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.

1. Dataprofiling

1.1 Tiền xử lí dữ liệu

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

```
show_id,type,title,director,cast,country,date_added,release_year,rating,duration,listed_in,description
s1,Movie,Dick Johnson Is Dead,Kirsten Johnson,,United States,"September 25, 2021",2020,PG-13,90 min,Documentaries,"As her father n
s2,TV Show,Blood & Water,,"Ama Qamata, Khosi Ngema, Gail Mabalane, Thabang Molaba, Dillon Windvogel, Natasha Thahane, Arno Greeff,
s3,TV Show,Ganglands,Julien Leclercq,"Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabiha Akkari, Sofia Lesaffre, Salim Kechiouche, N
s4,TV Show,Jailbirds New Orleans,,,,"September 24, 2021",2021,TV-MA,1 Season,"Docuseries, Reality TV","Feuds, flirtations and toil
s5,TV Show,Kota Factory,,"Mayur More, Jitendra Kumar, Ranjan Raj, Alam Khan, Ahsaas Channa, Revathi Pillai, Urvi Singh, Arun Kumar
s6,TV Show,Midnight Mass,Mike Flanagan,"Kate Siegel, Zach Gilford, Hamish Linklater, Henry Thomas, Kristin Lehman, Samantha Sloyan
s7,Movie,My Little Pony: A New Generation,"Robert Cullen, José Luis Ucha","Vanessa Hudgens, Kimiko Glenn, James Marsden, Sofia Car
s8,Movie,Sankofa,Haile Gerima,"Kofi Ghanaba, Oyafunmike Ogunlano, Alexandra Duah, Nick Medley, Mutabaruka, Afemo Omilami, Reggie C
s9,TV Show,The Great British Baking Show,Andy Devonshire,"Mel Giedroyc, Sue Perkins, Mary Berry, Paul Hollywood",United Kingdom,"S
s10,Movie,The Starling,Theodore Melfi,"Melissa McCarthy, Chris O'Dowd, Kevin Kline, Timothy Olyphant, Daveed Diggs, Skyler Gisondo
s11,TV Show,"Vendetta: Truth, Lies and The Mafia",,,,"September 24, 2021",2021,TV-MA,1 Season,"Crime TV Shows, Docuseries, Interna
s12,TV Show,Bangkok Breaking,Kongkiat Komesiri,"Sukollawat Kanarot, Sushar Manaying, Pavarit Mongkolpisit, Sahajak Boonthanakit, S
s13,Movie,Je Suis Karl,Christian Schwochow,"Luna Wedler, Jannis Niewöhner, Milan Peschel, Edin Hasanović, Anna Fialová, Marlon Boe
s14,Movie,Confessions of an Invisible Girl,Bruno Garotti,"Klara Castanho, Lucca Picon, Júlia Gomes, Marcus Bessa, Kiria Malheiros,
s15,TV Show,Crime Stories: India Detectives,,,,"September 22, 2021",2021,TV-MA,1 Season,"British TV Shows, Crime TV Shows, Docuser
s16,TV Show,Dear White People,,"Logan Browning, Brandon P. Bell, DeRon Horton, Antoinette Robertson, John Patrick Amedori, Ashley
s17,Movie,Europe's Most Dangerous Man: Otto Skorzeny in Spain,"Pedro de Echave García, Pablo Azorín Williams",,,"September 22, 202
s18,TV Show,Falsa identidad,,"Luis Ernesto Franco, Camila Sodi, Sergio Goyri, Samadhi Zendejas, Eduardo Yáñez, Sonya Smith, Alejan
s19,Movie,Intrusion,Adam Salky,"Freida Pinto, Logan Marshall-Green, Robert John Burke, Megan Elisabeth Kelly, Sarah Minnich, Hayes
s20,TV Show,Jaguar,,"Blanca Suárez, Iván Marcos, Óscar Casas, Adrián Lastra, Francesc Garrido, Stefan Weinert, Julia Möller, Alici
s21,TV Show,Monsters Inside: The 24 Faces of Billy Milligan,Olivier Megaton,,,"September 22, 2021",2021,TV-14,1 Season,"Crime TV S
s22,TV Show,Resurrection: Ertugrul,,"Engin Altan Düzyatan, Serdar Gökhan, Hülya Darcan, Kaan Taşaner, Esra Bilgiç, Osman Soykut, S
s23,Movie,Avvai Shanmughi,K.S. Ravikumar,"Kamal Hassan, Meena, Gemini Ganesan, Heera Rajgopal, Nassar, S.P. Balasubrahmanyam",,"Se
s24,Movie,Go! Go! Cory Carson: Chrissy Takes the Wheel,"Alex Woo, Stanley Moore","Maisie Benson, Paul Killam, Kerry Gudjohnsen, AC
s25,Movie,Jeans,S. Shankar,"Prashanth, Aishwarya Rai Bachchan, Sri Lakshmi, Nassar",India,"September 21, 2021",1998,TV-14,166 min,
s26,TV Show,Love on the Spectrum,,Brooke Satchwell,Australia,"September 21, 2021",2021,TV-14,2 Seasons,"Docuseries, International
s27,Movie,Minsara Kanavu,Rajiv Menon,"Arvind Swamy, Kajol, Prabhu Deva, Nassar, S.P. Balasubrahmanyam, Girish Karnad",,"September
s28,Movie,Grown Ups,Dennis Dugan,"Adam Sandler, Kevin James, Chris Rock, David Spade, Rob Schneider, Salma Hayek, Maria Bello, May
s29,Movie,Dark Skies,Scott Stewart,"Keri Russell, Josh Hamilton, J.K. Simmons, Dakota Goyo, Kadan Rockett, L.J. Benet, Rich Hutchm
s30,Movie,Paranoia,Robert Luketic,"Liam Hemsworth, Gary Oldman, Amber Heard, Harrison Ford, Lucas Till, Embeth Davidtz, Julian McM
s31,Movie,Ankahi Kahaniya,"Ashwiny Iyer Tiwari, Abhishek Chaubey, Saket Chaudhary","Abhishek Banerjee, Rinku Rajguru, Delzad Hiwal
s32,TV Show,Chicago Party Aunt,,"Lauren Ash, Rory O'Malley, RuPaul Charles, Jill Talley, Ike Barinholtz, Jon Barinholtz, Matthew C
s33,TV Show,Sex Education,,"Asa Butterfield, Gillian Anderson, Ncuti Gatwa, Emma Mackey, Connor Swindells, Kedar Williams-Stirling
s34,TV Show,Squid Game,,"Lee Jung-jae, Park Hae-soo, Wi Ha-jun, Oh Young-soo, Jung Ho-yeon, Heo Sung-tae, Kim Joo-ryoung, Tripathi
s35,TV Show,Tayo and Little Wizards,,"Dami Lee, Jason Lee, Bommie Catherine Han, Jennifer Waescher, Nancy Kim",,"September 17, 202
s36,Movie,The Father Who Moves Mountains,Daniel Sandu,"Adrian Titieni, Elena Purea, Judith State, Valeriu Andriuță, Tudor Smoleanu
s37,Movie,The Stronghold,Cédric Jimenez,"Gilles Lellouche, Karim Leklou, François Civil, Adèle Exarchopoulos, Kenza Fortas, Cyril
s38,TV Show,Angry Birds,,"Antti Pääkkönen, Heljä Heikkinen, Lynne Guaglione, Pasi Ruohonen, Rauno Ahonen",Finland,"September 16, 2
```

