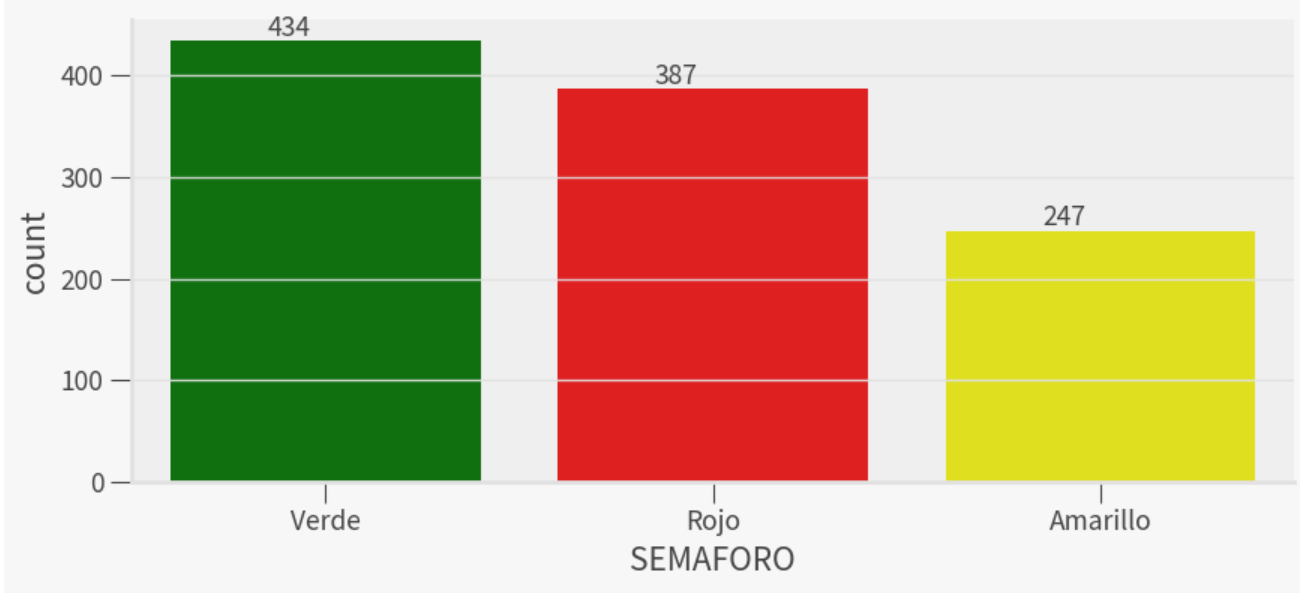
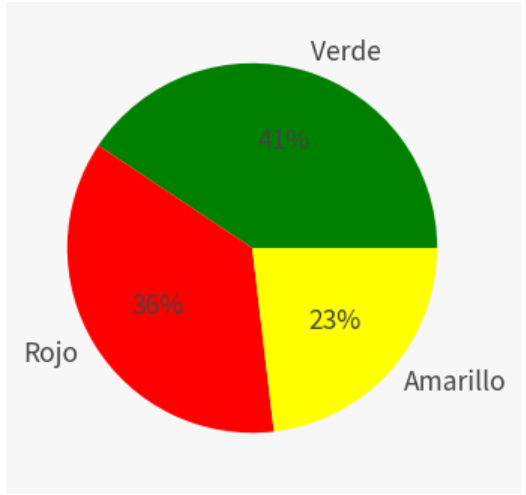
Para el siguiente análisis, utilizaremos la base de datos:

Datos\_de\_calidad\_del\_agua\_de\_sitios\_de\_monitoreo\_de\_aguas\_subterraneas\_2020.csv

