**PROYECTO de limpia** de datos del df attacks, los OBJETIVOS son:

* Saber cuales son los 5 paises con más ataques de tiburones en el año 2018.
* Cuales son las 5 actividades con más ataques de tiburones en el año 2018.

Con las premisas que mínimo tiene tiene que haber 2500 filas (el df original tiene 25723 filas) y 24 columnas, no se puede eliminar ninguna columna.

**PASOS PARA LIMPIAR MI DF**

1. Me quedo sólo con los años desde 1991 – 2023 porque aunque las gráficas serán sólo del año 2018, me interesa tener ese rango de años para futuras comparaciones.

Limpiamos la columna ‘Year’; eliminamos decimales con **math.isnan** y la ordenaremos desde el año más actual a menos con **sort\_values().**

Quedando 2609 filas y las 24 columnas.

1. Tenemos 5 columnas que hacen referencia a fechas y años, los valores están duplicados en dos de ellas, hay letras dentro de fechas…limpiamos:

* Usamos un patrón de fecha con **re.findall** para limpiar la columna ‘Case Number’.
* En las columnas ‘Case Number.1’ y ‘Case Number. 2’ lo sustituimos por ‘valor duplicado’ con la **igualdad** (df['Case Number.1'] = 'valor duplicado')

1. Filtramos las filas donde la columna "Species" contiene la palabra ‘shark’ y dejamos NaN en el resto de celdas.
2. Limpiamos la columna ‘Age’, dejando en blanco las celdas que no tengan número de 2 ó 1 dígito.
3. Limpiamos la columna ‘Activities’ dejando en blanco lo que no tiene sentido con **replace**.
4. Limpiamos la columna ‘Area’, aplicando una expresión regular y creando una **máscara booleana** para las condiciones 'miles' y 'km', dejando solo el nombre de las áreas; también eliminamos los números.
5. Hacemos limpia general: sustituimos valores ‘NaN’ por espacios en blanco con **replace**; este método lo utilizaremos muy a menudo para limpiar muchas celdas.También dejamos todos los valores de las celdas del df a la izquierda.
6. Eliminamos las filas donde en la columna ‘Year’, ‘Case Number’ y ‘Date’ no haya valores con **drop().**
7. Hacemos gráficos de barras y de tarta con **matplotlib.pyplot as plt.**

Iremos haciendo más limpias probando diferentes métodos para probar ☺