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:

```
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   show_id       8807 non-null   object
 1   type          8807 non-null   object
 2   title         8807 non-null   object
 3   director      6173 non-null   object
 4   cast          7982 non-null   object
 5   country       7976 non-null   object
 6   date_added    8797 non-null   object
 7   release_year  8807 non-null   int64
 8   rating        8803 non-null   object
 9   duration      8804 non-null   object
 10  listed_in     8807 non-null   object
 11  description   8807 non-null   object
dtypes: int64(1), object(11)
```

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”:

```
show_id,type,title,director,cast,country,date_added,release_year,rating,duration,listed_in,description
s1,Movie,Dick Johnson Is Dead,Kirsten Johnson,Không có thông tin,United States,"September 25, 2021",2020,PG-13,90 min,Documentaries,"As her father nears the end of his life, fil
s2,TV Show,Blood & Water,Không có thông tin,"Ama Qamata, Khosi Ngema, Gail Mabalane, Thabang Molaba, Dillon Windvogel, Natasha Thahane, Arno Greeff, Xolile Tshabalala, Getmore S
s3,TV Show,Ganglands,Julien Leclercq,"Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabiha Akkari, Sofia Lesaffre, Salim Kechiouche, Noureddine Farihi, Geert Van Rampelberg, Bakary
s4,TV Show,Jailbirds New Orleans,Không có thông tin,Không có thông tin,Không xác định,"September 24, 2021",2021,TV-MA,1 Season,"Docuseries, Reality TV","Feuds, flirtations and t
s5,TV Show,Kota Factory,Không có thông tin,"Mayur More, Jitendra Kumar, Ranjan Raj, Alam Khan, Ahsaas Channa, Revathi Pillai, Urvi Singh, Arun Kumar",India,"September 24, 2021",
s6,TV Show,Midnight Mass,Mike Flanagan,"Kate Siegel, Zach Gilford, Hamish Linklater, Henry Thomas, Kristin Lehman, Samantha Sloyan, Igby Rigney, Rahul Kohli, Annarah Cymone, Ann
s7,Movie,My Little Pony: A New Generation,"Robert Cullen, José Luis Ucha","Vanessa Hudgens, Kimiko Glenn, James Marsden, Sofia Carson, Liza Koshy, Ken Jeong, Elizabeth Perkins,
s8,Movie,Sankofa,Haile Gerima,"Kofi Ghanaba, Oyafunmike Ogunlano, Alexandra Duah, Nick Medley, Mutabaruka, Afemo Omilami, Reggie Carter, Mzuri","United States, Ghana, Burkina Fa
s9,TV Show,The Great British Baking Show,Andy Devonshire,"Mel Giedroyc, Sue Perkins, Mary Berry, Paul Hollywood",United Kingdom,"September 24, 2021",2021,TV-14,9 Seasons,"Britis
s10,Movie,The Starling,Theodore Melfi,"Melissa McCarthy, Chris O'Dowd, Kevin Kline, Timothy Olyphant, Daveed Diggs, Skyler Gisondo, Laura Harrier, Rosalind Chao, Kimberly Quinn,
s11,TV Show,"Vendetta: Truth, Lies and The Mafia",Không có thông tin,Không có thông tin,Không xác định,"September 24, 2021",2021,TV-MA,1 Season,"Crime TV Shows, Docuseries, Inte
s12,TV Show,Bangkok Breaking,Kongkiat Komesiri,"Sukollawat Kanarot, Sushar Manaying, Pavarit Mongkolpisit, Sahajak Boonthanakit, Suthipongse Thatphithakkul, Bhasaworn Bawronkira
s13,Movie,Je Suis Karl,Christian Schwochow,"Luna Wedler, Jannis Niewöhner, Milan Peschel, Edin Hasanović, Anna Fialová, Marlon Boess, Victor Boccard, Fleur Geffrier, Aziz Dyab,
s14,Movie,Confessions of an Invisible Girl,Bruno Garotti,"Klara Castanho, Lucca Picon, Júlia Gomes, Marcus Bessa, Kiria Malheiros, Fernanda Concon, Gabriel Lima, Caio Cabral, Le
s15,TV Show,Crime Stories: India Detectives,Không có thông tin,Không có thông tin,Không xác định,"September 22, 2021",2021,TV-MA,1 Season,"British TV Shows, Crime TV Shows, Docu
s16,TV Show,Dear White People,Không có thông tin,"Logan Browning, Brandon P. Bell, DeRon Horton, Antoinette Robertson, John Patrick Amedori, Ashley Blaine Featherson, Marque Ric
s17,Movie,Europe's Most Dangerous Man: Otto Skorzeny in Spain,"Pedro de Echave García, Pablo Azorín Williams",Không có thông tin,Không xác định,"September 22, 2021",2020,TV-MA,6
s18,TV Show,Falsa identidad,Không có thông tin,"Luis Ernesto Franco, Camila Sodi, Sergio Goyri, Samadhi Zendejas, Eduardo Yáñez, Sonya Smith, Alejandro Camacho, Azela Robinson,
s19,Movie,Intrusion,Adam Salky,"Freida Pinto, Logan Marshall-Green, Robert John Burke, Megan Elisabeth Kelly, Sarah Minnich, Hayes Hargrove, Mark Sivertsen, Brandon Fierro, Anto
s20,TV Show,Jaguar,Không có thông tin,"Blanca Suárez, Iván Marcos, Óscar Casas, Adrián Lastra, Francesc Garrido, Stefan Weinert, Julia Möller, Alicia Chojnowski",Không xác định,
s21,TV Show,Monsters Inside: The 24 Faces of Billy Milligan,Olivier Megaton,Không có thông tin,Không xác định,"September 22, 2021",2021,TV-14,1 Season,"Crime TV Shows, Docuserie
s22,TV Show,Resurrection: Ertugrul,Không có thông tin,"Engin Altan Düzyatan, Serdar Gökhan, Hülya Darcan, Kaan Taşaner, Esra Bilgiç, Osman Soykut, Serdar Deniz, Cengiz Coşkun, F
