Relatório de Sistemas de Informação Analíticos
Sistema Integrado de Data Warehousing para Análise de Filmes e Séries Mais Vistos na Netflix

2023/2024

Unidade Curricular: Sistemas de Informação Analíticos
Docente: Professor Carlos Miguel Francisco

Grupo:
Karina Kozyar nº 120160
Leonor Santana nº 120178
Matilde Alves nº 120172

Índice

Objetivo:

Desenvolver e implementar um conjunto de dados robusto que consolide informações sobre filmes e programas de TV disponíveis na Netflix. Este sistema permitirá análises detalhadas sobre o catálogo da Netflix, incluindo categorias, países onde está disponível, classificações e outros dados relevantes, facilitando a compreensão das preferências dos espectadores e o potencial de sucesso de diferentes tipos de conteúdo.

Descrição da empresa e do negócio:

A Netflix é uma empresa inovadora especializada em análise e consultoria na indústria do entretenimento em streaming, dedicada a ajudar produtores e distribuidores a compreender e otimizar o desempenho das suas séries através de análises de dados avançadas. Fundada num momento em que o streaming de conteúdo se tornou uma forma dominante de consumo de entretenimento, a Netflix posiciona-se na interseção entre tecnologia, análise de dados e criação de conteúdo.

A missão da Netflix é capacitar produtores e distribuidores com dados e insights precisos para promover decisões informadas na criação e distribuição de séries. Através da análise de grandes volumes de dados de visualização, feedback dos espectadores e tendências de mercado, a empresa visa facilitar a produção de conteúdo de alta qualidade que seja relevante para o público-alvo e impulsione o sucesso comercial.

Os serviços oferecidos incluem análises de audiência avançadas, consultoria de conteúdo, previsão de popularidade de séries e análises comparativas de desempenho.

A Netflix está comprometida com a inovação tecnológica. A empresa desenvolveu uma plataforma de armazenamento de dados proprietária que integra dados de diversas fontes, incluindo análises de audiência em tempo real, dados demográficos dos espectadores e informações de produção das séries. Esta plataforma permite análises complexas e a geração de modelos preditivos que oferecem uma visão clara das tendências de visualização e preferências do público.

Os clientes da Netflix variam desde pequenos estúdios independentes até grandes produtoras e distribuidoras de conteúdo, todas tentando alcançar o máximo de sucesso com as suas séries na plataforma da Netflix e outras plataformas de streaming.

A Netflix destaca-se como um parceiro essencial para empresas que procuram não apenas criar conteúdo de qualidade, mas também entender e atender às necessidades do público-alvo, impulsionando assim o sucesso comercial e artístico das suas séries.

Etapas:

1.Recolha de Dados:

O conjunto de dados "Filmes e Séries da Netflix" disponível no Kaggle foi compilado por Diego Enrique, um utilizador da plataforma. Este conjunto de dados contém informações sobre filmes e programas de TV disponíveis na Netflix, incluindo detalhes como título, tipo (filme ou programa de TV), realizador, elenco, país de produção, data de lançamento na Netflix, classificação, duração, entre outros em dois datasets designados de “titles” e “credits”.

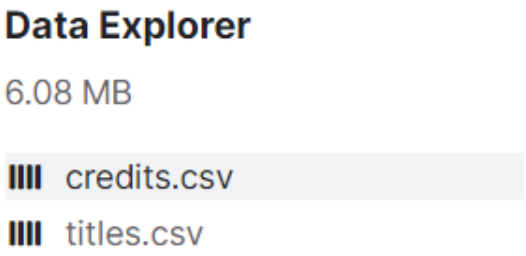

Figura 1 - Datasets escolhidos

Quanto à fonte dos dados específicos, é importante salientar que o Kaggle é uma plataforma que permite aos utilizadores partilharem conjuntos de dados de diversas origens. Assim, o autor do conjunto de dados provavelmente reuniu informações de várias fontes, como APIs públicas, sites da Netflix, bases de dados de filmes e séries, entre outros.

