M2.851 - Tipología y ciclo de vida de los datos - Práctica 2
Betancor Sánchez, Manuel - Aula 2
Navalón Hernández, María Dolores - Aula 1

Práctica 2: Limpieza y validación de los datos.

Índice:

1. Descripción del dataset.

En esta práctica vamos a analizar el impacto que tiene la celebración de los Juegos Olímpicos en los países anfitriones a nivel deportivo, usando como referencia las medallas conseguidas, así como la influencia supranacional, evaluándose también a nivel de continente.

Para ello utilizaremos la información suministrada por la web kaggle, la cual recopila información relativa a los deportistas que han participado en los Juegos Olímpicos modernos. El enlace de descarga, en el cual se pueden obtener dos ficheros csv, es el siguiente:

https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results

Para poder relacionar el país de los deportistas con el continente al que pertenece, vamos a complementar los datos con el dataset obtenido por el siguiente enlace:

https://www.kaggle.com/statchaitya/country-to-continent

Y para conocer qué país organiza cada edición de los Juegos Olímpicos hemos encontrado este otro dataset que relaciona cada ciudad sede con su país:

https://www.downloadexcelfiles.com/wo_en/download-excel-file-list-olympic-host-cities

La estructura de los diferentes ficheros obtenidos es la siguiente:

Estructura del fichero athlete_events.csv:

Columna	Descripción	Tipo
ID	Identificador único para cada atleta	ENTERO
Name	Nombre del atleta	TEXTO
Sex	Sexo. Valores M o F	TEXTO
Age	Edad	TEXTO
Height	Altura en centímetros	TEXTO
Weight	Peso en kilogramos	TEXTO
Team	Nombre del equipo	TEXTO

NOC	Código de 3 letras que identifica al Comité Olímpico del país	TEXTO
Games	Año y tipo de juegos	TEXTO
Year	Año de celebración	ENTERO
Season	Tipo de juego. Valores Summer (Verano) y Winter (invierno)	TEXTO
City	Ciudad organizadora de los juegos	TEXTO
Sport	Deporte	TEXTO
Event	Modalidad del deporte	TEXTO
Medal	Medalla conseguida. Valores Gold (Oro), Silver (Plata), Bronze (Bronce) o NA	TEXTO

Estructura del fichero countryContinent.csv:

Columna	Descripción	Tipo
country	Nombre del país	TEXTO
code_2	Código de país (2 letras)	TEXTO
code_3	Código del país (3 letras)	TEXTO
country_code	Código numérico del país	INTEGER
iso_3166_2	Código ISO 3166-2 del país	TEXTO
continent	Continente	TEXTO
sub_region	Región	TEXTO
region_code	Código numérico de la región	INTEGER
sub_region_code	Código numérico de la sub-región	INTEGER

Estructura del fichero list-host-cities-olympic-943j.csv:

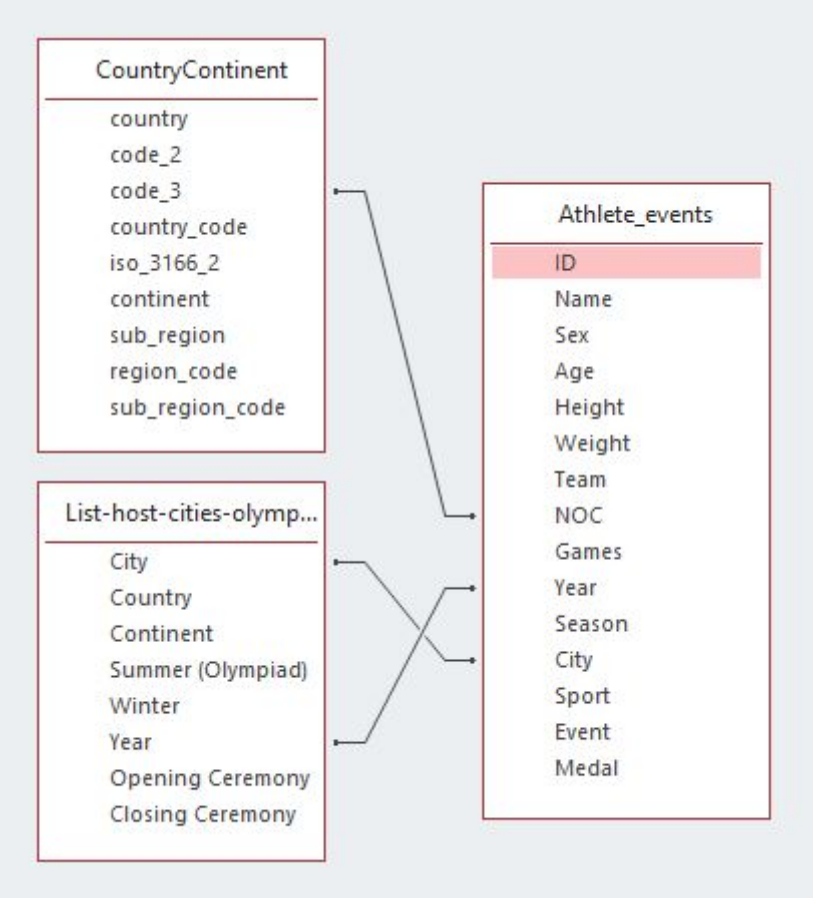
Columna	Descripción	Tipo
City	Nombre de la ciudad	TEXTO
Country	Nombre del país	TEXTO
Continent	Nombre del continente	TEXTO
Summer (Olympiad)	Edición de las Olimpiadas de verano	TEXTO
Winter	Edición de las Olimpiadas de invierno	TEXTO
Year	Año	INTEGER
Opening Ceremony	Fecha de la ceremonia de apertura	FECHA
Closing Ceremony	Fecha de la ceremonia de clausura	FECHA

