ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN KHOA CÔNG NGHỆ THÔNG TIN



DATA VISUALIZATION USING TABLEAU

Giảng viên hướng dẫn:

Nguyễn Ngọc Minh Châu

THÔNG TIN NHÓM

|  |  |  |  |
| --- | --- | --- | --- |
| ***Mã nhóm*** | ***MSSV*** | ***Họ và tên*** | ***Ghi chú*** |
| 22HTTT  Nhóm 11 | 22127107 | Nguyễn Thế Hiển | Trưởng nhóm |
| 22127260 | Bùi Công Mậu |  |
| 22127355 | Nguyễn Trần Đại Quốc |  |
| 22127400 | Thái Hữu Thọ |  |

**LƯU Ý: CÁC VÍ DỤ CHỈ DÙNG ĐỂ MÔ TẢ CHO PHẦN TƯƠNG ỨNG, KHÔNG DÙNG ĐỂ CHỈ DẪN SV LÀM BÀI.**

# Dataprofiling

### Tiền xử lí dữ liệu

Dữ liệu được cung cấp là từ kaggle: <https://www.kaggle.com/datasets/shivamb/netflix-shows>.



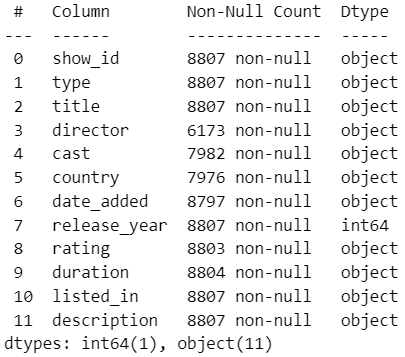
*File netflix\_titles.csv*

Dữ liệu gồm 8807 bộ dữ liệu ( không có header), có tổng cộng 11 thuộc tính được miêu tả như sau:

|  |  |
| --- | --- |
| Thuộc tính | Miêu tả |
| Show\_id | Mã xác định cho một bộ phim hoặc series. Mỗi bộ phim hoặc series có mã duy nhất. |
| Type | Loại của bộ phim hoặc series. Ví dụ, nó bao gồm các giá trị như "Phim" hoặc "Chương trình truyền hình". |
| Title | Tên phim. |
| Director | Đạo diễn của bộ phim hoặc series. |
| Cast | Danh sách các diễn viên đóng vai nhân vật chính trong bộ phim hoặc series. |

|  |  |
| --- | --- |
| Country | Danh sách các nước mà phim hay series được quay tại đó. |
| Date\_added | Ngày bộ phim hoặc series được chiếu trên Net-flix. |
| Release\_year | Năm phát hành của bộ phim hoặc series. |
| Rating | Xếp hạng của phim hoặc series dựa trên đánh giá của khán giả. Thường được đánh giá cùng với phân loại độ tuổi (ví dụ: PG-13, R). |
| Duration | Thời lượng của bộ phim tính theo thời gian phút. Tổng thời gian của 1 phần phim hoặc toàn bộ series. |
| Listed\_in | Thể loại của phim, series. |
| Description | Mô tả ngắn gọn về bộ phim hoặc loạt phim. Tóm tắt nhanh cốt truyện chính của bộ phim. |

Dữ liệu được cho là dữ liệu thô, có khá nhiều dữ liệu có missing value:



Dựa trên những thông tin trên, ta thực hiện quá trình tiền xử lí dữ liệu:

+ Đối với kiểu dữ liệu là dạng số (float), điền những ô trống bằng giá trị mean.

+ Đối với kiểu dữ liệu là dạng object, điền những ô trống bằng giá trị “Không xác định” hoặc “Không có thông tin” để nhận diện. Ở những trường hợp này, không thể điền giá trị cụ thể vào vì dễ làm sai lệch thông tin trong quá trình xử lí dữ liệu.

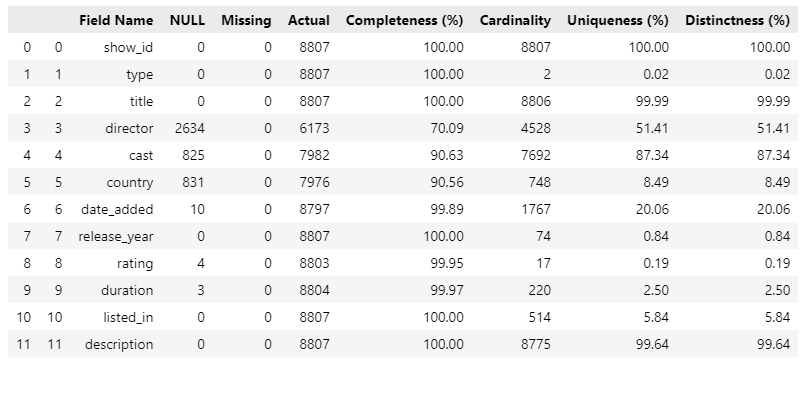
Sau quá trình tiền xử lí, ta thu được dữ liệu sạch vào file “netflix\_titles\_cleaned.csv”:



*File netflix\_titles\_cleaned.csv*

### Profiling toàn bộ dữ liệu

Tiến hành profiling toàn bộ dữ liệu thu được kết quả như sau:



Nhận xét:

* Trước khi preprocessing dữ liệu, các thuộc tính show\_id, type, title, release\_year, listed\_in, description có completeness 100%.78
* date\_added, rating, duration đều c 8ó completeness > 99.8% chỉ thiếu vài giá trị, dễ dàng tiền xử lý.
* duration có cardinality 220 độ đa dạng vừa đủ để thực hiện các phân tích thông tin.
* rating có cardinatily thấp thấy được rằng là 1 tập giá trị phân loại, dùng để phân loại dữ liệu theo 1 điều kiện nào đó.
* director chỉ có completeness là 70.09% gần 30% dữ liệu bị thiếu gây ảnh hưởng tới quá trình phân tích làm nhiễu dễ dàng bỏ sót thông tin.
* cast, country thiếu cũng khá nhiều với 9-10% dữ liệu.
  + - show\_id (100% duy nhất), title (99.99% duy nhất, 1 giá trị trùng), và description (99.64% duy nhất, 32 giá trị trùng) đảm bảo được tính xác định của từng bộ phim.
    - Sau khi preprocessing dữ liệu, thì đã xử lí được hết các giá trị NULL nên toàn bộ là 0 completeness tất cả các cột 100%.
    - Actual (dòng có giá trị) là 8,807, bằng với tổng số dòng.
    - country có 749 giá trị duy nhất (8.50%), cho thấy sự đa dạng về quốc gia sản xuất. listed\_in (thể loại) có 514 tổ hợp (5.84%), phản ánh các tựa phim thuộc nhiều thể loại khác nhau.
    - type (2 giá trị: Movie/TV Show), rating (18 giá trị), release\_year (74 giá trị), và duration (221 giá trị) có tính duy nhất thấp, phù hợp với vai trò phân loại, hỗ trợ phân tích xu hướng (ví dụ: theo năm, xếp hạng).

### Profiling từng thuộc tính

**Show\_id:**

A screenshot of a computer

AI-generated content may be incorrect.

#### Tổng quát:

Độ hoàn chỉnh (Completeness)

* Thuộc tính show\_id có độ hoàn chỉnh 100%, không có giá trị NULL.
* Dữ liệu đầy đủ, không cần xử lý thiếu dữ liệu.

Cardinality (Số lượng giá trị duy nhất)

* Có 8807 giá trị duy nhất,là primary key mỗi dữ liệu đều được xác định bởi 1 id.

Uniqueness & Distinctness

* Uniqueness: 100%
* Distinctness: 100%

#### Nhận xét:

#### Phân phối định dạng show\_id:

* show\_id có tổng cộng 4 định dạng chính dựa trên độ dài chuỗi:
* XXXXX – 7808 bản ghi (88.66%)
* XXXX – 900 bản ghi (10.22%)
* XXX – 90 bản ghi (1.02%)
* XX – 9 bản ghi (0.10%)
* Dù có sự chênh lệch về định dạng, nhưng không ảnh hưởng đến tính duy nhất của giá trị.

Thống kê số liệu:

* Độ dài chuỗi show\_id:
  + Ngắn nhất: 2 ký tự
  + Dài nhất: 5 ký tự
* Tổng số định dạng khác nhau: 4
* Kiểu dữ liệu: object (chuỗi)

Thuộc tính show\_id đóng vai trò như khóa chính trong tập dữ liệu.

### type:

A screenshot of a spreadsheet

AI-generated content may be incorrect.

#### Tổng quát:

Độ hoàn chỉnh (Completeness)

* Thuộc tính type có độ hoàn chỉnh 100%, không có giá trị NULL.
* Dữ liệu đầy đủ, không cần xử lý thiếu dữ liệu.

Cardinality (Số lượng giá trị duy nhất)

* Có 2 giá trị duy nhất, tương ứng với hai loại nội dung là Movie và TV Show.
* Điều này cho thấy đây là biến phân loại nhị phân, dễ dàng phân tích và trực quan hóa.

Uniqueness & Distinctness

* Uniqueness: 0.0227% → Mỗi giá trị (Movie hoặc TV Show) xuất hiện nhiều lần trong tập dữ liệu.
* Distinctness: 0.0227% → Dữ liệu có sự lặp lại rất lớn, do chỉ có 2 nhóm phân loại.

#### Nhận xét:

Phân phối loại nội dung

* Dữ liệu về type phân bố như sau:
  + Movie: 6131 bản ghi – chiếm 69.62%.
  + TV Show: 2676 bản ghi – chiếm 30.38%.
* Phim (Movie) là loại nội dung phổ biến hơn rõ rệt trong tập dữ liệu.
* Không có dấu hiệu outlier hay sai sót → Phân phối hợp lý.

Thống kê số liệu

* Số lượng định dạng: 2 (XXXXX và XXXXXX), phù hợp với độ dài của hai giá trị Movie và TV Show.
* Độ dài chuỗi:
  + Dài nhất: 7 ký tự (TV Show).
  + Ngắn nhất: 5 ký tự (Movie).

Dữ liệu dạng phân loại (Categorical)

* Toàn bộ dữ liệu type là kiểu object (chuỗi).
* Dữ liệu rõ ràng, dễ xử lý và phù hợp để phân tích phân loại, trực quan hóa (biểu đồ tròn, thanh).

type là một thuộc tính phân loại chất lượng cao, hoàn chỉnh và rõ ràng.

### title:

A screen shot of a graph

AI-generated content may be incorrect.

#### Tổng quát

Độ hoàn chỉnh (Completeness)

* Thuộc tính title có độ hoàn chỉnh 100%, không có giá trị NULL.
* Dữ liệu đầy đủ, không cần xử lý thiếu dữ liệu.

Cardinality (Số lượng giá trị duy nhất)

* Có 8806 giá trị duy nhất trên tổng số 8807 bản ghi, điều này cho thấy hầu hết tiêu đề là duy nhất.
* Chỉ có 1 giá trị bị trùng lặp (một bộ phim được remake), chiếm tỷ lệ rất nhỏ.

Uniqueness & Distinctness

* Uniqueness: 99.9886%
* Distinctness: 99.9886%

→ Gần như mỗi tiêu đề là duy nhất, đảm bảo độ đa dạng và phân biệt cao giữa các bản ghi.

#### Nhận xét

Dữ liệu dạng văn bản (Object)

* Toàn bộ dữ liệu trong title là kiểu object (chuỗi).
* Không có giá trị NULL → Đã được tiền xử lý.
* Không có giá trị trùng lặp thấy được các tiêu đề hoàn toàn không giống nhau trừ các tác phẩm được làm lại.