Para obter análises de visualização em tempo real, feedback dos espectadores em redes sociais e informações de produção das séries disponíveis na Netflix, pode ser necessário consultar diversas fontes de dados, como APIs de redes sociais (por exemplo, Twitter, Facebook, Reddit), sites de análises de mídia (por exemplo, IMDb, Rotten Tomatoes), bancos de dados de produção cinematográfica, entre outros.

Em resumo, reunir esses dados relevantes para análises de desempenho e preferências do público pode envolver uma abordagem multifacetada que inclui a consulta a várias fontes de dados e a integração de diferentes conjuntos de informações.

2. Modelação de Dados:

Desenvolver um modelo de dados dimensional que facilite análises rápidas e eficientes, com foco nos indicadores de popularidade, preferências de conteúdo e fatores de sucesso das séries. Este modelo permitirá uma compreensão mais profunda do comportamento dos espectadores e das suas tendências de streaming na Netflix.

3. Descrição dos dados

Dim Titles

Atributo	Descrição	Tipo de Atributo
ID	ID do título no JustWatch.	Varchar(200)
Title	O nome do título.	Varchar(200)
Show type	Série de TV ou filme.	Varchar(200)
Season	Número de temporadas se for uma série.	Interger(10)
Description	Uma breve descrição.	Varchar(200)
Release Year	O ano de lançamento.	
Age Certification	A certificação de idade.	Varchar(200)
Run Time	A duração do episódio (série) ou filme.	Interger(10)
Production Countries	Uma lista de países que produziram o título.	Varchar(200)
ID_Genres	O ID do género	Interger(10)
ID_Country	O ID do país de gravação	Interger(10)

Figura 2 - Entidade Dim_Titles

Dim_Genres

Atributo	Descrição	Tipo de Atributo
Genres	Género do filme/série	Varchar(200)
Id_genres	O ID do género	Interger(10)

Figura 3 - Entidade Dim_Genres

Dim_Person

Atributo	Descrição	Tipo de Atributo
Person_id	O ID da pessoa no JustWacth	Interger(10)
Name	Nome do ator ou diretor	Varchar(200)
Character_name	Nome da personagem	Varchar(200)
Role	Ator ou Diretor	Varchar(200)

Figura 4 - Entidade Dim_Person

Dim_Country

Atributo	Descrição	Tipo de Atributo
Country	País de Gravação	Varchar(200)
Id_country	O ID do país	Interger(10)

Figura 5 - Entidade Dim_country

Fact_IMDB

Atributo	Descrição	Tipo de Atributo
IMDB_ID	O ID do filme/serie no imbd	Varchar(200)
IMDB_votes	O ID do voto	Interger(10)
IMBD_Score	O ID da avaliação	Float (10)

Figura 6 - Entidade Fact_IMDB

Dim_Time_imdb

Atributo	Descrição	Tipo de Atributo
Date_time	Dia e horas em que foi feita a avaliação	Date

Figura 7 - Entidade Dim_Time_imdb

Começámos por desenvolver um modelo entidade-relacionamento para começar a planear a nossa base de dados e modelar os dados de forma estruturada de modo a garantir a integridade e consistência dos dados.

