

Individual Delivery

28/12/2020 – 13/01/2021

**─**

Juan Ignacio Guerrero Enríquez

Data Science

The Bridge

# General Vision

A partir del dataset de Full MovieLens (<https://www.kaggle.com/rounakbanik/the-movies-dataset?select=movies_metadata.csv>) que recoge metadatos sobre las 45.000 películas de esta plataforma, se realizará un estudio para comprobar la veracidad de la hipótesis que guiará este trabajo:

*-La industria del cine no es rentable*

El dataset incluye 24 columnas, de las cuales solo nos quedaremos con 10 relativas a: nombre de la película, fecha de estreno, idioma original, géneros, presupuesto, recaudación, países de producción, tiempo de exposición, media de puntuaciones y número de puntuaciones. Con los datos recogidos en estas columnas se pretende dar respuesta a la hipótesis planteada.

# Goals

El nivel que se pretende alcanzar con este trabajo, de los planteados en la guía, es el A.

# Specifications

Para alcanzar este objetivo se detallarán todos los pasos llevados a cabo y los requerimientos necesarios para ejecutar este proyecto:

## Software

-Python 3.9 o superior

-Jupyter Notebook

-IDE tal como Visual Studio Code

## Hardware

-8Gb de RAM o más

# Steps

## Research the context

El primer paso del trabajo consistió en buscar datos con los que poder trabajar y llevar a cabo un EDA (Exploratory Data Analysis). Por tanto, fue necesario acudir a un repositorio de bases de datos para hallar un dataset con el que trabajar. En este caso, a la plataforma a la que se acudió fue Kaggle.

## Get Data

De la dirección (<https://www.kaggle.com/rounakbanik/the-movies-dataset?select=movies_metadata.csv>) se descargó un CSV con información de metadatos de las 45 000 películas registradas por el portal Full MovieLens hasta el 2017.

## Data Wrangling

A partir de los datos se generó un dataframe desde el que poder acceder a los datos. De las columnas del dataframe original solo nos quedamos con diez.

## Data Mining / Clean Data

El minado y limpieza de los datos constituyó una parte muy importante del trabajo. El dataset requirió de un tratamiento intenso para hacer la información de las columnas accesibles y eliminar información faltante, extraña o anómala. Entre las funciones a llevar a cabo en estos apartados se encontraron: transformación de los objetos de columnas a clases con las que poder trabajar, extracción de infromación contenida en jsons dentro de columnas, interpretación de errores y corrección en el proceso de minado, detección de datos ausentes, et.

## Data Exploration

Visualización de las distribuciones contenidas en cada columna a través de gráficas. Este proceso ayudó en el proceso de limpieza de datos al facilitar la visualización de la información contenida en las columnas y a la comprensión de las mismas

1. Hypothesis Evaluation

Una vez que se tienen los datos listos para su tratamiento e interpretación, se procedió a generar visualizaciones que nos permitirán llegar a una conclusión sobre la hipótesis planteada.