Para la carga de los diferentes datasets se ha utilizado el siguiente código:

```
atletas <- read.csv("athlete_events.csv", encoding="utf-8")
paises <- read.csv("countryContinent.csv", encoding="utf-8")
ciudades <- read.csv("list-host-cities-olympic-943j.csv”, encoding="utf-8")
```

2. Integración y selección de los datos de interés a analizar.

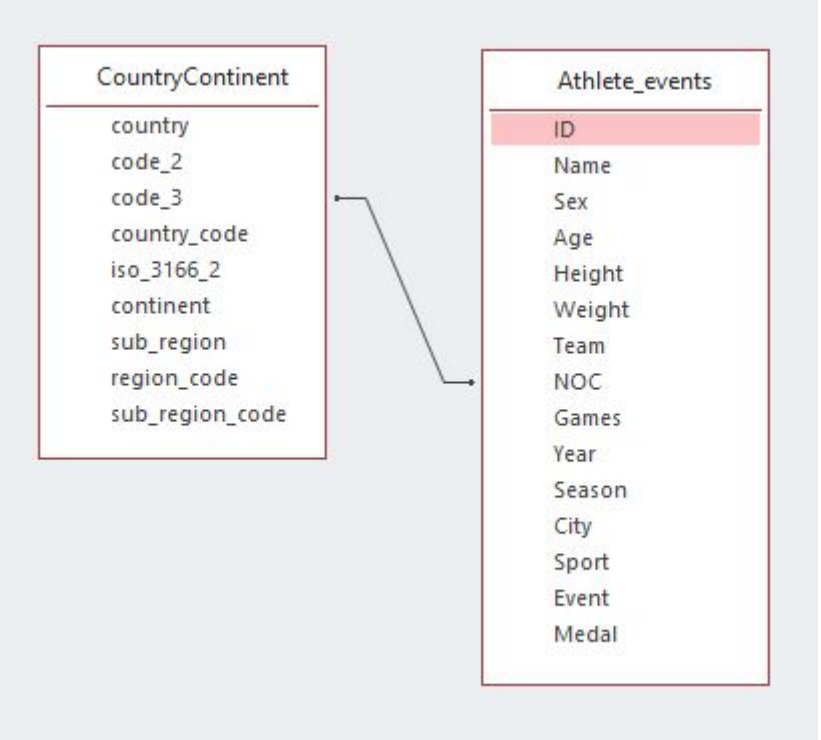
Las relaciones entre los diferentes datasets se puede resumir en el siguiente esquema:



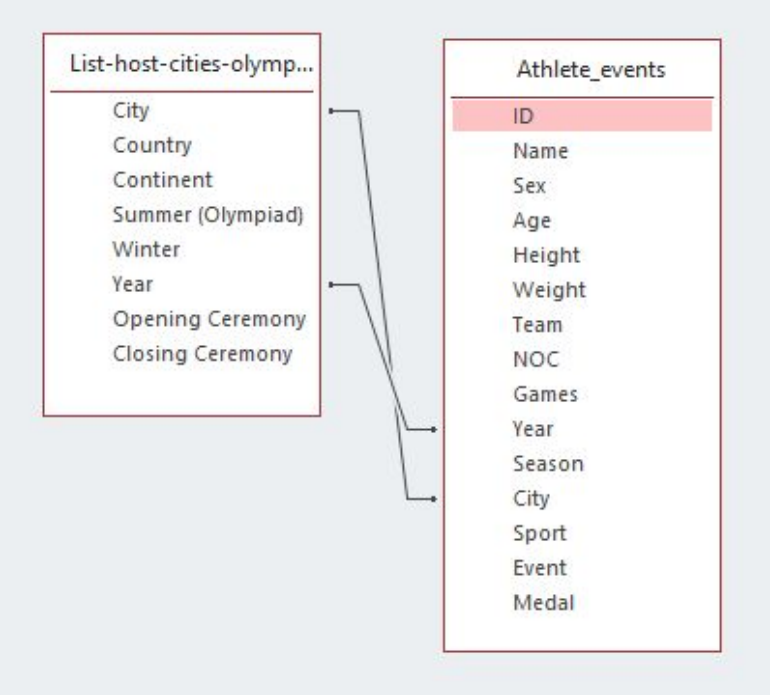
Analizando cada par de relaciones podemos comenzar con el país y continente de cada atleta, datos básicos para conocer el número de medallas obtenidas por cada país en los eventos, y que es la pieza fundamental de este trabajo.

