1. Título del dataset. Poned un título que sea descriptivo.

El título del dataset principal es Percepción_audiencia_cine_española. (en el código, sin tilde y usando n en lugar de ñ para garantizar que no hay problemas de codificación).

El del detaset de países será Paises_Latitud_Longitud.

2. Subtítulo del dataset. Agregad una descripción ágil de vuestro conjunto de datos por vuestro subtítulo.

Percepción_audiencia_cine_española:

> Relación entre recaudación de películas muy taquilleras en España y su valoración por la comunidad internacional.

Paises_Latitud_Longitud:

> Coordenadas GPS de un punto central del país autor de las películas que aparecen en Percepción_audiencia_cine_española.

3. Imagen. Agregad una imagen que identifique vuestro dataset visualmente

4. Contexto. ¿Cuál es la materia del conjunto de datos?

El conjunto de datos principal contiene información relacionada con la recaudación conseguida por las películas más taquilleras en España en los últimos años, así como una relación con el PIB per cápita de los mismos años. Con estos datos el objetivo es estudiar la recepción que tienen las películas entre la audiencia española.

5. Contenido. ¿Qué campos incluye? ¿Cuál es el periodo de tiempo de los datos y cómo se ha recogido?

Percepción_audiencia_cine_española:

> Los primeros 6 campos se extraen vía scrapping con la librería rvest directamente de la página del Ministerio de Cultura. Este sitio contiene las 25 películas más taquilleras desde 2002.
>
> El campo relPib lo obtenemos, también con rvest, de la página datosmacro.
>
> El campo puntuación lo obtenemos con la librería RJSONIO gracias a la API del sitio web TheMovieDataBase.

· [1] **Posición** (Integer). Es la posición en que quedó el film el año en que fue más taquillero. Como se recogen 25 títulos cada año, su valor va entre 1 y 25, siendo 1 el film más taquillero y 25 el menos.

· [2] **Largometraje** (Alfanumérico). Título de la película en España.

· [3] **Nacionalidad** (Factor). Nacionalidad de origen de la película.

· [4] **Distribuidora** (Alfanumérico). Compañía/Entidad que distribuye el film.

· [5] **Recaudación** (Double). En Euros. Dinero ingresado en concepto de entradas de cine.

· [6] **Año** (Integer). Año en que la película resultó taquillera.

· [7] **RelPib** (Double). Ratio con el PIB per cápita. Es el resultado de Recaudación/PIB(año) multiplicado por un factor de escala irrelevante para resultar en cantidades cómodas de manejar.

· [8] **Puntuación** (Double). Valoración que los usuarios de TMDB atribuyen a cada película.

· [9] **TasteInd** (Double). Es una medida de la acogida que tuvo la película en España. El cálculo se realiza a través de la regresión lineal de Recaudación vs Puntuación y año, de forma que para cada película calculamos qué recaudación se espera de ella según puntuación y año. La diferencia entre RepPib y el resultado es TasteInd. Las películas con valores positivos y altos de TasteInd son aquellas que tuvieron una muy buena acogida en relación con la valoración que tienen en TMDB, y la inversa, aquellas que tenga un TasteInd bajo o negativo son las más infravaloradas.

Hay varios campos que no son utilizados para esta práctica, como Posición, Nacionalidad y Distribuidora. He decido dejarlos porque aportan contexto y pueden ser útiles para que terceros realicen estudios sobre este dataset. Por ejemplo, alguien podría estar interesado en saber si el público español valora mejor el cine español, americano o inglés.

Paises_Latitud_Longitud:

· [1] **Nombre** (Factor). Nombre del país.

· [2] **Latitud** (Double). Este campo y el siguiente son las coordenadas de un punto más o menos céntrico del país correspondiente. Se obtiene de geocode. Como este servicio espera nombres en inglés, ha sido necesario construir una lookup de nombres de países, que relacionamos mediante el código ISO 3166-1 y las páginas de Wikipedia en inglés y español.

· [3] **Longitud** (Double).