s23,Movie,Avvai Shanmughi,K.S. Ravikumar,"Kamal Hassan, Meena, Gemini Ganesan, Heera Rajgopal, Nassar, S.P. Balasubrahmanyam",Không xác định,"September 21, 2021",1996,TV-PG,161
s24,Movie,Go! Go! Cory Carson: Chrissy Takes the Wheel,"Alex Woo, Stanley Moore","Maisie Benson, Paul Killam, Kerry Gudjohnsen, AC Lim",Không xác định,"September 21, 2021",2021,
s25,Movie,Jeans,S. Shankar,"Prashanth, Aishwarya Rai Bachchan, Sri Lakshmi, Nassar",India,"September 21, 2021",1998,TV-14,166 min,"Comedies, International Movies, Romantic Movie
s26,TV Show,Love on the Spectrum,Không có thông tin,Brooke Satchwell,Australia,"September 21, 2021",2021,TV-14,2 Seasons,"Docuseries, International TV Shows, Reality TV","Findir
s27,Movie,Minsara Kanavu,Rajiv Menon,"Arvind Swamy, Kajol, Prabhu Deva, Nassar, S.P. Balasubrahmanyam, Girish Karnad",Không xác định,"September 21, 2021",1997,TV-PG,147 min,"Com
s28,Movie,Grown Ups,Dennis Dugan,"Adam Sandler, Kevin James, Chris Rock, David Spade, Rob Schneider, Salma Hayek, Maria Bello, Maya Rudolph, Colin Quinn, Tim Meadows, Joyce Van
s29,Movie,Dark Skies,Scott Stewart,"Keri Russell, Josh Hamilton, J.K. Simmons, Dakota Goyo, Kadan Rockett, L.J. Benet, Rich Hutchman, Myndy Crist, Annie Thurman, Jake Brennan",L
s30,Movie,Paranoia,Robert Luketic,"Liam Hemsworth, Gary Oldman, Amber Heard, Harrison Ford, Lucas Till, Embeth Davidtz, Julian McMahon, Josh Holloway, Richard Dreyfuss, Angela S
s31,Movie,Ankahi Kahaniya,"Ashwiny Iyer Tiwari, Abhishek Chaubey, Saket Chaudhary","Abhishek Banerjee, Rinku Rajguru, Delzad Hiwale, Kunal Kapoor, Zoya Hussain, Nikhil Dwivedi,
s32,TV Show,Chicago Party Aunt,Không có thông tin,"Lauren Ash, Rory O'Malley, RuPaul Charles, Jill Talley, Ike Barinholtz, Jon Barinholtz, Matthew Craig, Bob Odenkirk, Mike Hage
s33,TV Show,Sex Education,Không có thông tin,"Asa Butterfield, Gillian Anderson, Ncuti Gatwa, Emma Mackey, Connor Swindells, Kedar Williams-Stirling, Alistair Petrie",United Kir
s34,TV Show,Squid Game,Không có thông tin,"Lee Jung-jae, Park Hae-soo, Wi Ha-jun, Oh Young-soo, Jung Ho-yeon, Heo Sung-tae, Kim Joo-ryoung, Tripathi Anupam, You Seong-joo, Lee Y
s35,TV Show,Tayo and Little Wizards,Không có thông tin,"Dami Lee, Jason Lee, Bommie Catherine Han, Jennifer Waescher, Nancy Kim",Không xác định,"September 17, 2021",2020,TV-Y7,1
s36,Movie,The Father Who Moves Mountains,Daniel Sandu,"Adrian Titieni, Elena Purea, Judith State, Valeriu Andriuță, Tudor Smoleanu, Virgil Aioanei, Radu Botar, Petronela Grigore
s37,Movie,The Stronghold,Cédric Jimenez,"Gilles Lellouche, Karim Leklou, François Civil, Adèle Exarchopoulos, Kenza Fortas, Cyril Lecomte, Michaël Abiteboul, Idir Azougli, Vince
s38,TV Show,Angry Birds,Không có thông tin,"Antti Pääkkönen, Heljä Heikkinen, Lynne Guaglione, Pasi Ruohonen, Rauno Ahonen",Finland,"September 16, 2021",2018,TV-Y7,1 Season,"Kid
s39,Movie,Birth of the Dragon,George Nolfi,"Billy Magnussen, Ron Yuan, Qu Jingjing, Terry Chen, Vanness Wu, Jin Xing, Philip Ng, Xia Yu, Yu Xia","China, Canada, United States","
s40,TV Show,Chhota Bheem,Không có thông tin,"Vatsal Dubey, Julie Tejwani, Rupa Bhimani, Jigna Bhardwaj, Rajesh Kava, Mousam, Swapnil",India,"September 16, 2021",2021,TV-Y7,3 Sea
s41,TV Show,He-Man and the Masters of the Universe,Không có thông tin,"Yuri Lowenthal, Kimberly Brooks, Antony Del Rio, Trevor Devall, Ben Diskin, Grey Griffin, David Kaye, Tom
s42,Movie,Jaws,Steven Spielberg,"Roy Scheider, Robert Shaw, Richard Dreyfuss, Lorraine Gary, Murray Hamilton, Carl Gottlieb, Jeffrey Kramer, Susan Backlinie, Jonathan Filley, Te
s43,Movie,Jaws 2,Jeannot Szwarc,"Roy Scheider, Lorraine Gary, Murray Hamilton, Joseph Mascolo, Jeffrey Kramer, Collin Wilcox Paxton, Ann Dusenberry, Mark Gruner, Barry Coe, Susa
s44,Movie,Jaws 3,Joe Alves,"Dennis Quaid, Bess Armstrong, Simon MacCorkindale, Louis Gossett Jr., John Putch, Lea Thompson, P.H. Moriarty, Dan Blasko, Liz Morris, Lisa Maurer",L
s45,Movie,Jaws: The Revenge,Joseph Sargent,"Lorraine Gary, Lance Guest, Mario Van Peebles, Karen Young, Michael Caine, Judith Barsi, Mitchell Anderson, Lynn Whitfield",United St
s46,Movie,My Heroes Were Cowboys,Tyler Greco,Không có thông tin,Không xác định,"September 16, 2021",2021,PG,23 min,Documentaries,"Robin Wiltshire's painful childhood was rescuec
s47,Movie,Safe House,Daniel Espinosa,"Denzel Washington, Ryan Reynolds, Vera Farmiga, Brendan Gleeson, Sam Shepard, Rubén Blades, Nora Arnezeder, Robert Patrick, Liam Cunningham
s48,TV Show,The Smart Money Woman,Bunmi Ajakaiye,"Osas Ighodaro, Ini Dima-Okojie, Kemi Lala Akindoju, Toni Tones, Ebenezer Eno, Eso Okolocha DIke, Patrick Diabuah, Karibi Fubara
s49,Movie,Training Day,Antoine Fuqua,"Denzel Washington, Ethan Hawke, Scott Glenn, Tom Berenger, Harris Yulin, Raymond J. Barry, Cliff Curtis, Dr. Dre, Snoop Dogg, Macy Gray, Ev
```

