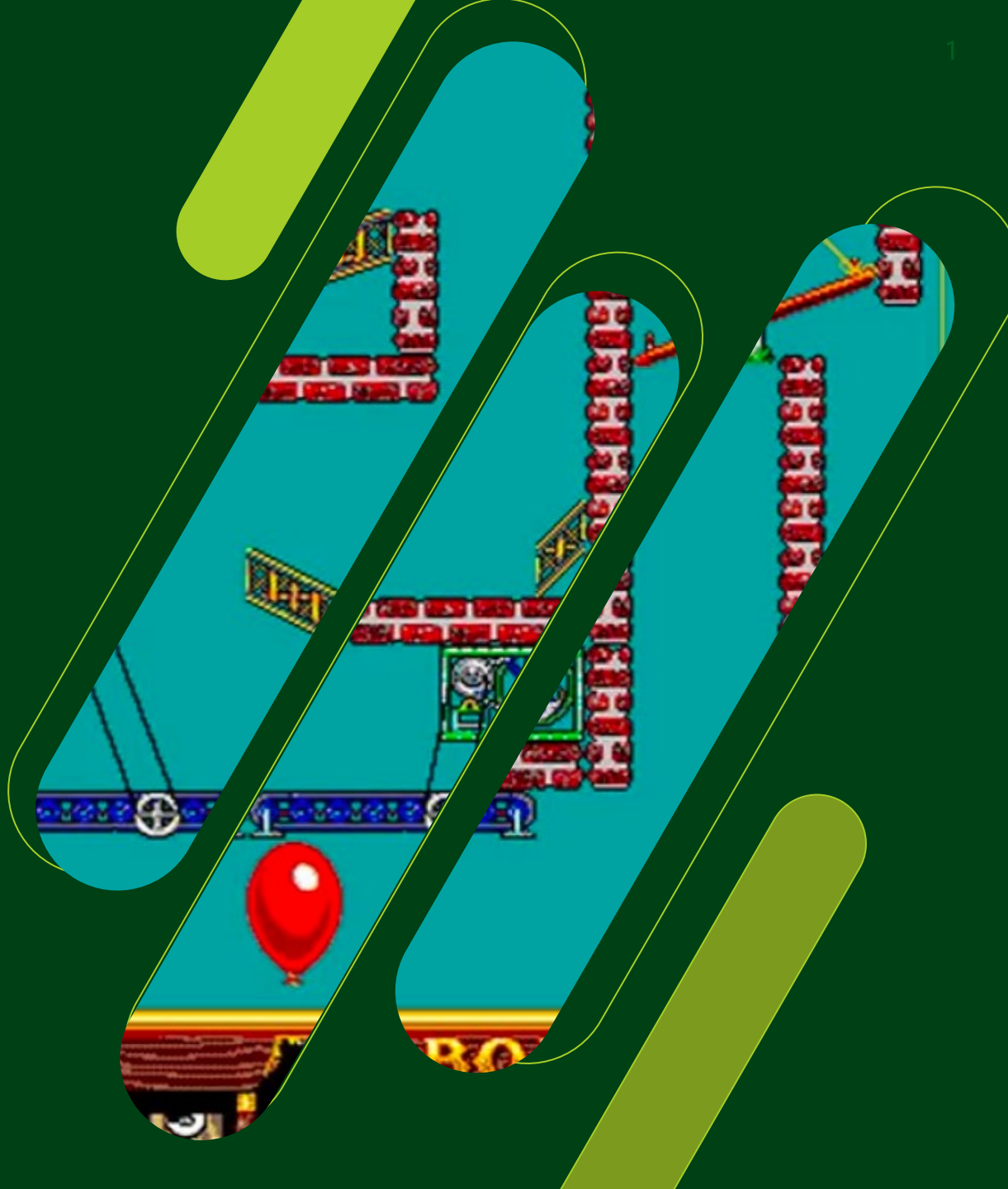
MBIA
Data Prep I

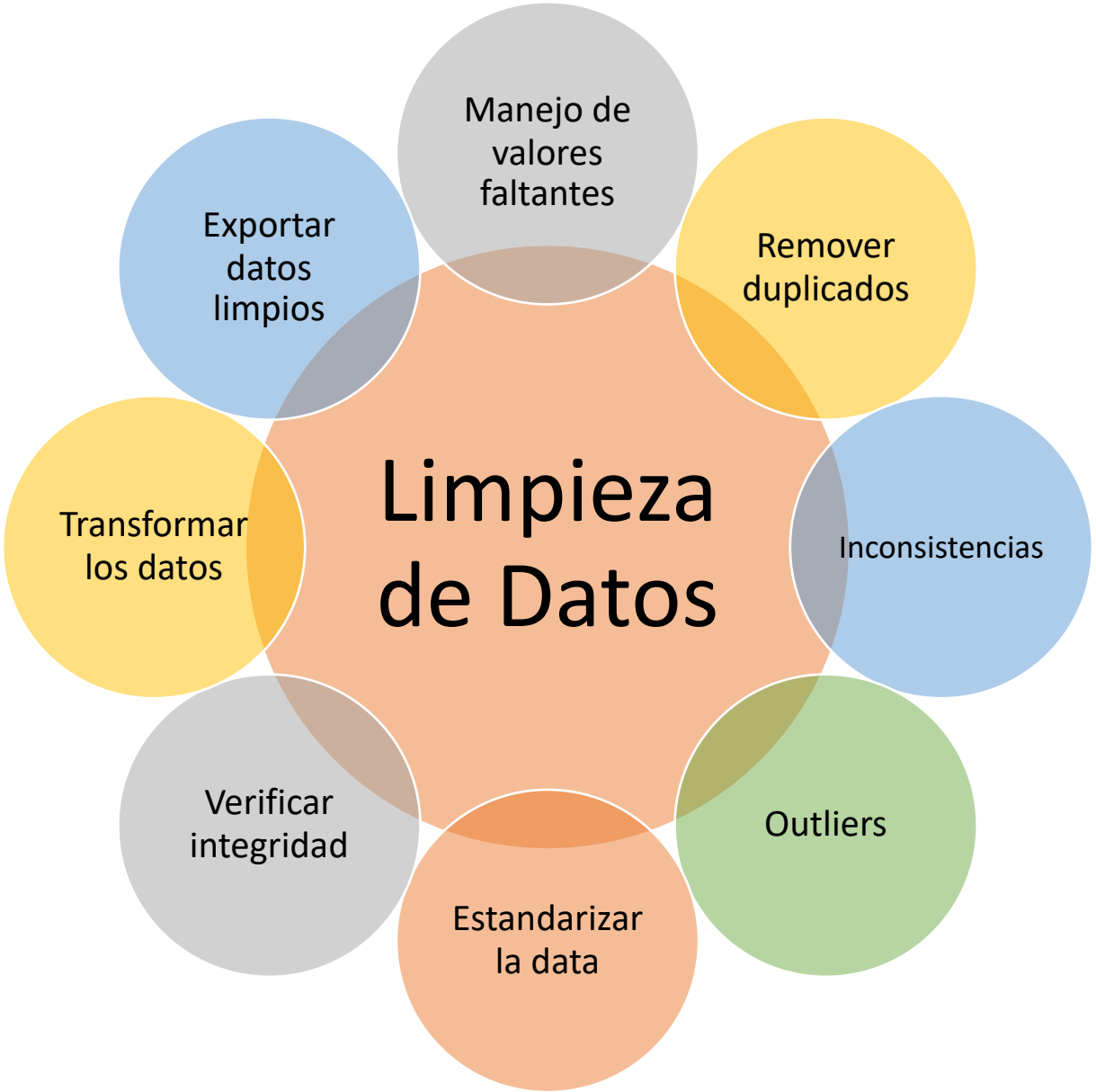
Tarea Data Cleansing

Abril 2023



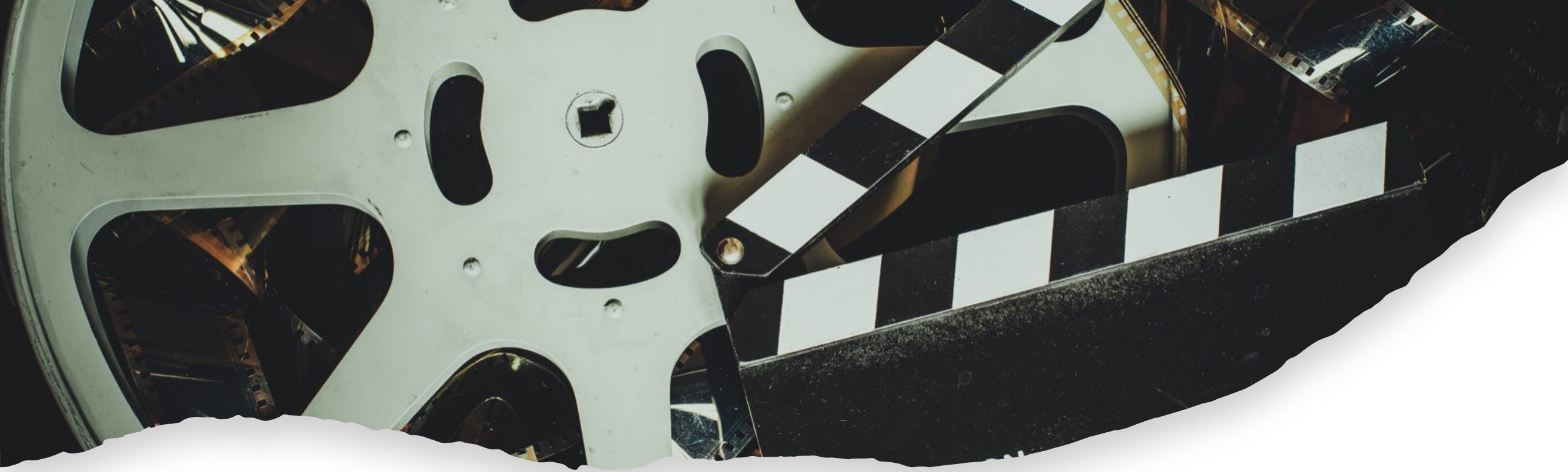
UVGMASTERS

Ejercicio Data Cleansing



Deberá trabajar con el conjunto de datos entregado, aplicar técnicas de Data Preparation y responder las preguntas planteadas

Se tomará en cuenta el procedimiento realizado para la preparación de los datos y la asertividad en las respuestas



Set de Datos

Deberá trabajar con un conjunto de datos sobre 100,000 calificaciones (1-5) que 900+ personas otorgaron a 1,600+ películas. Cada persona ha calificado al menos 20 películas.

También se cuenta con información demográfica de las personas (como edad, género y ocupación)

Los archivos relacionados pueden descargarse en Canvas.

Set de Datos

Archivos para trabajar el ejercicio:

1. movies_userdata.zip

Este archivo contiene varios archivos **CSV** relacionados con las calificaciones de películas. **Deben utilizarse los 8 archivos** contenidos dentro del .zip, los cuales comparten la misma estructura:

Campo	Descripción
userid	Código de persona que calificó una película
itemid	Código de película
raiting	Calificación otorgada a la película. Escala de 1 a 5. Siendo 5 la mejor calificación
timestamp	Fecha y hora del momento en que se realizó la calificación. Tomar en cuenta que está expresada en cantidad de segundos que han transcurrido desde el 1/ene/1970 a las 00:00 horas UTC

2. user.txt

Este archivo contiene la información acerca de las personas que han calificado las películas. Internamente está delimitado por **tabuladores** (tab) y no contiene ninguna fila inicial con encabezado. Su estructura es la siguiente:

Campo	Descripción
userid	Código de persona que calificó una película
age	edad
gender	género de la persona
occupation	ocupación
Zip code	Código postal (USA)

Set de Datos

Archivos para trabajar el ejercicio:

3. Item.txt

Este archivo contiene la información acerca de las películas que han sido calificadas. Internamente está delimitado por **tabuladores** (tab) y no contiene ninguna fila inicial con encabezado. Su estructura es la siguiente:

No. de campo	Campo	Descripción
1	movie id	Código de película
2	movie title	Nombre de película
3	release date	Fecha de película
4	Not used	Columna sin información debido a datos confidenciales omitidos
5	IMDb URL	Link a IMDB de la película
Del 6	unknown	Este campo identifica con 1 (sí) y 0 (no) si la película participa en este género
Al 24	western	Este campo identifica con 1 (sí) y 0 (no) si la película participa en este género

No. de campo	Género de Película
6	unknown
7	Action
8	Adventure
9	Animation
10	Children's
11	Comedy
12	Crime
13	Documentary
14	Drama
15	Fantasy
16	Film-Noir
17	Horror
18	Musical
19	Mystery
20	Romance
21	Sci-Fi
22	Thriller
23	War
24	western

Set de Datos – Consideraciones

Tomar en cuenta los siguientes aspectos al momento de utilizar los datos disponibles

1. Por error desde la fuente, pueden existir datos **duplicados** en los archivos de **calificaciones** de películas. Esto debe ser limpiado para evitar dar resultados equivocados. Los registros duplicados tienen exactamente la misma información en todos sus campos

