_FRONTERA SUR
CALIDAD_SDT_ra_Cultivos sensibles
CALIDAD_CONDUC_Indeseable para riego
CONT_ALC
CALIDAD_ALC_Indeseable como FAAP
CUMPLE_CON_ALC_1
DUR_mg/L
ORGANISMO_DE_CUENCA_PACIFICO SUR
CALIDAD_SDT_salin_Potable - Dulce
CALIDAD_CONDUC_Permisible para riego
SUBTIPO_POZO
CALIDAD_CONDUC_Excelente para riego
CALIDAD_COLI_FEC_Contaminada
ORGANISMO_DE_CUENCA_RIO BRAVO
CUMPLE_CON_COND_1
CUMPLE_CON_DUR_1
CONT_CF
CUMPLE_CON_CF_1
ORGANISMO_DE_CUENCA_GOLFO CENTRO
CALIDAD_MN_Sin efectos en la salud - Puede dar color al agua
CONDUCT_mS/cm
ORGANISMO_DE_CUENCA_PACIFICO NORTE
CALIDAD_SDT_ra_Indeseable para riego
SDT_M_mg/L
CONT_MN
CUMPLE_CON_MN_1
CONT_CONDUC
CALIDAD_FLUO_Potable - Optima
CALIDAD_PB_Potable - Excelente
CUMPLE_CON_PB_1
CONT_PB
CALIDAD_ALC_Baja
CONT_DT
PB_TOT_mg/L
SUBTIPO_MANANTIAL
CUMPLE_CON_FE_1
CONT_FE
CALIDAD_FE_Sin efectos en la salud - Puede dar color al agua
CALIDAD_N_NO3_Potable - Buena calidad
CALIDAD_SDT_ra_Excelente para riego



En este orden esta la correlación entre los grupos y cada columna. Podemos ver que lo que más peso tiene es:

1. El contenido de fluor
2. Cumple con FLUO_1
3. Semaforo Rojo
4. Contiene AS
5. Calidad AS No alta como FAAP