### director:

A screenshot of a computer

AI-generated content may be incorrect.

#### Tổng quát

Độ hoàn chỉnh (Completeness)

* Thuộc tính director có độ hoàn chỉnh 100%, không có giá trị NULL hay thiếu.
* Có hơn 29.9% bản ghi mang giá trị “Không có thông tin”, điều này tương đương với 2634 dữ liệu đã được tiền xử lý thay thế missing valuse.

Cardinality (Số lượng giá trị duy nhất)

* Có 4529 giá trị duy nhất trong tổng số 8807 bản ghi.
* Mức độ phân biệt tương đối cao, cho thấy dữ liệu phong phú về danh tính đạo diễn đồng thời cũng có nhiều đạo diễn thực hiện nhiều bộ phim khác nhau.

Uniqueness & Distinctness

* Uniqueness: 51.425%
* Distinctness: 51.425% → Dữ liệu có sự lặp lại đáng kể giữa các mức cholesterol.

→ Hơn một nửa số đạo diễn là duy nhất, số còn lại có thể là đạo diễn của nhiều phim hoặc TV shows khác nhau.

#### Nhận xét

Phân phối Đạo diễn

* Có hơn 4500 đạo diễn khác nhau, cho thấy dữ liệu rất đa dạng.
* Đạo diễn phổ biến nhất là:
* Rajiv Chilaka (19 lần)
* Raúl Campos, Jan Suter (18 lần)
* Suhas Kadav, Marcus Raboy (16 lần mỗi người)
* Rất nhiều đạo diễn chỉ xuất hiện đúng 1 lần (ví dụ: *Joe Menendez*, *Will Eisenberg*, *Mozez Singh*,...).
* Ngoài ra còn có 29,9% các bộ phim, series không có thông tin về đạo diễn.
* Điều này phản ánh **mức độ phân tán cao** của trường dữ liệu director.

Dữ liệu dạng văn bản (Object)

* Toàn bộ dữ liệu trong director đều là object
* Không có giá trị NULL → đã qua bước tiền xử lý.

Director rất giàu thông tin, phù hợp cho các phân tích dạng:

* Tìm top đạo diễn theo số lượng phim
* So sánh phong cách đạo diễn
* Phân tích tác động đạo diễn đến thể loại/phổ biến

### Cast

### A screenshot of a computer AI-generated content may be incorrect.

### Tổng quát:

Độ hoàn chỉnh (Completeness)

* Completeness cao (100%) không có giá trị NULL nào.
* Tuy nhiên, có 825 bản ghi (9.37%) mang giá trị "Không có thông tin" → chiếm tỷ lệ cao sẽ gây ra ảnh hưởng đến quá trình phân tích, trực quan hóa và so sánh.

Cardinality (Số lượng giá trị duy nhất)

* Có 7693 giá trị duy nhất trên tổng số 8807 bản ghi → cho thấy sự đa dạng và phong phú rất cao trong dữ liệu diễn viên, có thay đổi về số lượng diễn viên trong các bộ phim.

Uniqueness & Distinctness

* Uniqueness: 87.351%
* Distinctness: 87.351%

→ Phần lớn các bản ghi có dàn diễn viên khác nhau, chỉ một số ít chương trình/phim có cùng dàn diễn viên.

#### Nhận xét:

Dữ liệu dạng văn bản (Object)

* Kiểu dữ liệu: object
* Chiều dài trường:
  + Tối thiểu: 3 ký tự
  + Tối đa: 771 ký tự → có thể chứa danh sách dài các diễn viên (tên được phân cách bằng dấu phẩy hoặc cách khác).

Số định dạng khác nhau: 386 → cho thấy định dạng tên diễn viên không đồng nhất, có thể cần chuẩn hóa (đặc biệt nếu tách tên diễn viên riêng lẻ để phân tích mạng lưới hoặc tần suất).

→ Dễ nhận thấy rằng không có diễn viên nào xuất hiện quá nhiều lần, dữ liệu diễn viên phân tán mạnh → phù hợp với đặc điểm của phim và show truyền hình đa dạng.

### Country

A screenshot of a computer

AI-generated content may be incorrect.

#### Tổng quát:

Độ hoàn chỉnh (Completeness)

* Completeness cao (100%) không có giá trị NULL nào
* Tuy nhiên, có 831 bản ghi (9.4357%) mang giá trị "Không xác định".

Cardinality (Đa dạng giá trị)

* Cardinality: 749 → rất nhiều quốc gia hoặc tổ hợp quốc gia khác nhau (vd: “France, Spain”, “Mexico, United States…”)
* Uniqueness/Distinctness: 8.5046% → mức độ lặp lại không cao, dữ liệu rất phân tán, thể hiện độ phong phú về nguồn gốc sản xuất.

#### Nhận xét:

* Hoa Kỳ chiếm gần 32%, tiếp theo là Ấn Độ (~11%) và Vương Quốc Anh (~5%).
* Có nhiều bản ghi chứa nhiều quốc gia: ví dụ France, Spain, Germany → điều này cho thấy tính quốc tế và hợp tác đa quốc gia trong sản xuất nội dung.
* Dữ liệu không xác định tương đối lớn, gần 1/10 toàn bộ tập → điều này ảnh hưởng trực tiếp đến các phân tích có liên quan đến địa lý.

### Date\_added

A screenshot of a computer

AI-generated content may be incorrect.

#### Tổng quát:

Độ hoàn chỉnh (Completeness)

* Completeness cao (100%) không có giá trị NULL nào

Cardinality (Đa dạng giá trị)

* Cardinality: 1768, tức là có 1768 giá trị ngày khác nhau trong 8807 bản ghi.
* Uniqueness/Distinctness: 20.07% → mức trùng lặp vừa phải, cho thấy một số ngày có nhiều nội dung được thêm cùng lúc (phù hợp với hoạt động cập nhật theo đợt).

#### Nhận xét:

Dữ liệu dạng văn bản (Object)

Kiểu dữ liệu: object

Chiều dài chuỗi:

* + - Tối thiểu: 11 ký tự
    - Tối đa: 19 ký tự

→ cho thấy một số định dạng bao gồm cả phần thời gian hoặc độ dài tên tháng (ví dụ: "September 15, 2019").

Hoàn chỉnh (Completeness): 100% → không có giá trị bị thiếu (NULL hoặc Missing).

Cardinality: 1768 → có 1768 ngày khác nhau trong tổng số 8807 bản ghi.

Uniqueness/Distinctness: ~20.07% → mức độ trùng lặp tương đối cao, chứng tỏ nhiều bản ghi được thêm vào cùng một ngày → phản ánh các đợt cập nhật nội dung hàng loạt (batch upload).

Ngày phổ biến nhất: January 1, 2020 (1.24%)

Cột date\_added chỉ ra được thời điểm phim được đăng lên Net-flix.

### release\_year

A screenshot of a computer

AI-generated content may be incorrect.

#### Tổng quát:

* Completeness cao (100%) không có giá trị NULL nào.
* Có 74 giá trị duy nhất, tương ứng với 74 năm khác nhau (từ đến 2021) cho phép phân tích xu hướng phát hành phim qua các thời kỳ.
* Uniqueness / Distinctness: 0.8402% → phần lớn các năm có nhiều bản ghi lặp lại. Ví dụ: riêng năm 2018 đã có 1.147 bản ghi.

#### Nhận xét:

* Hơn 57% nội dung đến từ giai đoạn 2016–2020 → cho thấy xu hướng nền tảng ưu tiên cập nhật nội dung mới, hiện đại, bắt kịp thị trường.
* Một vài năm xa xưa (như 1959, 1961) chỉ có 1 bản ghi → đại diện cho nội dung kinh điển hoặc hiếm, ít được phân phối hoặc cập nhật lại.

### rating:

A screenshot of a data sheet

AI-generated content may be incorrect.

#### Tổng quát:

* Độ hoàn chỉnh (Completeness) đạt 100%.
* Có 18 giá trị duy nhất → thể hiện nhiều mức phân loại độ tuổi, từ trẻ em đến người lớn sự đa dạng về loại hình nội dung, phục vụ nhiều nhóm đối tượng khác nhau.
* Chỉ số Uniqueness / Distinctness: 0.2044% → mỗi giá trị rating đều xuất hiện nhiều lần trong tập dữ liệu, bởi vì chỉ có 18 giá trị.

#### Nhận xét chi tiết:.

* Nội dung chủ yếu tập trung vào các phân loại TV-MA và TV-14, hướng đến khán giả trưởng thành và thanh thiếu niên – đây là tệp người dùng chính của các nền tảng như Netflix.
* NC-17, UR, 74 min, 84 min, 66 min chỉ xuất hiện 1–3 lần → chiếm < 0.04% mỗi loại.

### Duration

A screenshot of a data sheet

AI-generated content may be incorrect.

#### Tổng quát:

* Dữ liệu hoàn chỉnh 100%, không có giá trị NULL hoặc bị thiếu.
* Có 221 giá trị duy nhất, thể hiện rõ sự đa dạng về thời lượng nội dung (bao gồm cả thời lượng phim lẻ lẫn số mùa của TV Show).
* Uniqueness / Distinctness: 2.5094% → mỗi giá trị được lặp lại nhiều lần, hầu như thường lượng của các phim từng thể loại thường như .

#### Nhận xét chi tiết:

* TV Show chiếm tỷ trọng lớn với các season (đặc biệt là 1 Season), cho thấy nội dung dạng nhiều tập rất phổ biến. → Trong khi đó, phim lẻ phổ biến nhất có thời lượng 90 và 94 phút, phù hợp với độ dài trung bình của phim điện ảnh.
* Các thời lượng như 189 min, 10 min, 3 min, 229 min, 191 min chỉ xuất hiện 1 lần.

.

### Discription

A screenshot of a computer

AI-generated content may be incorrect.

#### Tổng quát:

* Dữ liệu hoàn chỉnh 100%, không có giá trị NULL hoặc bị thiếu.
* Uniqueness & Distinctness đều đạt 99.6367% .
* Những mô tả trùng lặp có thể là 1 bộ phim được chiếu lại hoặc 1 bản đặc biệt của 1 bộ phim nào đó.

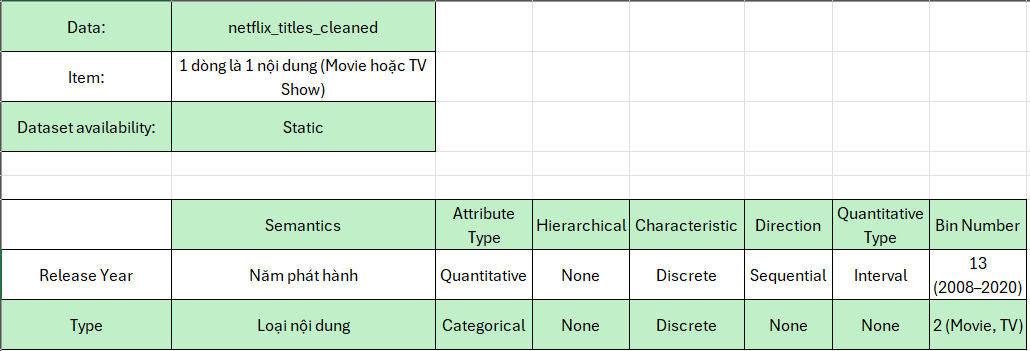
#### Nhận xét chi tiết:

* Mô tả ngắn nhất: 21 ký tự
* Mô tả dài nhất: 248 ký tự
* Tổng cộng có 104 mô tả có độ dài khác nhau → nội dung phim rất đa dạng không trùng lặp phù hợp với các thể loại.