File netflix_titles_cleaned.csv

1.2 Profiling toàn bộ dữ liệu

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

		Field Name	NULL	Missing	Actual	Completeness (%)	Cardinality	Uniqueness (%)	Distinctness (%)
0	0	show_id	0	0	8807	100.00	8807	100.00	100.00
1	1	type	0	0	8807	100.00	2	0.02	0.02
2	2	title	0	0	8807	100.00	8806	99.99	99.99
3	3	director	2634	0	6173	70.09	4528	51.41	51.41
4	4	cast	825	0	7982	90.63	7692	87.34	87.34
5	5	country	831	0	7976	90.56	748	8.49	8.49
6	6	date_added	10	0	8797	99.89	1767	20.06	20.06
7	7	release_year	0	0	8807	100.00	74	0.84	0.84
8	8	rating	4	0	8803	99.95	17	0.19	0.19
9	9	duration	3	0	8804	99.97	220	2.50	2.50
10	10	listed_in	0	0	8807	100.00	514	5.84	5.84
11	11	description	0	0	8807	100.00	8775	99.64	99.64

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).

1.3 Profiling từng thuộc tính

Show_id:

Metadata	
Attribute	Value
Field Name	show_id
Field Datatype	object
Field Length	1

Data Profiling Summary	
Attribute	Value
NULL	0
Missing	0
Actual	8807
Completeness (%)	100
Cardinality	8807
Uniqueness (%)	100
Distinctness (%)	100

Data Profiling Additional Statistics	
Attribute	Value
Data Types	1
Field Length (MIN)	2
Field Length (MAX)	5
Field Value (MAX)	N/A
Field Formats	4

show_id - Field Formats		
Format	Count	Percentage(%)
XXXXX	7808	88.6568
XXXX	900	10.2191
XXX	90	1.0219
XX	9	0.1022

show_id - Field Data Types		
Data Type	Count	Percentage(%)
object	8807	100

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:

Metadata	
Attribute	Value
Field Name	type
Field Datatype	object
Field Length	1

Data Profiling Summary	
Attribute	Value
NULL	0
Missing	0
Actual	8807
Completeness (%)	100
Cardinality	2
Uniqueness (%)	0.0227
Distinctness (%)	0.0227

Data Profiling Additional Statistics	
Attribute	Value
Data Types	1
Field Length (MIN)	5
Field Length (MAX)	7
Field Value (MAX)	N/A
Field Formats	2

type - Popular		
Value	Count	Percentage(%)
Movie	6131	69.6151
TV Show	2676	30.3849

type - Field Formats		
Format	Count	Percentage(%)
XXXXX	6131	69.6151
XXXXXXX	2676	30.3849

type - Field Data Types		
Data Type	Count	Percentage(%)
object	8807	100

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:

Metadata	
Attribute	Value
Field Name	title
Field Datatype	object
Field Length	3
Data Profiling Summary	
Attribute	Value
NULL	0
Missing	0
Actual	8807
Completeness (%)	100
Cardinality	8806
Uniqueness (%)	99.9886
Distinctness (%)	99.9886
Data Profiling Additional Statistics	
Attribute	Value
Data Types	1
Field Length (MIN)	1
Field Length (MAX)	104
Field Value (MAX)	N/A
Field Formats	77

title - Field Formats		
Format	Count	Percentage(%)
XXXXXXXXXXX	462	5.2458
XXXXXXXXX	460	5.2231
XXXXXXXXXXXX	452	5.1323
XXXXXXXXXX	450	5.1096
XXXXXXXXXXXXX	436	4.9506
XXXXXXX	416	4.7235
XXXXXXXXXXXXX	415	4.7122
XXXXXXXXX	410	4.6554
XXXXXXXXXXXXXX	374	4.2466
XXXXXXXXXXXXXX	369	4.1898
XXXXXXX	316	3.5881
XXXXXX	311	3.5313
XXXXXXXXXXXXXXXX	283	3.2134
XXXXXXXXXXXXXXXX	276	3.1339
XXXXXXXXXXXXXXXXX	256	2.9068
XXXXX	234	2.657
XXXXXXXXXXXXXXXXXXXX	221	2.5094
XXXXXXXXXXXXXXXXX	212	2.4072
XXXXXXXXXXXXXXXXXXX	176	1.9984
XXXXXXXXXXXXXXXXXXXXXXXX	173	1.9643
XXXXXXXXXXXXXXXXXXXX	161	1.8281
XXXXXXXXXXXXXXXXXXXX	161	1.8281
XXXX	157	1.7827
XXXXXXXXXXXXXXXXXXXXXXXX	146	1.6578
XXXXXXXXXXXXXXXXXXXXXX	132	1.4988
XXXXXXXXXXXXXXXXXXXXXXXXX	113	1.2831
XXXXXXXXXXXXXXXXXXXXXXXXXX	110	1.249
XXXXXXXXXXXXXXXXXXXXXX	108	1.2263
XXXXXXXXXXXXXXXXXXXXXXXXXXX	92	1.0446
XXXXXXXXXXXXXXXXXXXXXXXXXX	87	0.9879
XXXXXXXXXXXXXXXXXXXXXXXXX	81	0.9197
XXXXXXXXXXXXXXXXXXXXXXXXXX	76	0.8629
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	58	0.6586

title - Field Data Types		
Data Type	Count	Percentage(%)
object	8807	100
title - Top 5 Most Popular		
Value	Count	Percentage(%)
Consequences	2	0.0227
Ricardo O'Farrill Abrazo Genial	1	0.0114
Cooked	1	0.0114
Care Bears & Cousins	1	0.0114
Hannibal Buress: Comedy Camisado	1	0.0114
title - Top 5 Least Popular		
Value	Count	Percentage(%)
ROAD TO ROMA	1	0.0114
Good Time	1	0.0114
Captain Underpants Epic Choice-o-Rama	1	0.0114
We Bare Bears	1	0.0114
Zubaan	1	0.0114

