Práctica 2: ¿Cómo realizar la limpieza y análisis de datos?

# Descripción del dataset.

El dataset con el que vamos a trabajar es *puntuacionesFA.csv.* Este contiene los datos del usuario de Filmaffinity 968973. Se compone de 1941 filas, las cuales corresponden a las películas valoradas por el usuario y 14 columnas que reflejan:

1. titulo: Es el título de la película dado para el mercado español.
2. año: El año de estreno de la película.
3. duración: La duración en minutos del film.
4. pais: País donde se ha producido el metraje.
5. directores: Lista de director/es que realizaron la película.
6. guionistas: Lista de guionista/s que guionizaron la película.
7. musicos: Lista de musico/s implicado/s en la banda sonora del metraje.
8. fotografos: Lista de fotógrafo/s que participaron en la película.
9. actores: Lista de intérprete/s que actúan en la película.
10. productores: Lista de compañía/s que financian el metraje.
11. generos: Lista de genero/s entre los que se clasifica el metraje.
12. nota: La nota media de todos los usuarios de Filmaffinity que votaron el film.
13. votos: Número de votos que ha recibido la película por parte de los usuarios.
14. votacion: Votación del usuario del perfil del dataset.

Este dataset es el obtenido en la práctica de web scraping anterior. El análisis de este juego de datos pretende incorporar nuevas herramientas de análisis que la plataforma de Filmaffinity no suple, para realizar un estudio más exhaustivo de los gustos o preferencias del usuario.

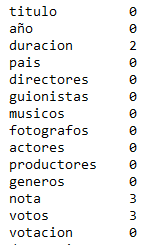
# Integración y selección.

Para este dataset no hemos considerado necesario añadir datos adicionales pero si crear varios subsets para el análisis posterior. Este conjunto de datos contiene varias columnas que acumulan datos en formato de lista, las cuales hemos transformado mediante una función propia que se ejecuta con un archivo 2\_extract\_subsets.py. Este ejecutable permite añadir el nombre de cualquier columna del dataset inicial para crear el archivo csv que se requiera.

[INSERTAR CAPTURA DEL EJECUTABLE]

# Limpieza de los datos.

Hemos detectado presencia de valores nulos en las columnas de *duración, nota* y *votos*.



Los nulos detectados en las columnas de nota y votos no las podemos sustituir por valores artificiales por lo que hemos decidido eliminar estas filas, debido a que no representan un alto porcentaje del total de registros.

En cuanto a los valores nulos en *duración* los hemos sustituido por una media del resto de valores de la columna.

Al realizar el análisis de valores extremos hemos detectado la presencia de estos en las columnas *año, duración, nota, votos* y *votación.* Como explicamos a continuación debido a la naturaleza de los datos estos valores extremos son normales.

-año: Los datos extremos se deben a que existen películas muy antiguas y el usuario ha visualizado aquellas más recientes.

-duracion: Se encuentran estos valores dispares ya que filmaffinity contiene los datos tanto de contometrajes como de documentales de larga duración.

-nota: Las valoraciones más bajas son las que causan estos valores atípicos, pero no se salen del rango 0-10.

-votos: Solo refleja la cantidad de votaciones de los usuarios de la plataforma.

-votacion: Al igual que la *nota* no se sale del rango 0-10.

# Análisis de los datos.

El análisis de los datos se va a realizar sobre los siguientes subsets:

-directores\_subset.csv: Contiene en desglose de los valores de la columna directores con sus valores correspondientes de titulo, votos, nota y votación.

-generos\_subset.csv: Contiene en desglose de los valores de la columna generos con sus valores correspondientes de titulo, votos, nota y votación.

-guionistas\_subset.csv: Contiene en desglose de los valores de la columna guionistas con sus valores correspondientes de titulo, votos, nota y votación.

-puntuacionesFA.csv: Dataset original.

# Contribuciones

|  |  |
| --- | --- |
| **Contribuciones** | **Firma** |
| Investigación previa | MR,CR |
| Redacción de las respuestas | MR,CR |
| Desarrollo del código | MR,CR |
| Participación en el vídeo | MR,CR |