# Abstraction

## Domain task 1: Phân tích xu hướng nội dung theo năm và loại hình.

### Data abastraction



* + 1. **Task abstraction**

#### Produce → Explore → Sum

* **Produce**
  + Kiểm tra giá trị thiếu trong cột Release Year và Type.
  + Loại bỏ dòng không có năm phát hành.
  + Gộp số lượng nội dung theo Release Year và Type.

#### Explore

#### Câu hỏi khám phá:

* + Có sự khác biệt giữa số lượng Movie và TV Show qua các năm không?
  + Năm nào có sự tăng trưởng đột biến?

**Hành động:**

* + Đếm số lượng nội dung theo từng năm (Release Year) chia theo loại (Movie, TV Show).
  + Vẽ biểu đồ Area Chart thể hiện sự thay đổi theo thời gian:
    - Trục X: Release Year
    - Trục Y: Count of Show ID
    - Màu sắc phân biệt Movie và TV Show

#### Sum

**Nhận xét:**

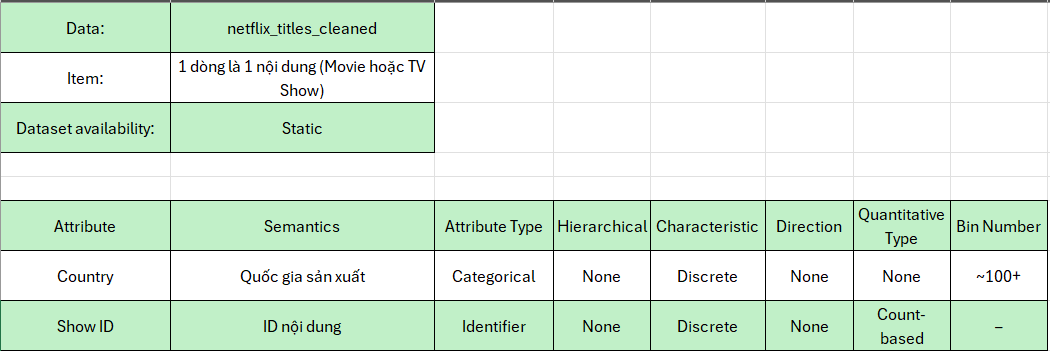
* + Từ năm 2016 trở đi, tổng số nội dung Netflix tăng nhanh rõ rệt.
  + Số lượng Movies luôn chiếm đa số, nhưng TV Shows có xu hướng tăng mạnh sau 2017.

**Tổng hợp phát hiện:**

* + Tăng trưởng nổi bật**:** Giai đoạn 2016–2019.
  + Movies chiếm tỷ trọng lớn nhưng TV Shows đóng vai trò ngày càng quan trọng trong chiến lược nội dung của Netflix.

## Domain task 2: Phân tích quốc gia sản xuất nhiều nội dung nhất

### Data abstraction



* + 1. **Task abstraction**

#### Produce → Explore → Sum

* **Produce**
  + Loại bỏ các giá trị thiếu hoặc "Không xác định" trong cột Country.
  + Tách các mục có nhiều quốc gia (nếu cần thiết) — ví dụ "India, United States" → tách thành 2 dòng (nếu bạn chọn làm kỹ hơn).
  + Đếm số lượng nội dung được sản xuất tại mỗi quốc gia.

#### Explore

#### Câu hỏi khám phá:

* + Quốc gia nào có nhiều nội dung Netflix nhất?
  + Có sự tập trung nội dung ở một vài nước hay phân tán?

Hành động:

* + Tạo biểu đồ Bar chart ngang (Horizontal Bar Chart):
    - Trục Y: Country
    - Trục X: Count of Show ID
    - Sắp xếp giảm dần theo số lượng
    - Chọn hiển thị Top 10 quốc gia

#### Sum

**Nhận xét:**

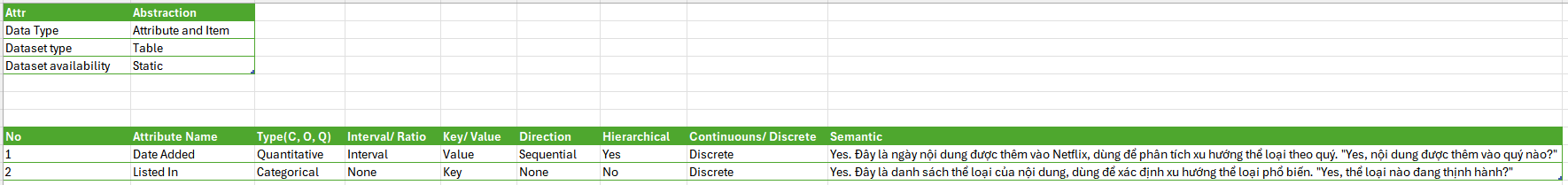
* + Hoa Kỳ (United States) là quốc gia có số lượng nội dung vượt trội — hơn 2000 mục.
  + Ấn Độ (India) và một số nước như Anh, Canada, Nhật cũng là những trung tâm sản xuất nội dung quan trọng.

**Tổng hợp phát hiện:**

* + Netflix có xu hướng tập trung nội dung từ một số quốc gia chính, chủ yếu nói tiếng Anh..
  + Điều này phản ánh chiến lược toàn cầu hóa nội dung nhưng vẫn ưu tiên những thị trường có sản lượng giải trí lớn.

## Domain task 3: Xu hướng thể loại phim được cập nhật nhiều trong các năm vừa qua

### Data abstraction



* + 1. **Task abstraction**

Phân tích **Analyze → Search → Query**

**Analyze:**

* **Consume:** Dữ liệu về ngày thêm nội dung (date\_added) và thể loại (listed\_in) của các bộ phim và chương trình trên Netflix đã có sẵn.
* **Biểu diễn:** Có thể dùng bảng thống kê hoặc biểu đồ (ví dụ: biểu đồ đường để thể hiện xu hướng thể loại qua các quý, hoặc biểu đồ cột để so sánh tỷ lệ các thể loại trong một quý cụ thể) → **Present**

**Search:**

* Không cần tìm kiếm một bộ phim hoặc chương trình cụ thể mà muốn quan sát xu hướng chung của các thể loại qua các quý.
* Không phù hợp với Lookup hay Locate do domain task không nhắm đến việc tìm kiếm một nội dung cụ thể mà tập trung vào việc khám phá mối quan hệ giữa thời gian (date\_added) và thể loại (listed\_in) → Explore

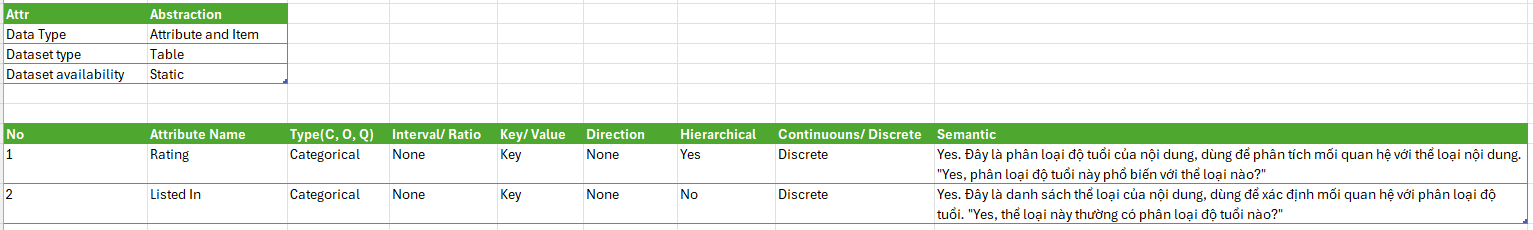
#### Query:

- Tìm được xu hướng của các thể loại qua các quý (ví dụ: thể loại nào đang tăng hoặc giảm qua thời gian) và tổng hợp số lượng hoặc tỷ lệ các thể loại theo từng quý → Summarize

**=> Present** → **Explore** → **Summarize**

## Domain task 4: Phân tích mối quan hệ giữa Phân loại Độ tuổi và Thể loại Nội dung

### Data abstraction



* + 1. **Task abstraction**

Phân tích theo **Analyze → Search → Query**

#### Analyze:

* **Consume**: Dữ liệu về phân loại độ tuổi (rating) và thể loại (listed\_in) của các nội dung trên Netflix đã có sẵn trong file netflix\_titles\_cleaned.csv.
* **Biểu diễn:** Có thể sử dụng bảng thống kê hoặc biểu đồ trực quan để trình bày mối quan hệ giữa phân loại độ tuổi và thể loại nội dung. Một Treemap có thể hiển thị số lượng nội dung cho từng cặp rating và listed\_in, với kích thước và màu sắc hình chữ nhật biểu thị số lượng. → **Present**

#### Search:

* Không cần tìm kiếm một bộ phim hoặc chương trình cụ thể mà tập trung vào việc quan sát mối quan hệ tổng thể giữa phân loại độ tuổi và thể loại nội dung.
* Không phù hợp với Lookup hay Locate vì domain task "Phân tích mối quan hệ giữa Phân loại Độ tuổi và Thể loại Nội dung" không nhằm tìm kiếm một nội dung cụ thể (như tra cứu một bộ phim) hay định vị một mục (như tìm vị trí của một chương trình). Thay vào đó, phân tích được mối liên hệ giữa rating và listed\_in, ví dụ: xem "TV-MA" thường gắn với thể loại nào → **Explore**

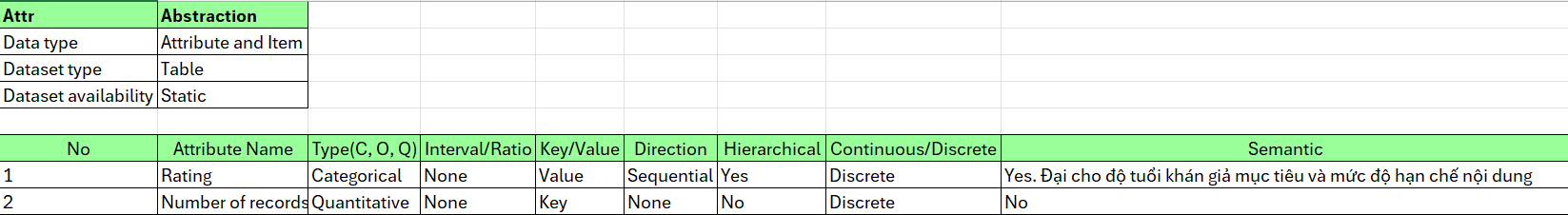
**Query:**

* Tìm được mối quan hệ giữa phân loại độ tuổi và thể loại nội dung (ví dụ: nội dung "TV-Y" thường thuộc "Kids' TV", hay "TV-MA" thường gắn với "TV Dramas" và "International TV Shows") và tổng hợp số lượng hoặc tỷ lệ các thể loại theo từng phân loại độ tuổi để rút ra kết luận → **Summarize**

=> **Present → Explore → Summarize**

## Domain task 5: So sánh số lượng nội dung giữa các nhóm độ tuổi khán giả

### Data abstraction



* + 1. **Task abstraction**

#### Phân tích theo quy trình Analyze → Search → Query

#### Analyze:

* **Consume**:
  + Dữ liệu về định mức độ tuổi (Rating) và số lượng nội dung (Number of records) của các bộ phim và chương trình trên Netflix đã có sẵn.
  + Chi tiết:
    - Rating là các giá trị phân loại (G, TV-G, TV-Y, ..., TV-MA, NC-17, NR, UR), đại diện cho độ tuổi khán giả mục tiêu và mức độ hạn chế nội dung.
    - Number of records là số lượng nội dung tương ứng với mỗi định mức độ tuổi, đã được tổng hợp sẵn.
* **Biểu diễn**:
  + Có thể dùng biểu đồ cột để thể hiện số lượng nội dung (Number of records) theo từng định mức độ tuổi (Rating).
  + Trong biểu đồ hiện tại, Rating được sắp xếp theo thứ tự tự nhiên (G → UR) để làm nổi bật xu hướng theo độ tuổi khán giả.