WordClouds de Twitter:

Las imágenes generadas se obtienen gracias a la API de Twitter, la librería twitteR y wordcloud. Solicitamos tweets que contengan las palabras de los títulos del top-5 de TasteIndicators, y plasmamos su contenido en forma de nube de palabras.

6. Agradecimientos. ¿Quién es propietario del conjunto de datos? Incluid citas de investigación o análisis anteriores.

Para esta práctica acudimos a diversas fuentes. Alguna de ellas es propietaria del conjunto de datos y otras no.

Ministerio de Educación, Cultura y deporte. Propietario de los datos. Concretamente el Insitituto de la Cinematografía y de las Artes Visuales.
Datosmacro. No es propietaria. Estas son sus fuentes.

Los datos de los que hacemos uso para esta práctica son propiedad de este portal, ya que se trata de las puntuaciones que los usuarios le ceden.
powered by Google Geocode. Esta API para la geolocalización y sus datos pertecen a Google.
La ISO es la fuente de los códigos de países que he utilizado para traducir los nombres de los países, si bien es Wikipedia el portal que los presenta.
Twitter. El contenido de los tweets extraídos pertenece, en principio, a los usuarios que los escribieron. Como no damos uso comercial al mismo y además tampoco lo relacionamos con el nombre de usuario ni ningún otro dato, entiendo que no estamos infringiendo ninguna regla.

7. Inspiración. ¿Por qué es interesante este conjunto de datos? ¿Qué preguntas le gustaría responder la comunidad?

En ocasiones me he preguntado por qué tengo la percepción de que en este país triunfa contenido audiovisual (tanto películas como series o programas diarios) de calidad, a mi juicio, dudosa. En este sentido, he querido estudiar, a través de los datos, si mi percepción es correcta o estoy equivocado. Podría haber realizado este trabajo en otras áreas como las series, los hits musicales, libros etc., pero en una primera fase del trabajo vi claro el camino por la rama de las películas y decidí tirar por ahí.

Considero que el conjunto principal Percepción_audiencia_cine_española es interesante porque nos permite ver, con un solo parámetro, cuáles son los gustos de la audiencia de España, o si más no, cuáles son las películas que tienen un mejor recibimiento.

La pregunta a que responde este dataset, formulada en una frase, sería "¿Pagamos los españoles acordemente con la calidad que la comunidad internacional atribuye a las películas?" La respuesta es dudosa, porque pese a que

```
> sum(peliculas$TasteInd)
[1] 0.01
```

Es decir que la sobrevaloración compensa la infravaloración de una forma casi exacta, ocurre lo siguiente:

```
> length(which(peliculas$TasteInd<0))
[1] 235
> length(which(peliculas$TasteInd>0))
[1] 138
```

Es decir que hay muchas más películas que tuvieron peor acogida de la merecida que las que fueron injustificadamente aplaudidas.

Por tanto, la respuesta no es clara, aunque sospecho que el primer resultado es consecuencia del ajuste lineal realizado y por tanto la segunda es la que para mí tiene más peso. En cualquier caso, realizar análisis sobre el dataset queda fuera del alcance de este trabajo y por tanto no voy a entrar en ese terreno.

8. Licencia. Seleccionad una de estas licencias y decid porqué la habéis seleccionado:

- **Released Under CC0: Public Domain License**
- **Released Under CC BY-NC-SA 4.0 License**
- **Released Under CC BY-SA 4.0 License**
- **Database released under Open Database License, individual contents under Database Contents License**
- **Other (specified above)**

- **Unknown License**

Released Under CC BY-NC-SA 4.0 License. La escojo para que otros puedan compartir y adaptar, pero sin uso comercial, puesto que se trata de un trabajo académico que no fue concebido para tal fin.

9. Código: Hay que adjuntar el código con el que habéis generado el dataset, preferiblemente con R o Python, que os ha ayudado a generar el dataset.

Enlace a github: https://github.com/ellibrorojo/stable

10. Dataset: Dataset en formato CSV.

Enlace a github: https://github.com/ellibrorojo/stable/blob/master/Percepcion_audiencia_cine_espanola.csv