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:

Metadata	
Attribute	Value
Field Name	director
Field Datatype	object
Field Length	3
Data Profiling Summary	
Attribute	Value
NULL	0
Missing	0
Actual	8807
Completeness (%)	100
Cardinality	4529
Uniqueness (%)	51.425
Distinctness (%)	51.425
Data Profiling Additional Statistics	
Attribute	Value
Data Types	1
Field Length (MIN)	2
Field Length (MAX)	208
Field Value (MAX)	N/A
Field Formats	80

director - Field Data Types		
Data Type	Count	Percentage(%)
object	8807	100
director - Top 5 Most Popular		
Value	Count	Percentage(%)
Không có thông tin	2634	29.908
Rajiv Chilaka	19	0.2157
Raúl Campos, Jan Suter	18	0.2044
Suhas Kadav	16	0.1817
Marcus Raboy	16	0.1817
director - Top 5 Least Popular		
Value	Count	Percentage(%)
Raymie Muzquiz, Stu Livingston	1	0.0114
Joe Menendez	1	0.0114
Eric Bross	1	0.0114
Will Eisenberg	1	0.0114
Mozez Singh	1	0.0114

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

Metadata	
Attribute	Value
Field Name	cast
Field Datatype	object
Field Length	3

ata Profiling Summary	
Attribute	Value
NULL	0
Missing	0
Actual	8807
Completeness (%)	100
Cardinality	7693
Uniqueness (%)	87.351
Distinctness (%)	87.351

ofiling Additional Statistics	
Attribute	Value
Data Types	1
Field Length (MIN)	3
Field Length (MAX)	771
Field Value (MAX)	N/A
Field Formats	386

cast - Field Data Types		
Data Type	Count	Percentage(%)
object	8807	100

cast - Top 5 Most Popular		
Value	Count	Percentage(%)
Không có thông tin	825	9.3675
David Attenborough	19	0.2157
Vatsal Dubey, Julie Tejwani, Rupa Bhimani, Jigna Bhardwaj, Rajesh Kava, Mousam, Swapnil	14	0.159
Samuel West	10	0.1135
Jeff Dunham	7	0.0795

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

Metadata	
Attribute	Value
Field Name	country
Field Datatype	object
Field Length	3

Data Profiling Summary	
Attribute	Value
NULL	0
Missing	0
Actual	8807
Completeness (%)	100
Cardinality	749
Uniqueness (%)	8.5046
Distinctness (%)	8.5046

Data Profiling Additional Statistics	
Attribute	Value
Data Types	1
Field Length (MIN)	4
Field Length (MAX)	123
Field Value (MAX)	N/A
Field Formats	72

country - Field Formats		
Format	Count	Percentage(%)
XXXXXXXXXXXXX	2862	32.4969
XXXXX	1599	18.156
XXXXXXXXXXXXXX	1276	14.4885
XXXXXX	763	8.6636
XXXXXXXXXX	310	3.5199
XXXXXXXXX	301	3.4177
XXXXXXX	248	2.8159
XXXXXXXXXXXXXXXXXXXX	228	2.5888
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	157	1.7827
XXXXXXXX	140	1.5896
XXXXXXXXXXXXXXXXXX	102	1.1582
XXXXXXXXXXXXXXXXXXXX	78	0.8857
XXXXXXXXXXXXXXXXXXXXXX	55	0.6245
XXXXXXXXXXXXXX	54	0.6131
XXXXXXXXXXXXXX	53	0.6018
XXXXXXXXXXXXXXXXXX	49	0.5564
XXXXXXXXXXX	47	0.5337
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	42	0.4769
XXXXXXXXXXXXXXXXXXXXXXXXXXXX	34	0.3861
XXXXXXXXXXXXXXXXXXXXXXXXXXXX	32	0.3633
XXXXXXXXXXXXXXXXXXXXXXXXX	28	0.3179
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	28	0.3179
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	28	0.3179
XXXXXXXXXXXXXXXXXXXXXXXXXXXX	27	0.3066
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	17	0.193
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	15	0.1703
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	15	0.1703
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	15	0.1703
XXXXXXXXXXXXXXXXX	14	0.159
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	14	0.159
XXXXXXXXXXXXXXXXXXXXXXXXXX	13	0.1476
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	12	0.1363
XXXXXXXXXXXXXXXX	12	0.1363
XXXXXXXXXXXXXX	11	0.1249
XXXXXXXXXXXXXXXXXXXXXX	11	0.1249
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	9	0.1022
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	8	0.0908
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	8	0.0908
XXXXXXXXXXXXXXXXXXXXXXXXXXXX	7	0.0795
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	6	0.0681

country - Field Data Types		
Data Type	Count	Percentage(%)
object	8807	100

country - Top 5 Most Popular		
Value	Count	Percentage(%)
United States	2818	31.9973
India	972	11.0367
Không xác định	831	9.4357
United Kingdom	419	4.7576
Japan	245	2.7819

country - Top 5 Least Popular		
Value	Count	Percentage(%)
Romania, Bulgaria, Hungary	1	0.0114
Uruguay, Guatemala	1	0.0114
France, Senegal, Belgium	1	0.0114
Mexico, United States, Spain, Colombia	1	0.0114
United Arab Emirates, Jordan	1	0.0114

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

Metadata	
Attribute	Value
Field Name	date_added
Field Datatype	object
Field Length	2

Data Profiling Summary	
Attribute	Value
NULL	0
Missing	0
Actual	8807
Completeness (%)	100
Cardinality	1768
Uniqueness (%)	20.0749
Distinctness (%)	20.0749

