# Open data

Aunque como hemos comentado en algunas ocasiones las fuentes de datos, aunque haya mucha, y cada vez más, promoción de los lugares con datos en abierto, sigue siendo un gran hándicap para los trabajos de análisis y visualización de los datos.

Además de otros enlaces que os he indicado, este es un buen portal de datos con mucha disponibilidad:

<https://www.kaggle.com/>

En él se hacen competiciones sobre predicciones **(analíticas predictivas)**. Este un tema de los más solicitados con los datos y os muestro algunos ejemplos de lo que se puede hacer con ellos como por ejemplo el estupendo trabajo realizado con la base de datos de películas, obtenido de la famosa web imdb.

<https://blog.nycdatascience.com/student-works/machine-learning/movie-rating-prediction/>

Realiza un índice sobre el trabajo muy interesante.

# Indice

## Primera pregunta / o pregunta inicial

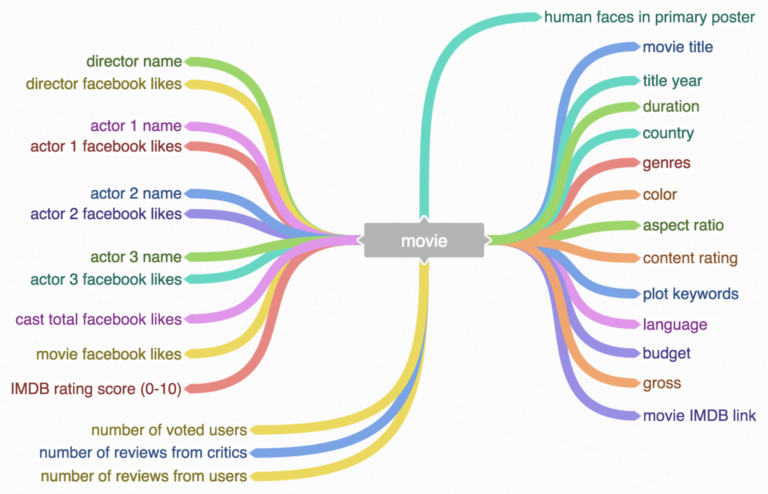
¿Cómo podemos decir la grandeza de una película antes de su estreno en el cine?

Depende del entorno, estamos hablando de la motivación, del porqué nos interesa realizar este trabajo. Luego irá enlazado con el objetivo.

## Scraping 5000+ películas de IMDB

Web scraping es la técnica utilizada mediante cualquier programa software para extraer información de sitios web.

Las herramientas usada en esta ocasión para extraer información de todas las películas 5000+ es una biblioteca de Python llamada "scrapy". A continuación se presentan algunos pasos breves (detallados en la web).

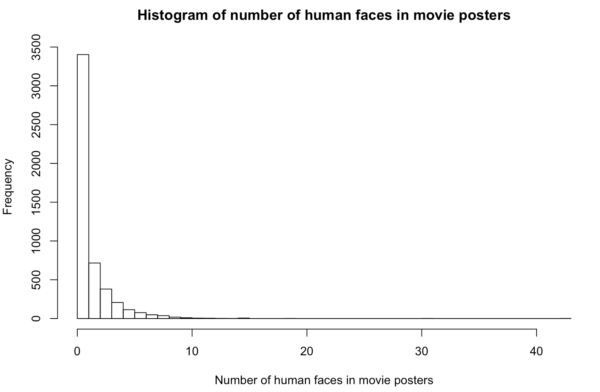
La imagen mostrada sirve para mostrar rápidamente las variables extraídas desde la web.

## Detección de rostros de carteles de películas

En el trabajo, los autores están especialmente interesados en saber la respuesta a esta pregunta: ¿El número de caras humanas en el póster de la película se correlacionará con la calificación de película?

Fijaros ahí la importancia de las **correlaciones**, elemento fundamental cuando se trabaja con gran volumen de datos, cuando a priori no se sabe lo que se puede obtener de los datos.

Histograma



## Calificación IMDB (rating score)

Rating score : Puntuación / calificación / valoración…

De las 28 variables, el estudio hace especial hincapié en saber cómo se correlaciona la calificación de IMDB con otras variables.

Varios subapartados con cada una de sus comparación con la calificación IMDB. Todos ellos con gráficos para la mejor interpretación

### IMDB score VS country

### IMDB score VS movie year

### IMDB score VS movie facebook popularity

### IMDB score VS director facebook popularity

### IMDB score VS top 3 actors/actresses facebook popularity

## Predicción de clasificación de películas

La predicción de las clasificaciones de películas en este trabajo se basa en los siguientes supuestos:

* La puntuación de IMDB refleja la grandeza de las películas. Cuanto más alto, mejor película.
* Es preferible ver películas buenas a malas para muchas personas.

Con esas 28 variables disponibles para todas las películas raspadas, ¿podemos predecir la clasificación de películas? Antes de comenzar, es necesario investigar la correlación de esas variables.

### Correlation analysis

### Dimensionality reduction

### Multiple linear regression

### Multiple linear regression

### Insights