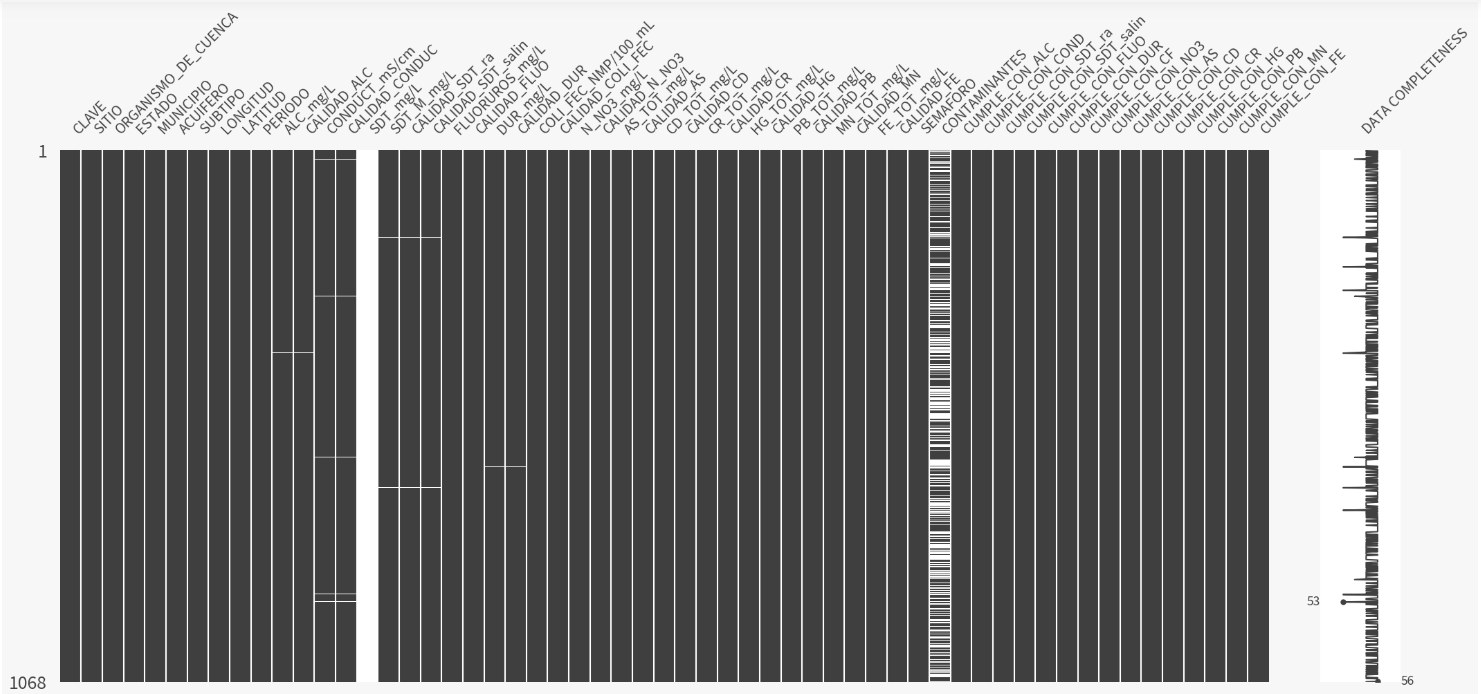
Tenemos un total de 57 columnas y 1068 datos, de las cuales 39 son categóricas y 18 son numéricas.

Adicionalmente, también sabemos que la variable ‘periodo’ solo es 2020. Así que más adelante se podría descartar ya que no aporta.

Procedemos primero a observar la variable semáforo que sería nuestra variable de salida. Tenemos la siguiente distribución:

Continuamos revisando los datos faltantes



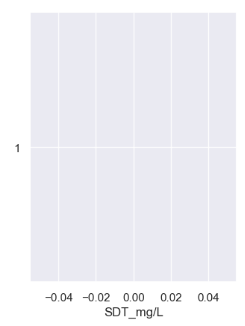
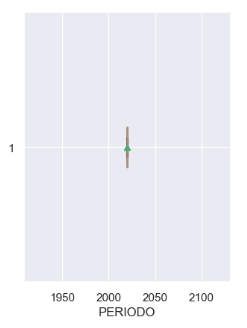
De esta gráfica podemos concluir que:

* La columna ‘SDT\_mg/L’ no tiene datos.
* Al parecer la columna ‘CONTAMINANTES’ tiene muchos datos faltantes, sin embargo vamos a analizar a más profundidad esta variable para identificar que sucede.