Data Profiling Additional Statistics	
Attribute	Value
Data Types	1
Field Length (MIN)	11
Field Length (MAX)	19
Field Value (MAX)	N/A
Field Formats	9

date_added - Top 5 Most Popular		
Value	Count	Percentage(%)
January 1, 2020	109	1.2377
November 1, 2019	89	1.0106
March 1, 2018	75	0.8516
December 31, 2019	74	0.8402
October 1, 2018	71	0.8062

date_added - Top 5 Least Popular		
Value	Count	Percentage(%)
December 4, 2016	1	0.0114
November 21, 2016	1	0.0114
November 19, 2016	1	0.0114
November 17, 2016	1	0.0114
January 11, 2020	1	0.0114

date_added - Field Formats		
Format	Count	Percentage(%)
XXXXXXXXXXXXXXXX	1589	18.0425
XXXXXXXXXXXXXXX	1567	17.7927
XXXXXXXXXXXX	1462	16.6004
XXXXXXXXXXXXX	1258	14.2841
XXXXXXXXXXXXXX	1178	13.3757
XXXXXXXXXXX	1040	11.8088
XXXXXXXXXXXXXXXXX	439	4.9847
XXXXXXXXXX	272	3.0885
XXXXXXXXXXXXXXXXX	2	0.0227

date_added - Field Data Types		
Data Type	Count	Percentage(%)
object	8807	100

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

Metadata	
Attribute	Value
Field Name	release_year
Field Datatype	int64
Field Length	1

Data Profiling Summary	
Attribute	Value
NULL	0
Missing	0
Actual	8807
Completeness (%)	100
Cardinality	74
Uniqueness (%)	0.8402
Distinctness (%)	0.8402

Data Profiling Additional Statistics	
Attribute	Value
Data Types	1
Field Length (MIN)	4
Field Length (MAX)	4
Field Value (MAX)	2021
Field Formats	1

release_year - Top 5 Most Popular		
Value	Count	Percentage(%)
2018	1147	13.0237
2017	1032	11.718
2019	1030	11.6952
2020	953	10.8209
2016	902	10.2419

release_year - Top 5 Least Popular		
Value	Count	Percentage(%)
1959	1	0.0114
1925	1	0.0114
1961	1	0.0114
1947	1	0.0114
1966	1	0.0114

release_year - Field Formats		
Format	Count	Percentage(%)
XXXX	8807	100

release_year - Field Data Types		
Data Type	Count	Percentage(%)
int64	8807	100

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:

Metadata		
Attribute	Value	
Field Name	rating	
Field Datatype	object	
Field Length	2	
Data Profiling Summary		
Attribute	Value	
NULL	0	
Missing	0	
Actual	8807	
Completeness (%)	100	
Cardinality	18	
Uniqueness (%)	0.2044	
Distinctness (%)	0.2044	
Data Profiling Additional Statistics		
Attribute	Value	
Data Types	1	
Field Length (MIN)	1	
Field Length (MAX)	14	
Field Value (MAX)	N/A	
Field Formats	7	
rating - Top 5 Most Popular		
Value	Count	Percentage(%)
TV-MA	3207	36.4142
TV-14	2160	24.5259
TV-PG	863	9.799
R	799	9.0723
PG-13	490	5.5638

rating - Top 5 Least Popular		
Value	Count	Percentage(%)
NC-17	3	0.0341
UR	3	0.0341
74 min	1	0.0114
84 min	1	0.0114
66 min	1	0.0114
rating - Field Formats		
Format	Count	Percentage(%)
XXXXX	7057	80.1294
X	840	9.5379
XXXX	527	5.9839
XX	370	4.2012
XXXXXXXX	6	0.0681
XXXXXXXXXXXXX	4	0.0454
XXXXXX	3	0.0341
rating - Field Data Types		
Data Type	Count	Percentage(%)
object	8807	100

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

Metadata	
Attribute	Value
Field Name	duration
Field Datatype	object
Field Length	2

Data Profiling Summary	
Attribute	Value
NULL	0
Missing	0
Actual	8807
Completeness (%)	100
Cardinality	221
Uniqueness (%)	2.5094
Distinctness (%)	2.5094

Data Profiling Additional Statistics	
Attribute	Value
Data Types	1
Field Length (MIN)	5
Field Length (MAX)	14
Field Value (MAX)	N/A
Field Formats	7

duration - Top 5 Most Popular		
Value	Count	Percentage(%)
1 Season	1793	20.3588
2 Seasons	425	4.8257
3 Seasons	199	2.2596
90 min	152	1.7259
94 min	146	1.6578

duration - Top 5 Least Popular		
Value	Count	Percentage(%)
189 min	1	0.0114
10 min	1	0.0114
3 min	1	0.0114
229 min	1	0.0114
191 min	1	0.0114

duration - Field Formats		
Format	Count	Percentage(%)
XXXXXX	3202	36.3574
XXXXXXX	2922	33.1782
XXXXXXXX	1793	20.3588
XXXXXXXXX	866	9.8331
XXXXXXXXXX	17	0.193
XXXXX	4	0.0454
XXXXXXXXXXXXX	3	0.0341

duration - Field Data Types		
Data Type	Count	Percentage(%)
object	8807	100

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

Metadata	
Attribute	Value
Field Name	description
Field Datatype	object
Field Length	3
Data Profiling Summary	
Attribute	Value
NULL	0
Missing	0
Actual	8807
Completeness (%)	100
Cardinality	8775
Uniqueness (%)	99.6367
Distinctness (%)	99.6367
Data Profiling Additional Statistics	
Attribute	Value
Data Types	1
Field Length (MIN)	61
Field Length (MAX)	248
Field Value (MAX)	N/A
Field Formats	104

description - Field Data Types		
Data Type	Count	Percentage(%)
object	8807	100

description - Top 5 Most Popular		
Value	Count	Percentage(%)
Paranormal activity at a lush, abandoned property alarms a group eager to redevelop the site, but the eerie events may not be as unearthly as they think.	4	0.0454
Challenged to compose 100 songs before he can marry the girl he loves, a tortured but passionate singer-songwriter embarks on a poignant musical journey.	3	0.0341
A surly septuagenarian gets another chance at her 20s after having her photo snapped at a studio that magically takes 50 years off her life.	3	0.0341
Multiple women report their husbands as missing but when it appears they are looking for the same man, a police officer traces their cryptic connection.	3	0.0341
Secrets bubble to the surface after a sensual encounter and an unforeseen crime entangle two friends and a woman caught between them.	2	0.0227

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.

2. Abstraction

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

2.1.1. Data abastraction

Data:	netflix_titles_cleaned						
Item:	1 dòng là 1 nội dung (Movie hoặc TV Show)						
Dataset availability:	Static						
	Semantics	Attribute Type	Hierarchical	Characteristic	Direction	Quantitative Type	Bin Number
Release Year	Năm phát hành	Quantitative	None	Discrete	Sequential	Interval	13 (2008–2020)
Type	Loại nội dung	Categorical	None	Discrete	None	None	2 (Movie, TV)

2.1.2. 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.