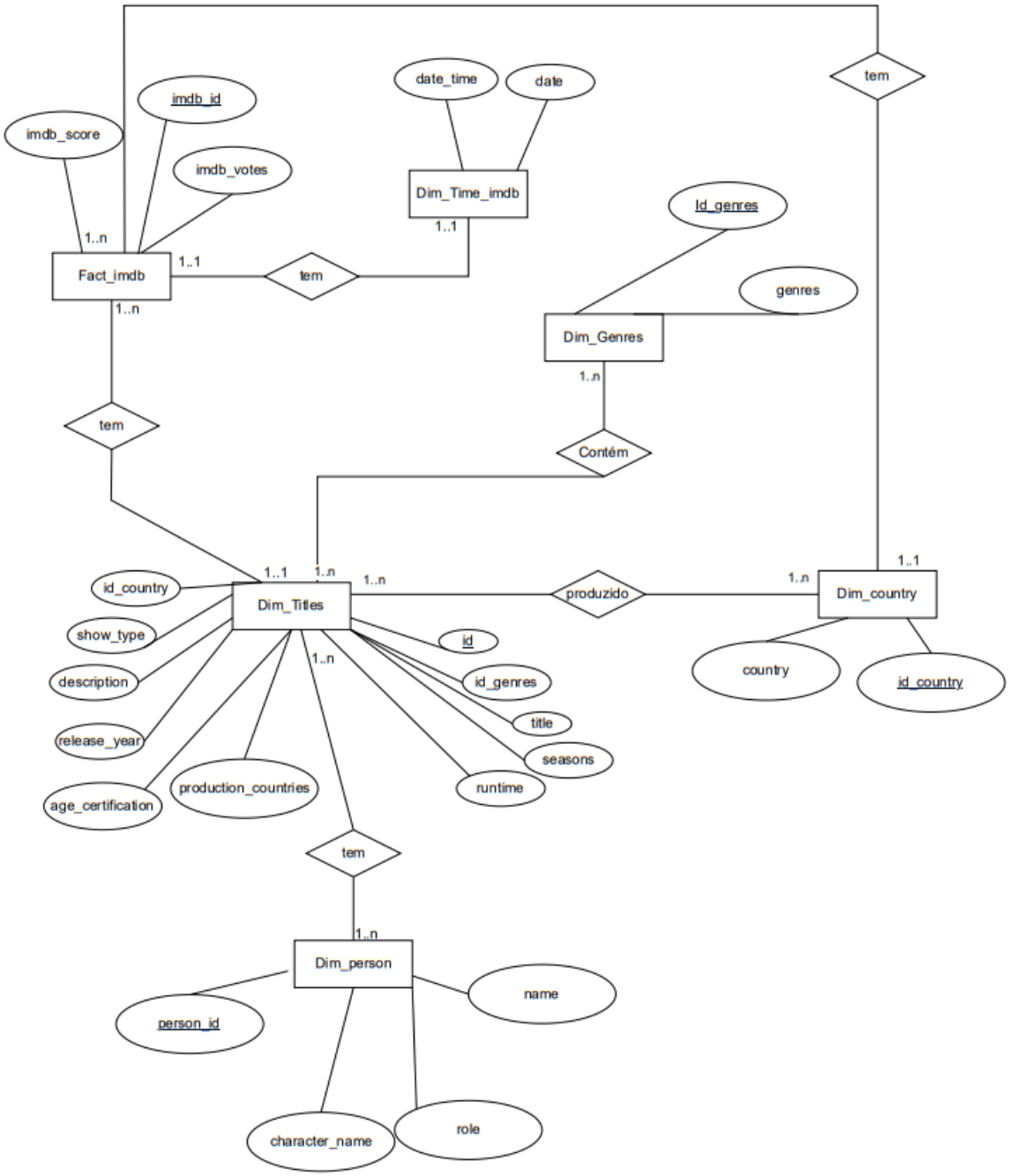

Figura 8 - Diagrama Entidade-Relacionamento de Séries Netflix

Posteriormente, foi criado um esquema dimensional, no formato de floco de neve, para garantir a normalização dos dados, mantendo-os organizados e estruturados. Dando assim um maior suporte a consultas complexas e consequentemente uma maior facilidade nas atualizações