Con las variables continuas, generamos los histogramas y boxplot para ver la distribución de los datos.

Podemos observar que los datos se encuentran en diferentes magnitudes y escalas.

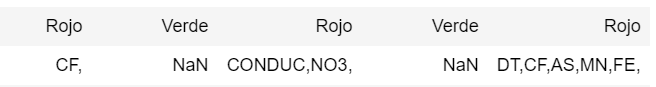
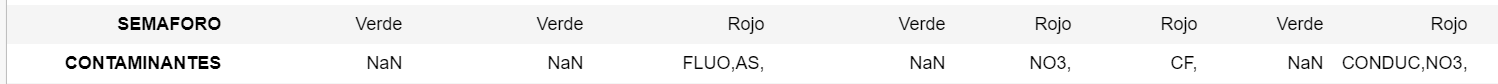
Adicionalmente, como se observó anteriormente, ‘PERIODO’ tiene un solo valor (2020) y ‘SDT\_mg/L’ no tiene valores



Se realiza una primera matriz de correlación entre las variables numéricas, se observa nuevamente que ‘PERIODO’ y ‘SDT\_mg/L’ y aportan

**Análisis de la variable CONTAMINANTES:**

Esta variable contiene una lista de contaminantes por cada toma



Se observa que efectivamente la variable SEMAFORO está directamente relacionada con la cantidad de CONTAMINANTES.

Análisis interno (borrar después): El rojo depende de la cantidad de contaminantes o del tipo de contaminante? Porque veo que hay muestras que apenas tienen 1 dato y ya es rojo.

Vamos a convertir esta lista de contaminantes por un número que representará la cantidad de contaminantes de cada muestra.

Los datos NaN es porque no tienen contaminantes, por lo tanto se convierte a 0.

Y graficamos para ver cuantos contaminantes en total tienen todas las muestras.



**Dataframe acondicionado:**

Se crea una columna SEMAFORO\_cat que convierte el color en un número de la siguiente manera:

'Verde':0,

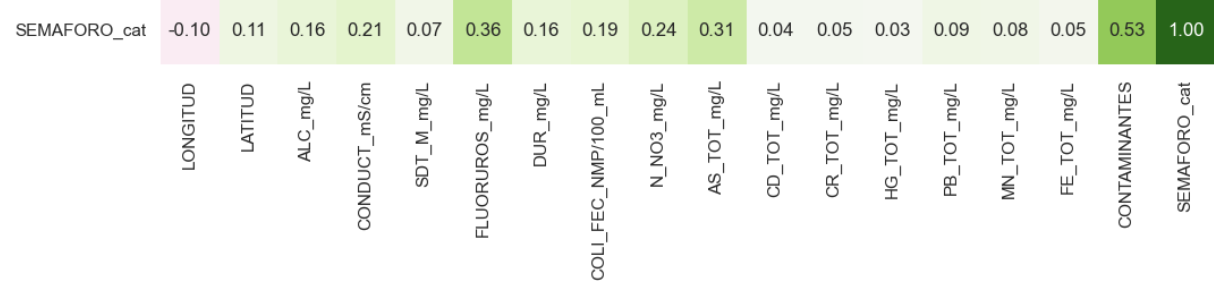
'Amarillo':1,

'Rojo':2

Adicionalmente, se eliminan las variables 'SEMAFORO', 'PERIODO', 'SDT\_mg/L' y guardamos el dataframe con un nuevo nombre.

Análisis interno: como analizamos las otras correlaciones entre variables como 'FE\_TOT\_mg/L' con 'MN\_TOT\_mg/L' por ejemplo.

Generamos nuevamente una matriz de correlación



Importante podemos destacar que la variable que mas relacionada está con ‘SEMAFORO\_cat’ es ‘CONTAMINANTES’.

VOY POR AQUI

Análisis de como impacta ORGANISMO\_DE\_CUENCA (ORGANISMO\_cat) con SEMAFORO\_cat

Análisis de como impacta ACUIFERO (ACUIFERO\_cat) con SEMAFORO\_cat

Como podemos mirar por acuifero cuantos están en rojo, amarillo, verde?