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

2.2.1. Data abstraction

Data:	netflix_titles_cleaned
Item:	1 dòng là 1 nội dung (Movie hoặc TV Show)
Dataset availability:	Static

Attribute	Semantics	Attribute Type	Hierarchical	Characteristic	Direction	Quantitative Type	Bin Number
Country	Quốc gia sản xuất	Categorical	None	Discrete	None	None	~100+
Show ID	ID nội dung	Identifier	None	Discrete	None	Count-based	–

2.2.2. 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.

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

2.3.1. Data abstraction

Attr	Abstraction
Data Type	Attribute and Item
Dataset type	Table
Dataset availability	Static

No	Attribute Name	Type(C, O, Q)	Interval/ Ratio	Key/ Value	Direction	Hierarchical	Continuouns/ Discrete	Semantic
1	Date Added	Quantitative	Interval	Value	Sequential	Yes	Discrete	Yes. Đây là ngày nội dung được thêm vào Netflix, dùng để phân tích xu hướng thể loại theo quý. "Yes, nội dung được thêm vào quý nào?"
2	Listed In	Categorical	None	Key	None	No	Discrete	Yes. Đây là danh sách thể loại của nội dung, dùng để xác định xu hướng thể loại phổ biến. "Yes, thể loại nào đang thịnh hành?"

2.3.2. 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

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

2.4.1. Data abstraction

Attr	Abstraction
Data Type	Attribute and Item
Dataset type	Table
Dataset availability	Static

No	Attribute Name	Type(C, O, Q)	Interval/ Ratio	Key/ Value	Direction	Hierarchical	Continuouns/ Discrete	Semantic
1	Rating	Categorical	None	Key	None	Yes	Discrete	Yes. Đây là phân loại độ tuổi của nội dung, dùng để phân tích mối quan hệ với thể loại nội dung. "Yes, phân loại độ tuổi này phổ biến với thể loại nào?"
2	Listed In	Categorical	None	Key	None	No	Discrete	Yes. Đây là danh sách thể loại của nội dung, dùng để xác định mối quan hệ với phân loại độ tuổi. "Yes, thể loại này thường có phân loại độ tuổi nào?"

2.4.2. 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**

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

2.5.1. Data abstraction

Attr	Abstraction
Data type	Attribute and Item
Dataset type	Table
Dataset availability	Static

No	Attribute Name	Type(C, O, Q)	Interval/Ratio	Key/Value	Direction	Hierarchical	Continuous/Discrete	Semantic
1	Rating	Categorical	None	Value	Sequential	Yes	Discrete	Yes. Đại cho độ tuổi khán giả mục tiêu và mức độ hạn chế nội dung
2	Number of records	Quantitative	None	Key	None	No	Discrete	No

2.5.2. 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.

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

2.6.1. Data abstraction

4	Longitude (generated)	Quantitative	Interval	Value	Sequential	No	Continuous	Yes. Kinh độ địa lý quốc gia

2.6.2. 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.

2.7. Domain task 7: Phân phối diễn viên theo thể loại

2.7.1. Data abstraction

Attr	Abstraction
Data type	Attribute and Item
Dataset type	Table
Dataset availability	Static

No	Attribute Name	Type(C,O,Q)	Interval/Ratio	Key/Value	Direction	Hierarchial	Continuos/Discrete	Semantic
1	Cast	Quantitative	Ratio	Value	None	No	Discrete	Yes. Số lượng diễn viên
2	Listed in	Categorical	None	Key	None	Yes(Genres)	Discrete	Yes. Danh sách thể loại

2.7.2. 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

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

2.8.1. Data abstraction

Attr	Abstraction
Data type	Attribute and Item
Dataset type	Table
Dataset availability	Static

No	Attribute Name	Type(C,O,Q)	Interval/Ratio	Key/Value	Direction	Hierarchial	Continuos/Discrete	Semantic
1	Rating	Categorical	None	Key	None	Yes(Rating)	Discrete	Yes. Phân loại độ tuổi
2	Release year	Quantitative	Interval	Key	Increasing	Yes(Years)	Discrete	Yes. Năm phát hành

2.8.2. 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 → Summa

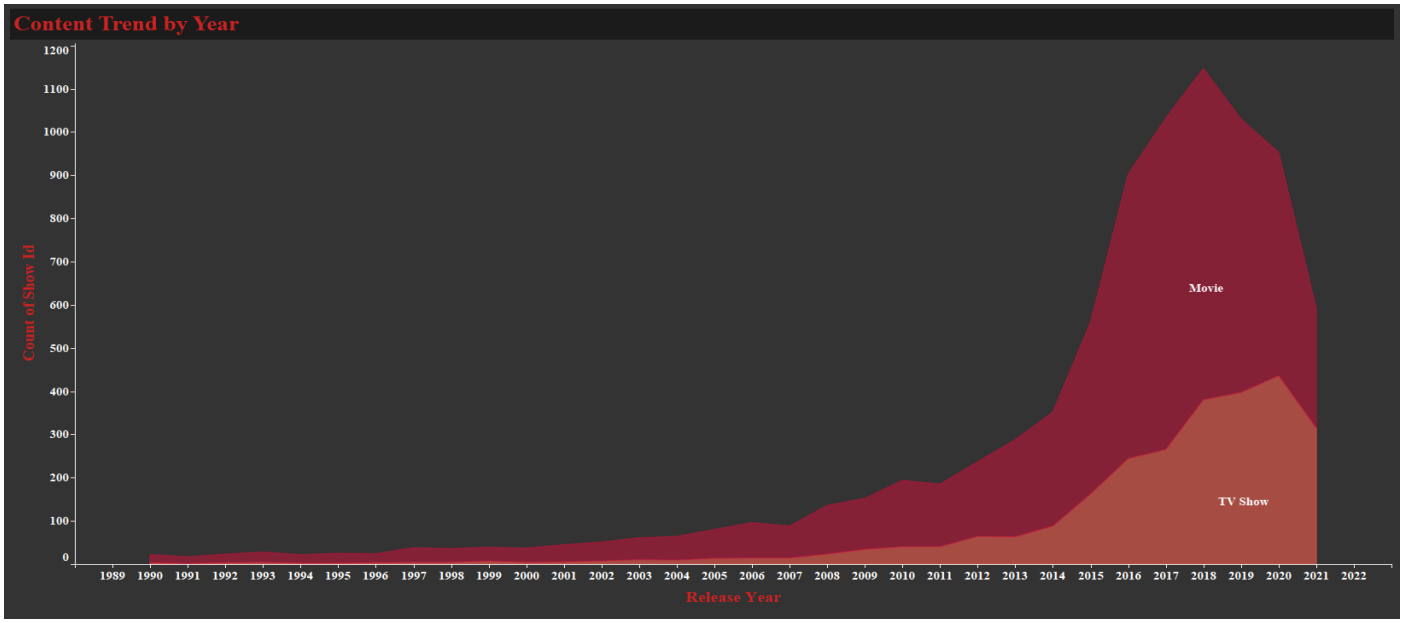
3. Thiết kế idiom

3.1.Domain task 1

3.1.1. 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)	
Encode	Mark: area	
	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*	

3.1.2. Biểu đồ



3.1.3. 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.