→ **Present**: Dữ liệu đã sẵn sàng để trình bày trực tiếp dưới dạng biểu đồ cột.

**Search:**

* **Không cần tìm kiếm một định mức độ tuổi cụ thể**:
  + Biểu đồ không tập trung vào việc tìm kiếm một định mức cụ thể (ví dụ: chỉ xem số lượng nội dung của TV-MA), mà quan sát xu hướng và phân bố chung trên toàn bộ các định mức độ tuổi (G → UR).
* **Không phù hợp với Lookup hay Locate**:
  + Domain task không nhắm đến việc tìm kiếm một giá trị cụ thể (Lookup, ví dụ: số lượng nội dung của TV-MA) hay xác định vị trí của một giá trị (Locate, ví dụ: TV-MA nằm ở đâu trong danh sách).
  + Thay vào đó, domain task tập trung vào việc khám phá mối quan hệ giữa số lượng nội dung (Number of records) và độ tuổi khán giả mục tiêu (thể hiện qua thứ tự Rating).

→ **Explore**: Nhiệm vụ là khám phá xu hướng và phân bố trên toàn bộ dữ liệu.

**Query:**

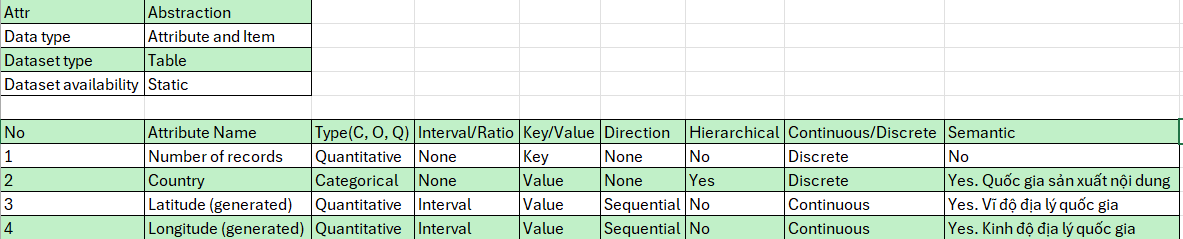
* **Tìm xu hướng và phân bố**:
  + Domain task yêu cầu tìm xu hướng của số lượng nội dung qua độ tuổi khán giả (từ trẻ em đến người lớn) và tổng hợp số lượng nội dung theo từng định mức độ tuổi.
  + Ví dụ:
    - Số lượng nội dung tăng hay giảm như thế nào khi độ tuổi khán giả tăng (G → UR)?
    - Định mức độ tuổi nào có số lượng nội dung lớn nhất?

→ **Summarize**: Tóm tắt xu hướng và phân bố trên toàn bộ dữ liệu

**=> Present → Explore → Summarize**.

## Domain task 6: So sánh số lượng nội dung giữa các quốc gia sản xuất

### Data abstraction



* + 1. **Task abstraction**

#### Phân tích theo quy trình Analyze → Search → Query

#### Analyze

**Consume:**

* **Dữ liệu về quốc gia sản xuất (Country), số lượng nội dung (CNT), và vị trí địa lý (Latitude, Longitude):** Dữ liệu đã có sẵn từ tập dữ liệu netflix\_titles\_cleaned.csv.
  + **Country**: Giá trị phân loại (ví dụ: United States, India, South Korea), đại diện cho quốc gia sản xuất nội dung.
  + **CNT (Number of records)**: Số lượng nội dung tương ứng với mỗi quốc gia, đã được tổng hợp sẵn (ví dụ: United States: 2,818, India: 912).
  + **Latitude và Longitude**: Các trường được tạo tự động trong Tableau để định vị các quốc gia trên bản đồ.

**Chi tiết:**

* **Country** là giá trị phân loại, phân biệt các quốc gia sản xuất nội dung.
* **CNT** là giá trị định lượng, biểu thị số lượng nội dung được sản xuất bởi mỗi quốc gia.
* **Latitude và Longitude** là giá trị thứ tự, được sử dụng để ánh xạ vị trí địa lý của các quốc gia trên bản đồ.

**Biểu diễn:**

* Bản đồ hiện tại sử dụng **biểu đồ bản đồ (Map)** để thể hiện số lượng nội dung (CNT) theo quốc gia (Country), với các quốc gia được định vị bằng Latitude và Longitude.
* **Mã hóa:** Số lượng nội dung (CNT) được mã hóa bằng màu sắc (thang màu từ vàng nhạt đến đỏ đậm), trong đó màu đậm hơn biểu thị số lượng nội dung lớn hơn (ví dụ: United States với màu đỏ đậm có 2,818 tiêu đề).
* **Tập trung:** Biểu đồ tập trung vào sự phân bố địa lý và so sánh số lượng nội dung giữa các quốc gia.

**→ Present:** Dữ liệu đã sẵn sàng để trình bày trực tiếp dưới dạng bản đồ với màu sắc mã hóa số lượng nội dung.

#### Search

**Không cần tìm kiếm một quốc gia hoặc khu vực cụ thể:**

* Bản đồ không tập trung vào việc tìm kiếm số lượng nội dung của một quốc gia cụ thể (ví dụ: chỉ xem United States) hay một khu vực cụ thể (chỉ xem Bắc Mỹ), mà quan sát sự phân bố chung trên toàn bộ các quốc gia và khu vực địa lý.
* Domain task nhắm đến việc so sánh số lượng nội dung giữa các quốc gia và khám phá các khu vực địa lý có sản xuất nội dung lớn.

**Không phù hợp với Lookup hay Locate:**

* Domain task không nhắm đến việc tìm kiếm một giá trị cụ thể (Lookup, ví dụ: số lượng nội dung của United States) hay xác định vị trí của một quốc gia trên bản đồ (Locate, ví dụ: United States nằm ở đâu).
* Thay vào đó, nhiệm vụ tập trung vào việc khám phá mối quan hệ giữa số lượng nội dung (CNT) và vị trí địa lý (Country), đồng thời so sánh giữa các quốc gia và khu vực.

**→ Explore:** Nhiệm vụ là khám phá sự phân bố và so sánh trên toàn bộ dữ liệu.

#### Query

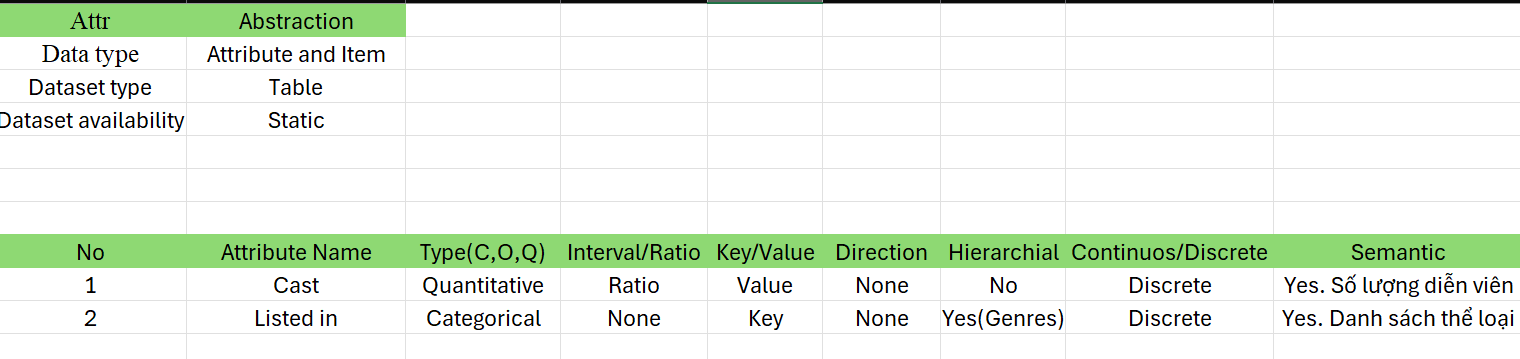
**Tìm sự phân bố và so sánh:**

* Domain task yêu cầu tìm sự phân bố số lượng nội dung theo quốc gia và so sánh giữa các khu vực địa lý (ví dụ: Bắc Mỹ, Nam Á, Đông Á).
* Ví dụ:
  + Khu vực nào sản xuất nhiều nội dung nhất (ví dụ: Bắc Mỹ với United States và Canada)?
  + Sự chênh lệch về số lượng nội dung giữa các quốc gia lớn như United States và India so với các quốc gia nhỏ hơn như Guatemala hay Cambodia là bao nhiêu?

**→ Summarize:** Tóm tắt sự phân bố và so sánh trên toàn bộ dữ liệu.  **=> Present → Explore → Summarize.**

## Domain task 7: Phân phối diễn viên theo thể loại

### Data abstraction



* + 1. **Data abstraction**

#### Analyze:

* **Consume:**
  + Dữ liệu về số lượng diễn viên (Cast), thể loại (Listed in).
  + Chi tiết:
    - Listed in là các giá trị phân biệt thể loại.
    - Số lượng diễn viên (cast).

#### → Present Search:

* **Không cần tìm kiếm thể loại cụ thể**:
  + Biểu đồ không tập trung vào việc tìm kiếm số lượng diễn viên của một thể loại cụ thể mà quan sát phân phối số lượng diễn viên theo các thể loại.
* **Không phù hợp với Lookup hay Locate**:
  + Domain task không nhắm đến việc tìm kiếm một giá trị cụ thể hay xác định vị trí của một giá trị (Lookup hay Locate).
  + Thay vào đó, domain task tập trung vào việc khám phá phân phối số lượng diễn viên theo thể loại phim.

→ **Explore**: Nhiệm vụ là khám phá phân phối và so sánh trên toàn bộ dữ liệu.

#### Query:

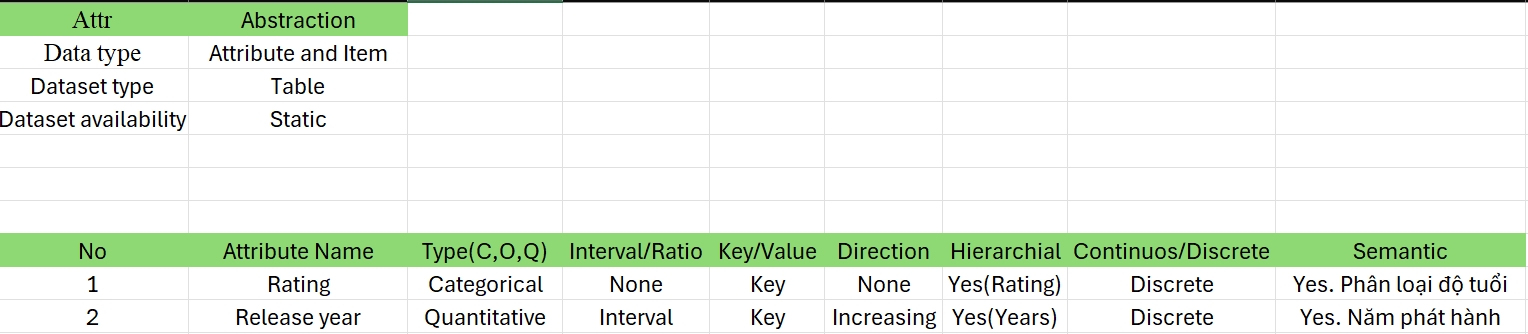
* **Tìm phân phối và so sánh**:
  + Domain task yêu cầu tìm phân phối của diễn viên theo từng thể loại để có thể so sánh và đánh giá
  + Ví dụ: Dựa trên trực quan hóa, đưa ra được thể loại nào có số lượng diễn viên nhiều nhất, ít nhất?

