TP Base de datos: ETL

Fecha: 9 de Junio 2024

Autores

miembro	padrón
Marcos Vrljicak	96693
Martín del Castillo	106296
Matias Valmaggia	105621
Bruno Contreras	105634
Ignacio Sugai	109549
Nicolás Carreño	106442

Introducción

Para la resolución de este trabajo se utilizó un set de datos obtenido de Kaggle. El set de datos detalla las medallas olímpicas obtenidas por diferentes naciones a lo largo del tiempo, tanto en los juegos olímpicos de verano como los de invierno.

Link al set de datos: https://www.kaggle.com/datasets/shreyaskeote23/national-olympic-committee-2022-medals-csv/data

Este set de datos se sanitizó y adaptó según criterios que se detallarán más adelante, para luego cargarse en una base de datos SQLite. Tanto el saneamiento del set de datos como la carga en SQLite se hicieron mediante scripts en Python. ChatGPT asistió en el desarrollo de ambas rutinas.

Análisis del set de datos

En cuanto a los numéricos de la cantidad de medallas y atletas del set de datos, había algunas irregularidades en el tipo de datos. Los valores mayores a 1000 se mostraban entrecomillados y con una coma `(,)` para marcar las centenas. Para solucionar esto se le aclaró a la función read_csv que el caracter `,` se utiliza para las centenas.

El dataset original contenía columnas para el total de las medallas tanto para los juegos de verano como los de invierno. Para mantener la tabla normalizada, se eliminaron estas columnas. De ser necesario, podrían obtenerse en base a los campos individuales.

No hay valores nulos. Sí hay valores 0 pero estos son legítimos, ya que simplemente significa que el país no tiene medallas del tipo indicado.

El campo que debió adaptarse un poco más fue el campo `team`. Este campo en el set de datos original se conforma del nombre del país, seguido por un código de país entre paréntesis, seguido en algunos casos por una serie de códigos entre corchetes cuyo significado no es aclarado por el autor del dataset.

Para hacer que el set de datos sea más uniforme y tenga una estructura final más ordenada en la base de datos, el campo `team` se modificó de la siguiente manera:

- El nombre del país se pasó a un campo nuevo llamado `country`
- El código del país se pasó a un campo nuevo llamado `country_code`. Además, se le sacaron los paréntesis.
- Los códigos entre corchetes fueron completamnete quitados del dataset

Por último, se quitó una columna sin nombre que contenía un valor autonumérico que no cumplía ninguna función aparente.

Carga a base de datos SQLite

Para la carga a la base de datos SQLite simplemente se utilizó la libreria sqlite3 de Python junto con pandas. La funcion to_sql de pandas recibe el objecto conexión de sqlite y escribe sobre él el dataframe entero. La única modificación que se hace al dataset en este paso es el nombre de las columnas (campos) para que sean más legibles. Abreviaciones como SOG se modificaron por términos explicitos como summer_olympics, y algunas siglas redundantes se quitaron.

Haciendo queries al archivo resultante olympic_medals_2022.db se puede verificar fácilmente que la información está limpia y sana. Por ejemplo, se puede fácilmente hacer un select *, ya que el set de datos tiene tan sólo 150 registros.