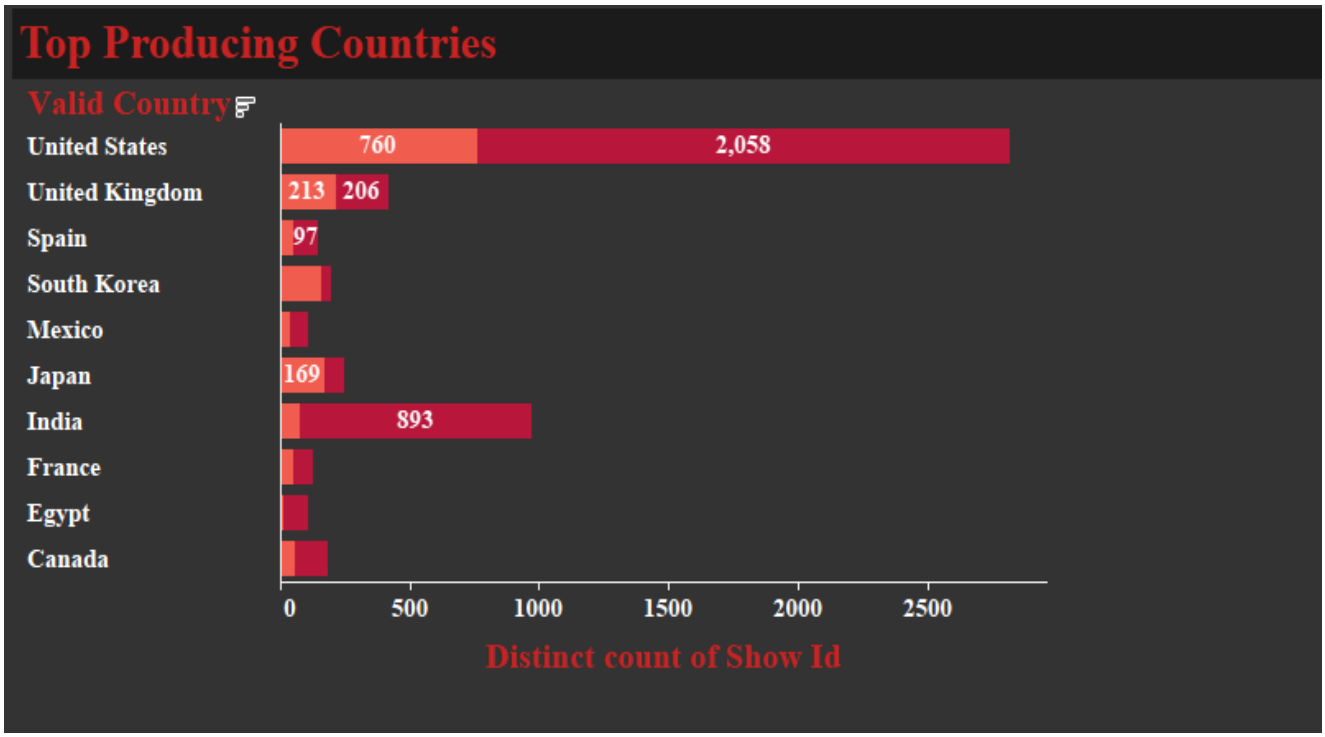
- **Tính hiệu quả**
  - Accuracy: Trục Y biểu diễn số lượng giúp đọc giá trị chính xác.
  - Discriminablity: Màu sắc phân biệt rõ Movie và TV Show.
  - Separability: Hai loại nội dung hiển thị cùng biểu đồ giúp dễ so sánh tỷ trọng theo từng năm.
- **Phân tích biểu đồ**
  - Giúp xác định độ tuổi có nguy cơ mắc bệnh tim cao nhất và xu hướng theo độ tuổi.

3.2.Domain task 2

3.2.1. 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)	
Encode	Mark: bar	
	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)	

3.2.2. Biểu đồ



3.2.3. Đánh giá

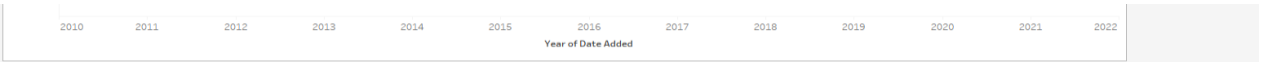
- **Tính biểu đạt**
  - Biểu tròn giúp thể hiện rõ tỉ lệ từng nhóm.
  - Dễ nhận diện mức độ chênh lệch giữa nam và nữ.
- **Tính hiệu quả**
  - Accuracy: giúp dễ dàng hình dung tỷ lệ giữa các nhóm.
  - Discriminablity: sử dụng màu sắc giúp phân biệt giữa có/không có bệnh tim.
  - Separability: không xét 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.

3.3.Domain task 3

3.3.1. 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)	

3.3.2. Biểu đồ



3.3.3. 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.

3.4.Domain task 4

3.4.1. 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)	

3.4.2. Biểu đồ

3.4.3. 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.

3.5.Domain task 5

3.5.1. 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 giả Xu hướng phân phối theo định mức độ tuổi	
SCALE	Keys: 14 Level: Categorical	

3.5.2. Biểu đồ

3.5.3. 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 biến 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.

3.6.Domain task 6

3.6.1. 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

3.6.2. Biểu đồ

3.6.3. 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).

3.7.Domain task 7

3.7.1. 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	

3.7.2. Biểu đồ

3.7.3. 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.

3.8.Domain task 8

3.8.1. 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

3.8.2. Biểu đồ

3.8.3. 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.