GROUP 2

NGUYỄN HỒNG ĐỨC

PHẠM ĐỨC TOÁN

ĐINH MAI

## Miêu tả dữ liệu:

**Đường dẫn**: \data\movie\_dataset.csv

**Nội dung**: Thông tin về phim từ năm 1917 – 2017

**Số bản ghi:** 4803

**Trường dữ liệu:**

* budget: ci phí sản xuất
* genres: thể loại
* homepage: trang chủ
* *id: id (khóa chính)*
* keywords: từ khóa miêu tả
* original\_language: ngôn ngữ
* original\_title: tên gốc
* overview: tóm tắt nội dung
* popularity: chỉ số sự phổ biến
* production\_companies: công ty sản xuất phim
* production\_countries: quốc gia sản xuất phim
* release\_date: ngày phát hành
* revenue: lợi nhuận
* runtime: thời lượng
* spoken\_languages: ngôn ngữ sử dụng trong phim
* status: đã phát hành hay chưa
* tagline: tựa đề quảng cáo
* title: tên
* vote\_average: đánh giá trung bình
* vote\_count: số lượng đánh giá
* cast: dàn diễn viên
* crew: danh sách nhân viên sản xuất phim
* director: diễn viên

## Làm sạch/chuẩn bị dữ liệu:

**Miêu tả vấn đề và cách giải quyết**

**Đường dẫn:** \data\_cleaning.ipynb

1. **Ký tự unicode:**

Vấn đề:

Một số ký tự bị mã hóa unicode.

Cách giải quyết:

Tìm những trường có ký tự unicode (‘\u’) rồi dùng .decode("unicode-escape") để giải mã.

1. **Trường budget/revenue:**

Vấn đề:

Một số lượng lớn (1500 bản ghi) với trường quan trọng budget hoặc revenue bằng 0, hoặc quá nhỏ (nhỏ hơn 10) do tập dữ liệu thiếu, sai hoặc không có thông tin...

Cách giải quyết:

Đường dẫn: \budget\_revenue\_crawl

Viết chương trình crawl dữ liệu budget và revenue từ trang the-numbers.com và google.com. Cuối cùng còn khoảng 400 bản ghi lỗi -> chấp nhận bỏ ra khỏi dataset

1. **Tạo bảng dim\_genres:**

Vấn đề:

Tên thể loại cách nhau bởi dấu cách.

Cách giải quyết:

Tạo một hàm tách các tên thể loại, chú ý tên “Science Fiction”. Kết quả là một bảng **dim\_genres**.

1. **Tạo bảng dim\_prod\_comp:**

Vấn đề:

Mỗi ô trường production\_companies có định dạng giống file json

Cách giải quyết:

Dùng hàm read\_json để chuyển thành dataframe rồi append lại .Kết quả là một bảng **dim\_prod\_comp**.

1. **Tạo bảng dim\_prod\_comp:**

Vấn đề:

Mỗi ô trường production\_countries có định dạng giống file json.

Cách giải quyết:

Dùng hàm read\_json để chuyển thành dataframe rồi append lại. Kết quả là một bảng **dim\_prod\_ctry**.

1. **Tạo bảng day, month, year:**

Cách giải quyết:

Tách trường release\_date

1. **Tạo bảng dim\_lang:**

Vấn đề:

Mỗi ô trường spoken\_languages có định dạng giống file json.

Cách giải quyết:

Dùng hàm read\_json để chuyển thành dataframe rồi append lại, cho thêm trường original\_language. Kết quả là một bảng **dim\_lang.**

1. **Tạo bảng dim\_crew:**

Vấn đề:

Mỗi ô trường spoken\_languages có định dạng giống file json. Tuy nhiên read\_json chỉ đọc được value dùng ngoặc kép (“”). Dữ liệu trường này lại viết ở ngoặc đơn (‘’). Ngoài ra một số bản ghi có thêm (“”) (ví dụ, như tên ‘Thomas “Noah” Yorke’)

Cách giải quyết:

Tạo hàm json\_clean để làm sạch lỗi. Dùng hàm read\_json để chuyển thành dataframe rồi append. Kết quả là một bảng **dim\_crew.**

1. **Trường cast:**

Vấn đề:

Tên diễn viên ngăn cách bằng dấu cách. Là tên riêng có thể có 1,2,3 4 kí tự nên không thể tách ra như genres được

Cách giải quyết:

Đường dẫn: \actor\_crawl

Tìm được dataset của imdb với 400 000 bản ghi là tên diễn viên. Viết chương trình tách tên dần dần từ bản cast -> Kết quả là một bảng **dim\_cast.**

1. **Tạo trường inflated\_budget và inflated\_revenue**

Cách giải quyết:

Đường dẫn: \data\inflation\_data.csv

Tìm inflation\_data trên mạng rồi tạo hàm tính lạm phát.

1. **Tạo trường profitability:**

Cách giải quyết:

Tạo trường profitablity = revenue/budget \* 100%

1. **Tạo trường weighted\_rating:**

Cách giải quyết:

Công thức: WR = (v / (v+m)) \* R + (m / (v+m)) \* C

WR: weighted\_rating

R: vote\_average

v: vote\_count

m: lượng vote tối thiểu để có trong top bảng xếp hạng (chọn là 54)

C: trung bình đánh giá của cả dataset

1. **Xóa cách trường không cần thiết**

['genres', 'original\_language', 'production\_companies', 'production\_countries', 'spoken\_languages', 'crew', 'status']

## Insight và dashboard:

Sau phần clean dữ liệu, chuẩn bị phân tích và trực quan hóa:

Diagram

Description automatically generated

Một số ý kiến về insight:

1. Tương quan giữa budget và revenur
2. Top 10 film có doanh thu/chi phí cao nhất năm
3. Tương quan giữa rating và profitability
4. Diễn viên/Đạo diễn có nhiều film có lãi nhất
5. Độ dài film qua từng năm
6. Thể loại film có doanh thu cao nhất