Pandas - Project Ironhack Data Bootcamp

Miguel Ángel Ávalos

Data Part Time Barcelona Dic 2019

Project Description

En este proyecto se basa en utilizar técnicas de data cleaning y data manipulation mediante el uso de la librería pandas de Python para limpiar un dataset basado en datos de ataque de tiburones a nivel mundial y con varios años de registros. Al tratarse de un dataset con tantos registros y presenta numerosos campos sin valores (null values) y con filas y columnas duplicadas o con información poco relevante para un posterior análisis.

Dataset

El dataset original proviene de la página kaggle.com y se encuentra bajo el nombre de global shark attacks. En dicho dataset podemos encontrar registros ataques de tiburones de varios años, diferenciados por localización, sexo, edad, actividad realizada, heridas causadas.
Las columnas que presenta el dataset son las siguientes:

- `'Case Number'` : identificador del registro,
- `'Date'` : fecha del ataque,
- `'Year'` : año del ataque,
- `'Type'` : tipo de ataque (provocado o no)
- `'Country'` : país donde se ha producido el ataque,
- `'Area'` : subzona del país,
- `'Location'` : zona más reducida,
- `'Activity'` : actividad que realizaba el/los accidentado/s
- `'Name'` : nombre de la persona relacionada con el ataque,
- `'Sex '` : género,
- `'Age'` : edad/es
- `'Injury'` : tipo de heridas sufridas durante,
- `'Fatal (Y/N)'` : si este fue mortal o no,
- `'Time'` : hora a la que se produjo,
- `'Species '` : tipo de tiburón,
- `'Investigator or Source'` : origen de la referencia sobre el ataque.
- `'pdf'` : documento sobre la referencia del ataque,
- `'href formula'` : link del documento sobre el ataque,
- `'href'` : columna duplicada (`'href formula'`)
- `'Case Number.1'` : columna duplicada (`'Case Number'`),

- `'Case Number.2'` : columna duplicada (`'Case Number'`),
- `'original order'` : índices del dataset,
- `'Unnamed: 22'` : sin información
- `'Unnamed: 23'` : sin información

Trabajo realizado

La mayor parte del tiempo se ha dedicado a la exploración y comprensión de las diferentes columnas para discernir cuáles eran susceptibles de ser tratadas y cuáles contenían información duplicada.
Una vez determinadas cuales eran las columnas y filas con información duplicada, se han realizado como tareas principales, la transformación de los null values, tratando de recuperar registros en algunos casos, la exclusión de registros de *Type* Invalid, ya que la información que contenía información sobre ataques no relacionados con tiburones.
Además, se han filtrado y categorizado los datos de *Injuries* para agruparlos el tipo de herida en función de su gravedad. Finalmente, se ha extraído de la columna *Date*, el mes en el que se ha producido los ataques para explorar si puede ser una variable que pueda estar relacionada con los ataques de tiburones.

Resultados

El dataset final se ha guardado con el nombre de *clean_df.csv*.