→ **Summarize**: Tóm tắt phân phối dữ liệu và so sánh trên toàn bộ dữ liệu.

**=> Present → Explore → Summarize**

## Domain task 8: Sự phân bố và xu hướng phân loại độ tuổi theo năm phát hành

### Data abstraction



* + 1. **Task abstraction**

#### Analyze:

* **Consume** 
  + Dữ liệu về số lượng phân loại độ tuổi (rating), năm phát hành (Released year).
  + Chi tiết:
    - Release year năm phát hành của các phim có các phân loại độ tuổi.
    - Số lượng về phân loại độ tuổi (rating).

#### → Present

#### Search:

* **Không cần tìm kiếm thể loại cụ thể**:
  + Biểu đồ không tập trung vào việc tìm kiếm phân loại độ tuổi hay theo năm phát hành cụ thể, mà tập trung vào phân bố và xu hướng của các phân loại độ tuổi theo năm phát hành.
* **Không phù hợp với Lookup hay Locate**:
  + Domain task không nhắm đến việc tìm kiếm một giá trị cụ thể hay xác định vị trí của một giá trị (Lookup hay Locate).
  + Thay vào đó, domain task tập trung vào việc khám phá phân phối số lượng phim có phân loại độ tuổi theo năm phát hành

→ **Explore**: Nhiệm vụ là khám phá phân phối và xu hướng trên toàn bộ dữ liệu.

* **Query:**
* **Tìm phân phối và so sánh**:
  + Domain task yêu cầu tìm phân phối của số lượng phim có phân loại theo độ tuổi theo năm phát hành.
  + Ví dụ:
    - Dựa trên trực quan hóa, đưa ra được xu hướng các phân loại theo độ tuổi qua từng năm phát hành?
    - Năm nào, thể loại nào có số lượng phân loại theo độ tuổi phát hành cao nhất, thấp nhất ?

→ **Summarize**: Tóm tắt phân phối và xu hướng dữ .

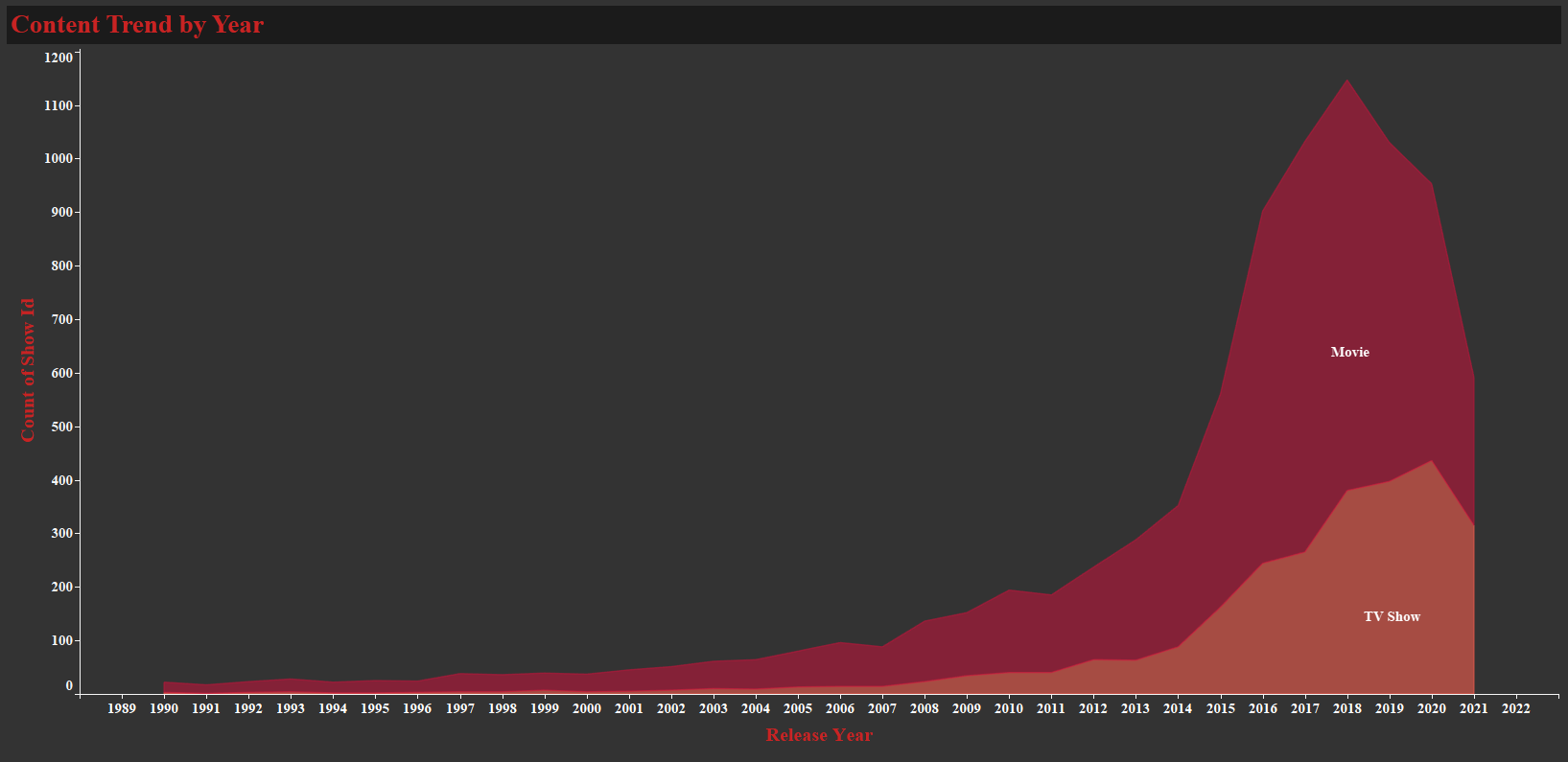
**=> Present → Explore → Summarize**

1. **Thiết kế idiom** 
   1. **Domain task 1**
      1. **Idiom**

|  |  |  |
| --- | --- | --- |
| **Idiom** | **Bar charts** | |
| **Data** | Content Type: C (Categorical, TV Show) | |
| Release Year: O (Ordinal - Năm phát hành) | |
| Count of Show ID: Q (Quantitative - số lượng nội dung) | |
|  | Mark: area | |
| **Encode** | Channel | O: pos ngang |
| Q: pos dọc |

|  |  |  |
| --- | --- | --- |
|  |  | C: màu sắc (phân biệt giữa Movie và TV Show) |
| **TASK** | *Theo dõi xu hướng tăng/giảm số lượng nội dung theo thời gian*  *So sánh giữa Movie TV và TV Show qua từng năm* | |
| **SCALE** | *Keys: ~13 (2008-2020) Level: Ordinal* | |

* + 1. **Biểu đồ**



* + 1. **Đánh giá**

#### Tính biểu đạt

* + - * + Biểu đồ biểu diễn đầy đủ các năm phát hành và số lượng nội dung tương ứng.
        + Phân biệt rõ hai loại nội dung bằng màu sắc nổi bật.
        + Diện tích và độ dốc thể hiện tốc độ tăng trưởng.

#### Tính hiệu quả

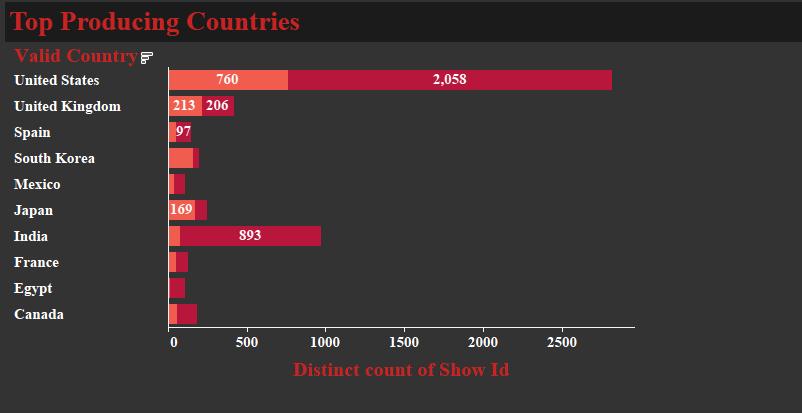
* + - * + Accuracy: Trục Y biểu diễn số lượng giúp đọc giá trị chính xác.
        + Discriminablity: Màu sắc phân biệt rõ Movie và TV Show.
        + Separability: Hai loại nội dung hiển thị cùng biểu đồ giúp dễ so sánh tỷ trọng theo từng năm.

#### Phân tích biểu đồ

* + - * + Giúp xác định độ tuổi có nguy cơ mắc bệnh tim cao nhất và xu hướng theo độ tuổi.
  1. **Domain task 2**
     1. **Idiom**

|  |  |  |
| --- | --- | --- |
| **Idiom** | **Pie charts** | |
| **Data** | Country: O (Ordinal -> tên quốc gia) | |
| Content Type: C (Categorical – Movie, TV Show) | |
| Count of Show ID: Q (Quantitative - số lượng nội dung) | |
|  | Mark: bar | |
| **Encode** | Channel | O: pos dọc (quốc gia - trục Y) |
| Q: pos ngang (số lượng - trục X) |
| C: màu sắc (phân biệt giữa Movie và TV Show) |
| **TASK** | So sánh mức độ sản xuất nội dung giữa các quốc gia.  Quan sát quốc gia nào chiếm ưu thế và loại hình phổ biến tại từng nước. | |
| **SCALE** | Keys: 10 (Top 10 quốc gia)  Level: Ordinal (xếp hạng theo số lượng) | |

* + 1. **Biểu đồ**



* + 1. **Đánh giá**

#### Tính biểu đạt

* + - * + Biểu tròn giúp thể hiện rõ tỉ lệ từng nhóm.
        + Dễ nhận diện mức độ chênh lệch giữa nam và nữ.

