El ataque de los tiburones

En este ejercicio vamos a tratar con el dataset "Global Shark Attacks" que recoge una base de datos de incidentes humano-tiburón. Más información sobre el dataset aquí: http://www.sharkattackfile.net/index.htm (http://www.sharkattackfile.net/index.htm)

1. Descarga de datos

Hay una descripción en Kaggle, esta es la página Web: https://www.kaggle.com/teajay/global-shark-attacks (https://www.kaggle.com/teajay/global-shark-attacks)

Y el enlace a la descarga directa es el siguiente: https://www.kaggle.com/teajay/global-shark-attacks/downloads/global-shark-attacks.zip (https://www.kaggle.com/teajay/global-shark-attacks/downloads/global-shark-attacks.zip) Nótese que kaggle requiere identificarse, por lo que no pueden descargarse sin autentificación los ficheros.

Se pide lo siguiente:

- Cargad el fichero en un `DataFrame` de `pandas` .
- Comparad las columns cargadas con las del fichero Excel. ¿Hay información que deberíamos quitar? Nota: con `columns` podemos obtener la lista de columnas del DataFrame.

2. Observación de valores nulos

El siguiente paso es observar los datos, los valores que toman y si hay valores nulos.

- Identificad las columnas en las que hay valores nulos y las que no y cuántas en cada una. Por ejemplo, parece que las hay en la columna Species. Nota: usar `isnull()`
- Para las columnas con información categórica, observar las etiquetas de cada categoría, por ejemplo, las especies de tiburones o el tipo de incidente.
- Concretamente, obtener las 5 especies de tiburón con más incidentes registrados en la base de datos. Nota: Puede hacers con `groupby()` o dividiendo el array por especies, y contando cada subconjunto.

3. Más valores nulos

Vamos a examinar la ocurrencia de eventos por años y países. Para ello:

- Queremos utilizar como índice (país, año) primero. Probar a establecerlo con `set_index()` . ¿Hay algún problema de valores nulos en el índice? ¿Son NA o también hay años no válidos? ¿Cómo se pueden eliminar?

- Obtener la cuenta de los incidentes en USA en 2014 y 2015. Utilizad `loc` e `iloc` en la misma expresión.

4. Distribución y correlaciones (i)

Ahora queremos observar la distribución de los incidentes, primero por países y luego por años.

- Mostrar los tres países que tienen más incidentes.
- Dibujar la distribución por años de los incidentes, señalar el año en el que más se han producido.
- Para los dos países con más incidentes, mostrar su evolución por años.

5. Distribución y correlaciones (ii)

A continuación, queremos examinar el sexo y la edad de los implicados en incidentes.

Mostrad la distribución de los incidentes por sexo, e intentad obtener la media de la edad de los implicados en los incidentes.

Observad:

- Valores nulos o valores que parecen incorrectos.
- El tipo de los valores de la columna "Age". Si fuese necesario cambiarlos, buscar formas de convertir de valores de cadena a valores numéricos en pandas.

6. La curiosidad del data scientist

Finalmente, ¿qué más se puede examinar en el dataset que sea interesante?