2. Por un error en la plataforma que capturó las **calificaciones**, algunas de ellas no se encuentran dentro del rango 1 a 5. Se ha definido que estas calificaciones deben ser ignoradas por completo en los análisis a realizar

3. Tip: después de eliminar los duplicados y las calificaciones erróneas, en total deberá tener **100,000 filas** entre todos los archivos de calificaciones

4. La fecha en los archivos de calificaciones viene expresada en número de segundos transcurridos a partir de la fecha 1/ene/1970 tiempo UTC, hora 0:00. Deberá cualcular la fecha y hora de cada evento tomando en cuenta este aspecto. Puede aplicarse el concepto relacionado con Unix time: https://en.wikipedia.org/wiki/Unix_time

5. Se otorgarán algunos puntos extras si la última pregunta a contestar se calcula convirtiendo las horas al tramo horario conforme al zip code de las personas que han calificado

Preguntas

Responder a cada una de las siguientes cuestionantes con base a los datos preparados

1. ¿Cuáles son las 3 películas más populares? Es decir, las que más calificaciones han recibido, sin importar la calificación otorgada
2. ¿Cuáles son las 10 películas mejor calificadas por los usuarios?
3. ¿Cuántas películas hay de cada género?. Si una película pertenece a más de un género, puede contarse más de una vez
4. ¿Qué género de películas tiene las calificaciones más altas?
5. ¿Cuál es la edad media de todos los usuarios que calificaron las películas?
6. ¿Existe alguna relación entre la edad del usuario y la calificación que otorgó a la película?
7. ¿Cuáles son las películas más populares entre los usuarios mayores de 50 años?
8. ¿Hay algún género de película que sea más popular entre los usuarios mujeres que entre los usuarios hombres?
9. ¿Cuál es la película mejor calificada durante fines de semana, en horario de 8 a 10 P.M?

Entrega

Tomar en cuenta los siguientes aspectos para entregar como parte de esta tarea

1. Podrá utilizar cualquier herramienta(s) de preparación de datos y/o lenguaje de programación que desee

2. El formato de entrega deberá ser un archivo PDF que contenga los siguientes elementos:

    I. Pasos/Lógica/algoritmo utilizado para limpiar, prepar e integrar los archivos. Incluir explicación breve y print-screens de la herramienta utilizada

    II. Breve resumen de lo que se observa al realizar data-profiling sobre los datos compartidos

    III. Respuesta a las preguntas planteadas. El formato de estas respuestas debe ser pensando en que la audiencia que las lea será de nivel ejecutivo en una empresa (y que no habla lenguaje técnico). Cuidar la presentación y ortografía.

Set de Datos – Citation

El formato y contenido de los archivos originales pudieron ser modificados con fines didácticos

SUMMARY & USAGE LICENSE
==========

MovieLens data sets were collected by the GroupLens Research Project at the University of Minnesota.

The data was collected through the MovieLens web site (movielens.umn.edu) during the seven-month period from September 19th, 1997 through April 22nd, 1998. This data has been cleaned up – users who had less than 20 ratings or did not have complete demographic information were removed from this data set. Detailed descriptions of the data file can be found at the end of this file.

Neither the University of Minnesota nor any of the researchers involved can guarantee the correctness of the data, its suitability for any particular purpose, or the validity of results based on the use of the data set. The data set may be used for any research purposes under the following conditions:

* The user may not state or imply any endorsement from the University of Minnesota or the GroupLens Research Group.

* The user must acknowledge the use of the data set in publications resulting from the use of the data set (see below for citation information).

* The user may not redistribute the data without separate permission.

* The user may not use this information for any commercial or revenue-bearing purposes without first obtaining permission from a faculty member of the GroupLens Research Project at the University of Minnesota.

If you have any further questions or comments, please contact GroupLens <grouplens-info@cs.umn.edu>.

Further information on the GroupLens Research project, including research publications, can be found at the following web site: http://www.grouplens.org/

GroupLens Research currently operates a movie recommender based on collaborative filtering: http://www.movielens.org/

CITATION
==========

To acknowledge use of the dataset in publications, please cite the following paper:

F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets:
History and Context. ACM Transactions on Interactive Intelligent
Systems (TiiS) 5, 4, Article 19 (December 2015), 19 pages.
DOI=http://dx.doi.org/10.1145/2827872