#### Tính hiệu quả

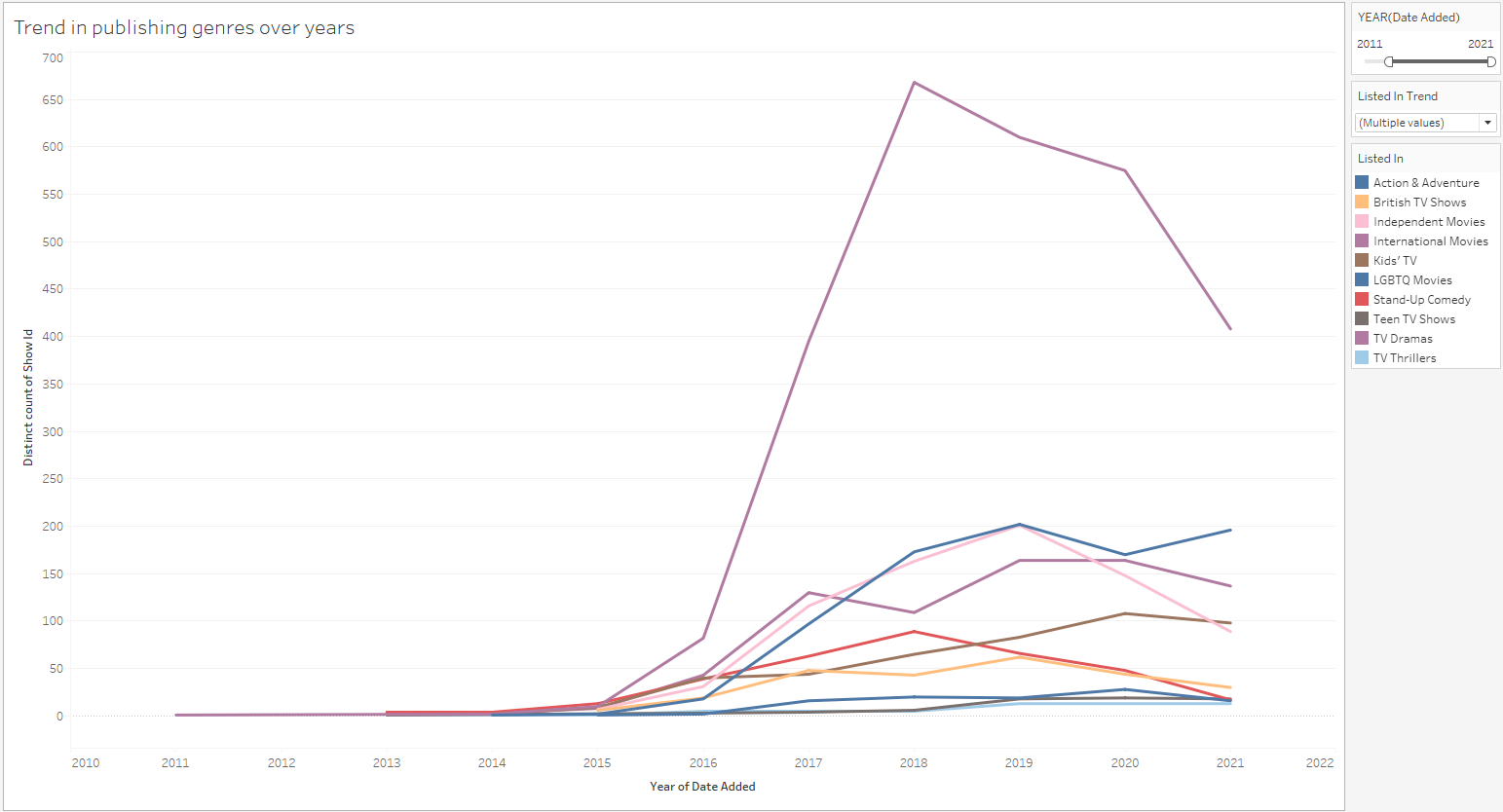
* + - * + Accuracy: giúp dễ dàng hình dung tỷ lệ giữa các nhóm.
        + Discriminablity: sử dụng màu sắc giúp phân biệt giữa có/không có bệnh tim.
        + Separability: không xét vì không có thuộc tính nào sử dụng 2 channel để biểu diễn.

#### Phân tích biểu đồ

* + - * + Giúp xác định giới tính nào có tỷ lệ mắc bệnh tim cao hơn và mức độ chênh lệch giữa hai giới.
  1. **Domain task 3**
     1. **Idiom**

|  |  |  |
| --- | --- | --- |
| **Idiom** | **Lines** | |
| **Data** | Count: Q (Quantitative) | |
| Listed In: C (Categorical) | |
| Date Added: Q (Quantitative) | |
| **Encode** | Mark: area chart | |
| Channel | Q: pos dọc. Số lượng nội dung (Count) được mã hóa bằng vị trí dọc trên trục Y |
| Q: pos ngang. Thời gian (date\_added) được mã hóa bằng vị trí ngang trên trục X. |
| C: Color: Các cặp listed\_in (ví dụ: Kids' TV") được mã hóa bằng màu sắc. |
| **TASK** | So sánh số lượng nội dung của từng thể loại được đăng tải lên netflix qua các năm.  Xu hướng thể loại được đăng tải lên netflix qua các quý. | |
| **SCALE** | Keys:10  Level: Categorical (Listed In), Quantitative (Date Added) | |

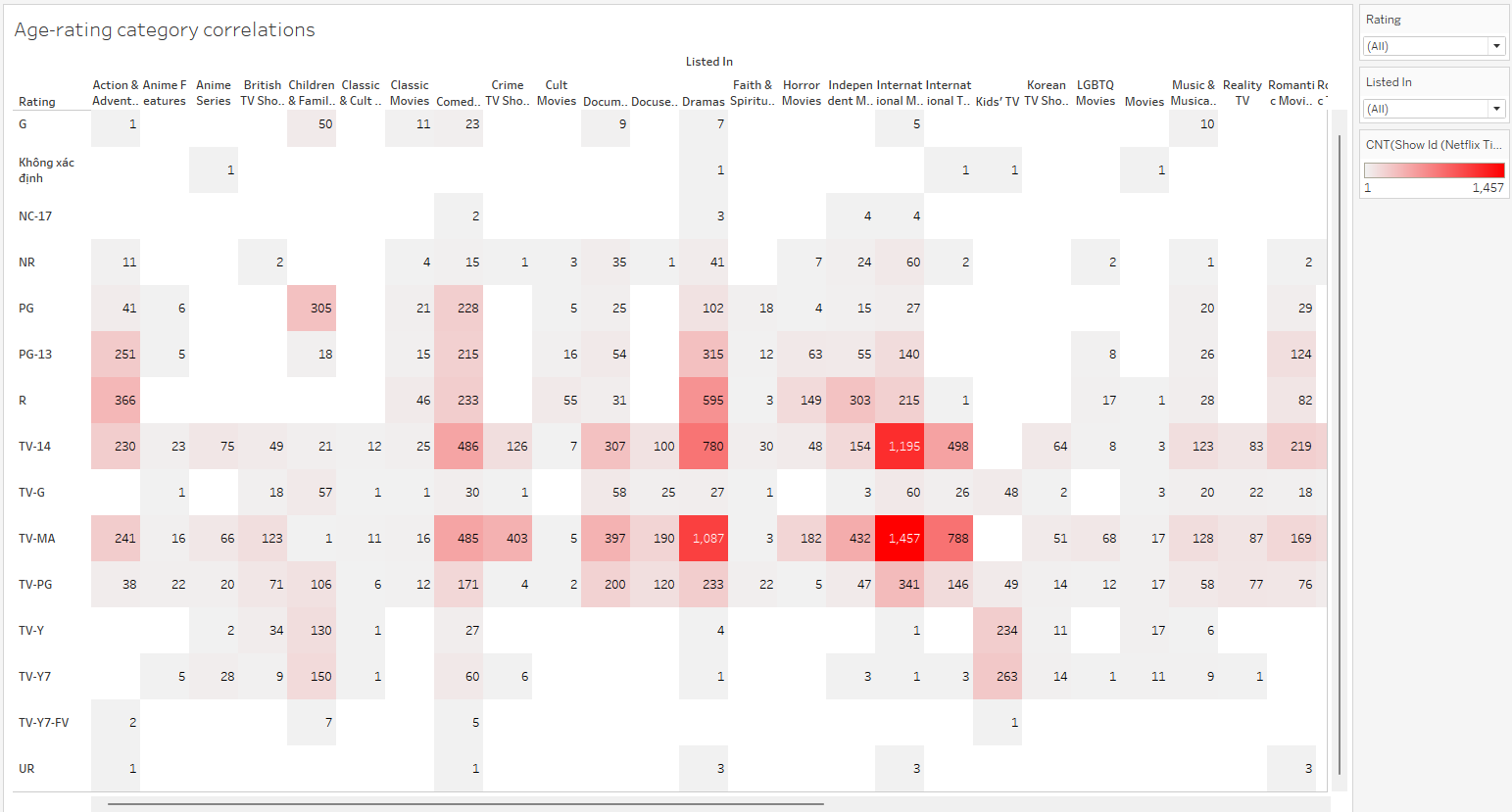
* + 1. **Biểu đồ**



* + 1. **Đánh giá**
* **Tính Biểu đạt**
  + - Số lượng nội dung được hiển thị rõ ràng trên mỗi điểm. Lines số lượng nội dung được biểu thị bằng chiều cao vùng trên trục Y.
    - Xu hướng thể loại được hiển thị đầy đủ, được biểu thị bằng các đường màu khác nhau. Legend được hiện thị đầy đủ giúp người xem dễ nhận biết.
    - Hỗ trợ phân tích xu hướng, Lines rất hiệu quả trong việc hiển thị xu hướng theo thời gian. Người xem có thể thấy số lượng nội dung của listed\_in thay đổi qua các quý.
* **Tính hiệu quả**
  + - Accuracy: Area Chart sử dụng channel Position dọc (trục Y) để biểu thị số lượng nội dung (Count)
    - Discriminability: Màu sắc là channel tốt để phân biệt các danh mục phân loại (categorical).
    - Separability: Không xét vì không có thuộc tính nào sử dụng 2 channel để biểu diễn cùng một dữ liệu.
* **Phân tích biểu đồ**
  + Số lượng nội dung tăng mạnh từ 2015, đạt đỉnh vào khoảng 2019-2020 (khoảng 160-170 nội dung mỗi quý), sau đó giảm dần vào 2021.
  + Thể loại phổ biến:
    - Qua các quý thấy được thể loại Documentaries là có sự biến động nhiều nhất. Đặc biệt là quý 2 năm 2018 có 34 bộ phim được đăng tải lên nền tảng nhiều nhất trong khoảng thời gian từ năm 2009-2021.
    - Các thể loại “Children & Family Movies”, “Comedies”, “Dramas”, “International Movies” và “Romantic Movies” luôn ổn định nhưng không chiếm số lượng quá lớn.
* Xu hướng:
  + Trước 2015, số lượng nội dung rất ít, chủ yếu là “Children & Family Movies”, “Comedies” và “Stand-Up Comedy”.
  + Từ 2015-2019, các thể loại như "International TV Shows", "TV Dramas", và "Comedies" tăng mạnh, phản ánh chiến lược mở rộng nội dung quốc tế và giải trí của Netflix.
  + Sau 2020, số lượng nội dung giảm.
  1. **Domain task 4**
     1. **Idiom**

|  |  |  |
| --- | --- | --- |
| **Idiom** | **Horizontal Bars** | |
| **Data** | Rating: C (Categorical) | |
| Listed In: C (Categorical) | |
| Count: Q (Quantitative) | |
| **Encode** | Mark: rectangle | |
| Channel | Q: Label (Chữ số hiện thị số lượng) |
| C: Color (Màu sắc biểu thị số lượng nội dung) |
| **TASK** | So sánh số lượng nội dung theo phân loại độ tuổi và thể loại nội dung.  Nhận diện các cặp phân loại độ tuổi và thể loại nội dung phổ biến nhất. | |
| **SCALE** | Keys: 12 (Rating), 30 (Listed In)  Level: Categorical (cho Rating và Listed In), Quantitative (cho Count) | |