Para conocer el país podemos relacionar el código del comité organizador (NOC) del csv de Athlete_events con la variable code_3 del csv CountryContinent. A través de este último dataset podemos ver tanto el país, en la variable country, como el continente, en la variable continent.

El siguiente gráfico refleja las relaciones citadas en los párrafos anteriores:



En relación con la ciudad anfitriona enlazamos el campo City del repositorio Athlete_events con el campo City del csv List-host-cities-olympic-943j, así mismo para seleccionar el evento concreto se enlazará también el año, uniendo los campos Year de ambos repositorios. Desde List-host-cities-olympic-943j podemos obtener tanto el nombre del país como el continente de la ciudad organizadora, utilizando los campos Country y Continent..



Los datos a utilizar por los siguientes ficheros en nuestro análisis se limitarán, centrándonos en el objetivo a conseguir, a los siguientes campos:

Fichero athlete_events.csv:

Columna	Descripción	Tipo
NOC	Código de 3 letras que identifica al Comité Olímpico del país	TEXTO
Year	Año de celebración	ENTERO
Season	Tipo de juego. Valores Summer (Verano) y Winter (invierno)	TEXTO
City	Ciudad organizadora de los juegos	TEXTO
Sport	Deporte	TEXTO
Event	Modalidad del deporte	TEXTO
Medal	Medalla conseguida. Valores Gold (Oro), Silver (Plata), Bronze (Bronce) o NA	TEXTO

Fichero countryContinent.csv:

Columna	Descripción	Tipo
country	Nombre del país	TEXTO
code_3	Código del país (3 letras)	TEXTO
continent	Continente	TEXTO
sub_region	Región	TEXTO

Fichero list-host-cities-olympic-943j.csv:

Columna	Descripción	Tipo
City	Nombre de la ciudad	TEXTO
Country	Nombre del país	TEXTO

Continent	Nombre del continente	TEXTO
Year	Año	INTEGER

El código utilizado para esta selección de columnas ha sido el siguiente:

```
atletas <- atletas[,c("NOC","Year","Season","City","Sport","Event","Medal")]
paises <- paises[,c("country","code_3","continent","sub_region")]
ciudades <- ciudades[,c("City","Country","Continent","Year")]
```

3. Limpieza de los datos.

Encontramos un caso grave dentro del csv list-host-cities-olympic-943j, ya que en los años en los que se han efectuado dos competiciones olímpicas, por coincidir invierno y verano en el mismo ejercicio, en el conjunto de datos sólo pone el año en el primer evento ocurrido, dejando como NA el posterior.

Siguiendo el manual de Calvo M., Subirats L. y Pérez D., en este tipo de circunstancias, cuando los datos son pocos y conocidos, lo mejor es una modificación manual.

En vez de realizar este cambio a mano hemos optado por solucionarlo a través de código, ya que después de analizar el dataset encontramos un patrón para la sustitución de valores que implementandolo con código corregía el error de una manera más simple y rápida.

El código desarrollado para la eliminación de los datos nulos es el siguiente:

```
for(i in 3:nrow(ciudades)){
  if (is.na(ciudades[i,"Year"])) {
    ciudades[i,"Year"] <- ciudades[i-1,"Year"]
  }
}
```

En el mismo conjunto de datos también pudimos ver que los tres últimos registros no disponían de información, eliminandolos también a través de código.

```
ciudades <- ciudades[1:(nrow(ciudades)-3),]
```

Analizando los diferentes dataset también podemos ver que no coinciden los códigos, en algunos países, de los Comités Olímpicos Organizadores (NOC) y de los países.

Por ejemplo, en Alemania el NOC es GER, mientras que el código code_3 del país es DEU.

Es por ello que para solucionar este problema hemos generado un diccionario que relaciona ambos códigos.

Gracias a este diccionario hemos podido generar un sólo conjunto de datos, sin datos nulos, que nos va a permitir realizar el análisis de datos.

4. Análisis de los datos.

4.1 Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).

4.2 Comprobación de la normalidad y homogeneidad de la varianza.

4.3 Aplicación de pruebas estadísticas para comparar los grupos de datos.

5. Representación de los resultados a partir de tablas y gráficas.

6. Resolución del problema.

7. Código

8. Contribuciones.

9. Bibliografía.

Calvo M., Subirats L., Pérez D. (2019). *Introducción a la limpieza y análisis de los datos.* Editorial UOC

Dalgaard P. (2002). *Introductory Statistics with R*. New York: Springer-Verlag.

Squire M. (2015). *Clean Data.* Birmingham: Packt Publishing.

Vries A., Meys J. (2015). *R For Dummies, 2nd Edition.* New Jersey: John Wiley & Sons.

VV.AA. (2000). *Introducción a R. Versión 1.0.1.* R Development Core Team.