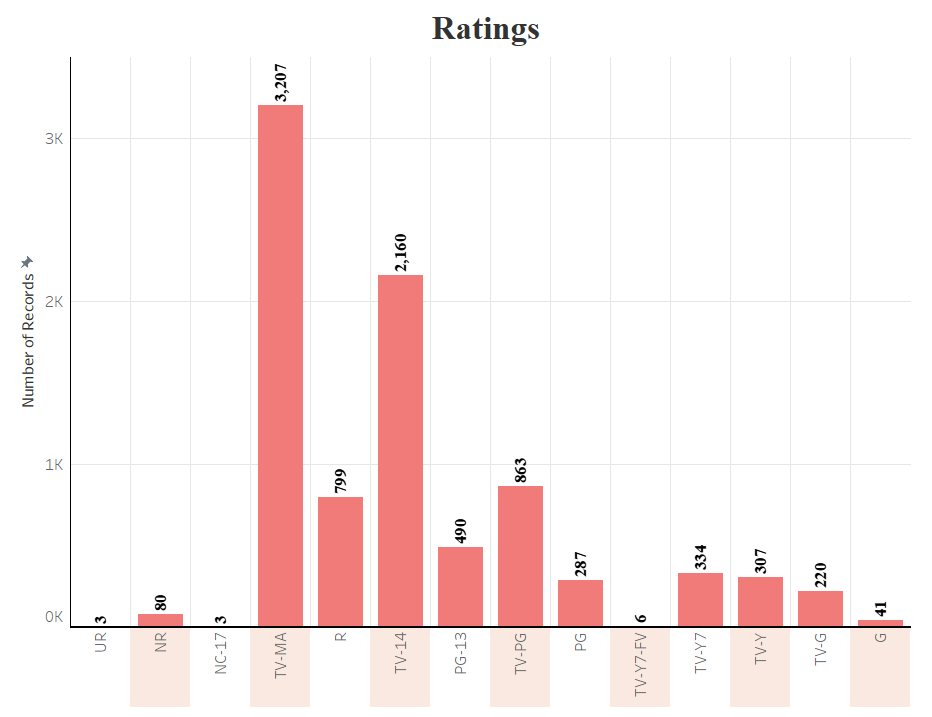
* + 1. **Biểu đồ**



* + 1. **Đánh giá**
* **Tính biểu đạt:**
  + Horizontal Bars sử dụng màu sắc và chữ số để biểu thị số lượng nội dung, với những nội dung có số lượng cao thì độ tương quan cũng cao biểu đạt tốt để nhận diện xu hướng tổng thể, nhưng vẫn còn không tốt khi biểu thị giá trị chính xác.
  + Số lượng nội dung dựa trên các cặp rating và listed\_in được hiện thị đầy đủ thông tin trên biểu đồ.
  + Horizontal Bars giúp so sánh trực quan số lượng nội dung giữa các cặp thông tin đồng thời dễ dàng nhận biết được mối quan hệ của các thuộc tính.
* **Tính hiệu quả:**
  + Accuracy: Sử dụng channel và Color để biểu thị số lượng nội dung, với những nơi có số lượng cao thấy được mối quan hệ cao giữa 2 thuộc tính.
  + Discriminability: Các phân loại độ tuổi (rating) và thể loại (listed\_in) được phân biệt rõ ràng nhờ Position dọc và Position ngang trên trục Y và X. Mỗi ô đại diện cho một cặp duy nhất, không có sự chồng lấn.
  + Separability: Không xét vì không có thuộc tính nào sử dụng 2 channel để biểu diễn cùng một dữ liệu. Rating dùng Position dọc, Listed In dùng Position ngang, và Count dùng Color. Không có sự chồng lấn channel gây nhiễu.
* **Phân tích biểu đồ:**
  + Cặp phổ biến nhất: Các ô đậm nhất có giá trị hơn 1000 là:
    - TV-14 với International Movies
    - TV-MA với International Movies
* Thấy được các bộ phim có độ tuổi TV-14 và TV-MA có hầu hết các loại nội dung.
* Ở xếp hạng độ tuổi PG, PG-13 và R chỉ có các thể loại phim Action & Adventure, Children & Family Movie, Comedies và Dramas còn các thể loại khác gần như là không xuất hiện.
* Với các xếp hạng độ tuổi TV-Y, TV-Y7 chỉ có Children & Family Movie, Comedies, Kids’ TV và TV Comedies.
* Các nhóm tuổi trẻ em thì số lượng, thể loại phim bị giới hạn nên số lượng ít.
* 2 nhóm còn lại thể loại phong phú trải dài hầu hết các thể loại.
  1. **Domain task 5**
     1. **Idiom**

|  |  |  |
| --- | --- | --- |
| **Idiom** | **Bar charts** | |
| **Data** | Rating: C (Categorical) | |
| Listed In: C (Categorical) | |
| **Encode** | Mark: bar chart | |
| Channel | Q: pos dọc |
| C: pos ngang |
| **TASK** | So sánh số lượng nội dung dựa trên thứ hạng đánh giá kèm độ tuổi của khán giá  Xu hướng phân phối theo định mức độ tuổi | |
| **SCALE** | Keys: 14  Level: Categorical | |

* + 1. **Biểu đồ**



* + 1. **Đánh giá**

#### Tính biểu đạt

* + - * + Số lượng nội dung được hiển thị rõ ràng trên mỗi đỉnh cột, giúp người xem dễ dàng nắm bắt giá trị chính xác.
        + Các loại đánh giá (Rating) được hiển thị đầy đủ, không bị cắt xén, giúp dễ dàng nhận diện từng danh mục.
        + Hỗ trợ phân tích xu hướng: người xem có thể thấy số lượng nội dung tăng từ trẻ em (G:41) đến thanh thiếu niên (TV-14: 2160) và người lớn (TV-MA: 3207)
        + Tiêu đề đúng nội dung biểu đồ, giúp người xem hiểu ngay được mục đích của hình ảnh hóa.

#### Tính hiệu quả

* + - * + Accuracy: sử dụng channel vị trí (trục Y) giúp đọc số liệu chính xác.
        + Discriminablity: các định mức độ tuổi được phân biệt rõ ràng nhờ vị trí trên trục X.
        + Separability: không xét vì không có thuộc tính nào sử dụng 2 channel để biểu diễn.

#### Phân tích biểu đồ

* + - * + TV-MA là loại đánh giá phổ biến nhất với 3207 nội dung, sau đó là TV-14 (2160) và TV-PG (863).
        + Các loại khác như UR và NC-17 có số lượng rất ít (3 mỗi loại) cho thấy nội dung này không phổ biển trên Netflix.
        + Phân nhóm:

Nhóm trẻ em (G, TV-G, TV-Y, TV-Y7, TV-Y7-FV): 908 nội dung.

Nhóm thanh thiếu niên (PG, TV-PG, PG-13, TV-14): 3,800 nội dung.

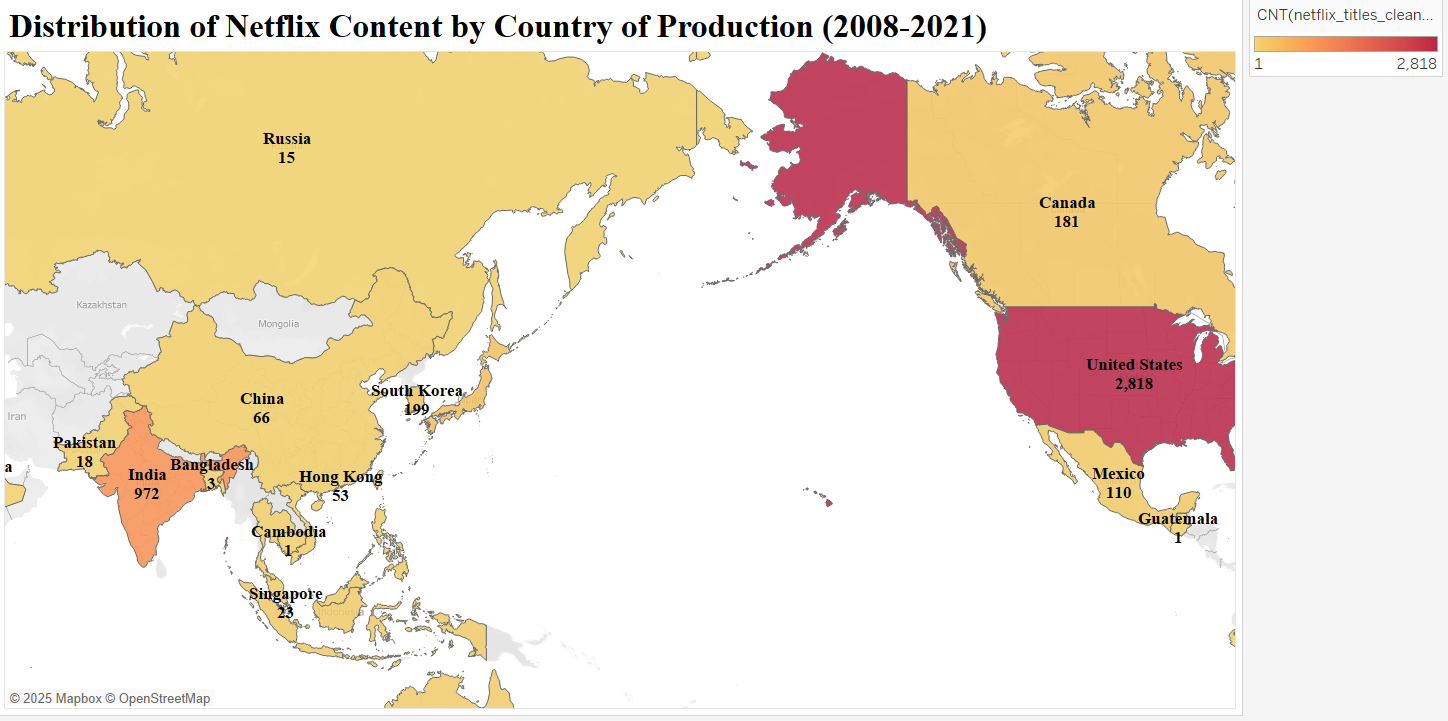
Nhóm người lớn (R, TV-MA, NC-17): 4,009 nội dung.

Nhóm không xếp hạng (NR, UR): 83 nội dung.

* 1. **Domain task 6**
     1. **Idiom**

|  |  |  |
| --- | --- | --- |
| **Idiom** | **Line charts** | |
| **Data** | Country: C | |
| Latitude (generated): Q1 | |
| Longitude (generated): Q2 | |
| Count: Q | |
| **Encode** | Mark: line | |
| Channel | Q1: pos dọc |
| C: màu sắc |
| Q: title |
| Q2: pos ngang |
| **TASK** | So sánh số lượng nội dung giữa các quốc gia qua các khu vực địa lý | |
| **SCALE** | Level: Nominal, Interval | |

* + 1. **Biểu đồ**



* + 1. **Đánh giá**

#### Tính biểu đạt (Expressiveness)

* Biểu đồ bản đồ thể hiện rõ sự phân bố số lượng nội dung Netflix theo quốc gia, với màu sắc mã hóa số lượng tiêu đề (CNT).
* Tiêu đề “Distribution of Netflix Content by Country of Production (2008-2021)” phản ánh đúng nội dung của biểu đồ, giúp người xem hiểu mục đích của hình ảnh trực quan.
* Nhãn quốc gia (Country) hiển thị trực tiếp trên bản đồ, được mã hóa bằng vị trí địa lý và màu sắc, giúp người xem dễ dàng nhận diện các quốc gia và số lượng nội dung tương ứng.

### Tính hiệu quả (Effectiveness)

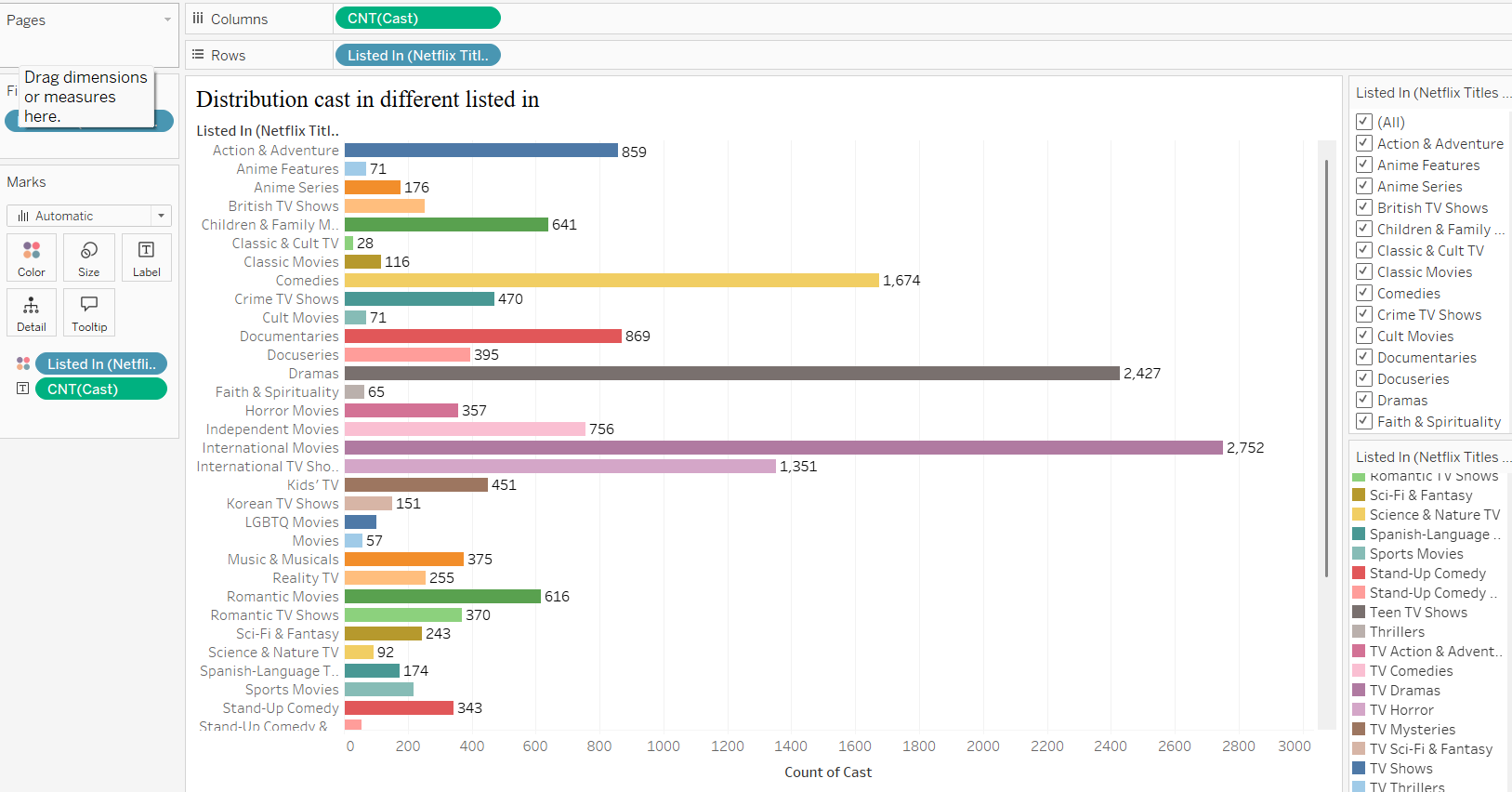
* **Accuracy**: Sử dụng channel vị trí (Latitude, Longitude) trên bản đồ để định vị quốc gia một cách chính xác, và màu sắc để biểu thị số lượng nội dung (CNT), giúp người xem dễ dàng so sánh.
* **Discriminability**: Sử dụng thang màu từ vàng nhạt đến đỏ đậm để phân biệt số lượng nội dung, dễ dàng nhận diện các quốc gia có nhiều nội dung (màu đỏ đậm) và ít nội dung (màu vàng nhạt).
* **Separability**: Không xét vì không có thuộc tính nào sử dụng 2 channel để biểu diễn cùng lúc.

#### Phân tích biểu đồ

* Thể hiện rõ sự phân bố địa lý của nội dung Netflix, với Hoa Kỳ (2,818 tiêu đề) và Ấn Độ (912 tiêu đề) là hai quốc gia dẫn đầu về số lượng nội dung.
* Bắc Mỹ (Hoa Kỳ, Canada, Mexico) là khu vực có sản lượng nội dung lớn nhất, trong khi Nam Á (Ấn Độ, Pakistan, Bangladesh) và Đông Á (Hàn Quốc, Nhật Bản, Hồng Kông) cũng đóng góp đáng kể. Các quốc gia nhỏ hơn như Guatemala và Campuchia có rất ít nội dung (1 tiêu đề).
  1. **Domain task 7**
     1. **Idiom**

|  |  |  |
| --- | --- | --- |
| **Idiom** | **Bar charts** | |
| **Data** | Cast: Q | |
| Listed in: C | |
| **Encode** | Mark: bar | |
| Channel | Q: pos ngang |
| C: màu sắc |
| C: pos dọc |
| **TASK** | Phân phối số lượng diễn viên theo từng thể loại phim | |
| **SCALE** | Keys: 42  Level: Nominal | |

* + 1. **Biểu đồ**



* + 1. **Đánh giá**

#### Tính biểu đạt

* + - * + Biểu đồ cột thể hiện rõ số lượng diễn viên đóng theo từng thể loại
        + Tiêu đề phản ảnh đúng nội dung của biểu đồ, cho người xem biết mục đích của hình ảnh trực quan.
        + Nhãn “Listed in” và “count (cast)” biểu đạt rõ nội dung mà người dùng cần làm.

#### Tính hiệu quả

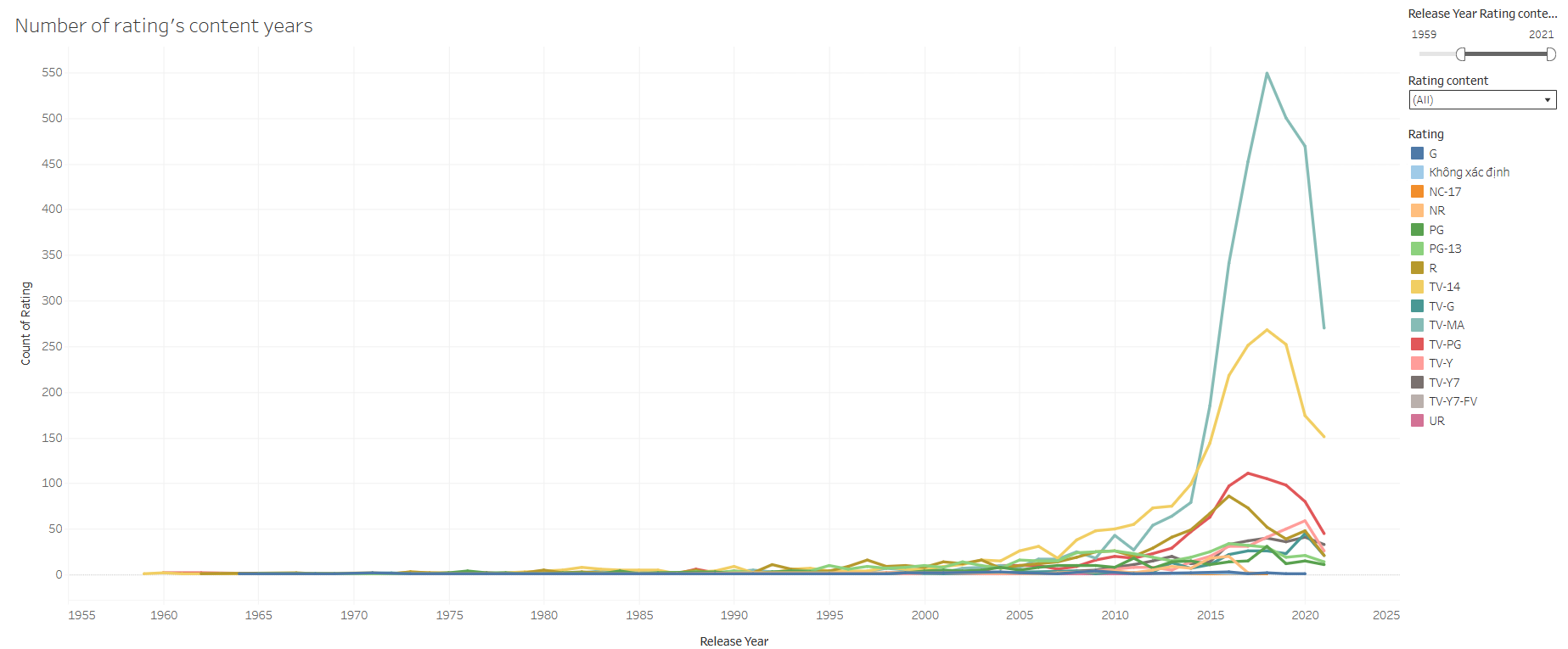
* + - * + Accuracy: sử dụng channel vị trí (trục Y) giúp phân biệt thể loại chính xác.
        + Discriminablity: Sử dụng màu sắc để phân biệt thể loại phim.
        + Separability: không xét vì không có thuộc tính nào sử dụng 2 channel để biểu diễn.

#### Phân tích biểu đồ

* + - * + Thể hiện rõ số lượng diễn viên tham gia theo từng thể loại phim.
        + International movies có tổng số lượng diễn viên cao nhất trong khi đó TV shows có số lượng ít nhất.
  1. **Domain task 8**
     1. **Idiom**

|  |  |  |
| --- | --- | --- |
| **Idiom** | **Line charts** | |
| **Data** | Rating: C | |
| Release Year: Q | |
| **Encode** | Mark: line | |
| Channel | C: pos dọc |
| C: màu sắc |
| Q: pos ngang |
| **TASK** | Phân phối và xu hướng phim có phân loại độ tuổi được phát hành theo năm. | |
| **SCALE** | Keys: 18  Level: Nominal | |

* + 1. **Biểu đồ**



* + 1. **Đánh giá**

#### Tính biểu đạt

* + - * + Biểu đồ cột thể hiển rõ số lượng phim có phân loại độ tuổi phát hành theo năm.
        + Tiêu đề phản ảnh đúng nội dung của biểu đồ, cho người xem biết mục đích của hình ảnh trực quan.
        + Nhãn “rating” giúp người dùng dễ phân biệt phân loại độ tuổi nào.

#### Tính hiệu quả

* + - * + Accuracy: sử dụng channel vị trí (trục Y) giúp phân biệt phân loại độ tuổi.
        + Discriminablity: Sử dụng màu sắc để phân biệt phân loại độ tuổi.
        + Separability: không xét vì không có thuộc tính nào sử dụng 2 channel để biểu diễn.

#### Phân tích biểu đồ

* + - * + Thể hiện rõ xu hướng phim có phân loại độ tuổi phát hành theo từng năm.
        + TV-MA có xu hướng tăng và giảm nhẹ tuy nhiên dẫn đầu qua từng năm phát hành. Trong khi đó, các độ tuổi với tiêu đề “66 min”, “74 min”, “84 min” chỉ có phát hành một năm duy nhất và là số lượng phim có phân loại độ tuổi thấp nhất.