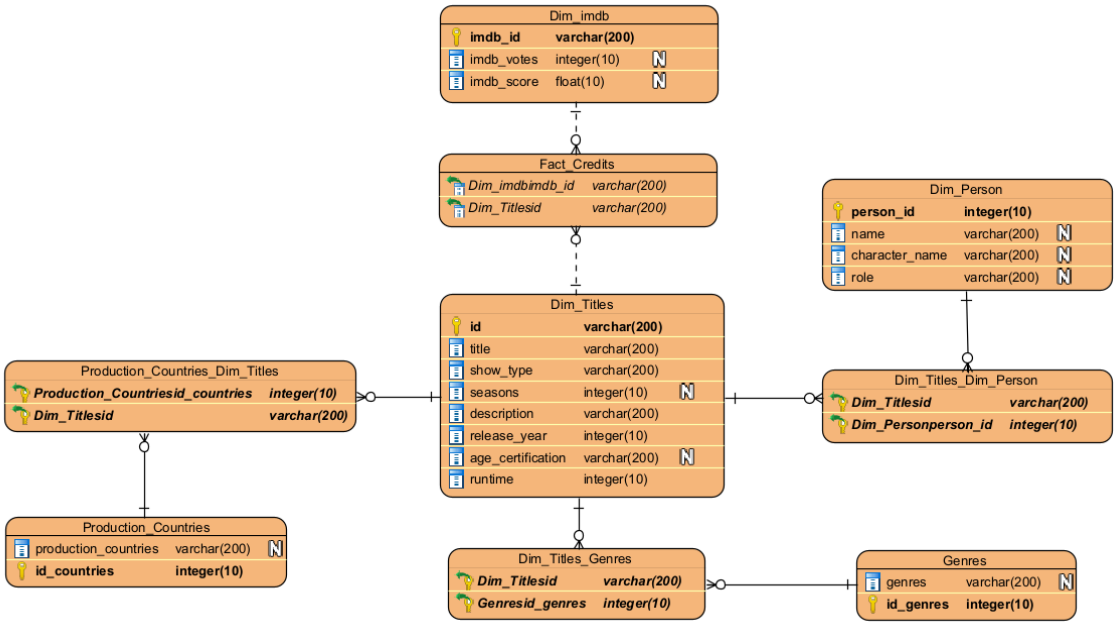

Figura 9 - Modelo dimesnional

A tabela de factos contém duas chaves estrangeiras, sendo estas a chave primária id da tabela dimensão IMDb e a chave primária da tabela dimensão Títulos. A tabela dimensão IMDb, além do código identificador, existem dados da quantidade de votos em determinado filme ou série e a pontuação final realizada. A tabela dimensão Títulos contém dados como: título, tipo (filme/série), número de temporadas (que pode não existir caso seja filme), descrição, data de emissão, certificado de idade recomendada e duração em minutos. As tabelas dimensão dos países produtores, géneros e pessoas, apresentam uma relação de muitos para muitos com a tabela referente aos títulos, assim, foram criadas terceiras tabelas para cada uma destas relações.

4. Limpeza e Preparação de Dados

A limpeza e preparação de dados foi realizada em Python (.py) e Spoon (.ktr).

Data cleaning:

Python (.py)
Começámos por criar um número identificador para cada género de filme/série de forma a associar um ou mais géneros a um filme/série. A coluna “genres” em “titles.csv” estava distribuída da seguinte forma:
genres: [‘documentation´]; [‘drama’, ‘sport’]; [‘romance’, ‘comedy’]; [‘crime’, ‘drama’, ‘comedy’, ‘music’], [‘fantasy’, ‘comedy’]; …

genres
['documentation']
['drama', 'sport']
['romance', 'comedy']
['crime', 'drama', 'comedy', 'music']
['drama', 'sport']
['fantasy', 'comedy']

Figura 10 - Coluna "genres" em "titles.csv"

Após a extração de todos os géneros únicos desta coluna, criou-se um dicionário com “genre_ids” e “genre”, convertemos o dicionário para um dataframe e de seguida, para um ficheiro excel posteriormente convertido para csv. Agora ficamos com um novo ficheiro csv “genres”:
genre_id = genre

0 = action; 1= animation; 2 = comedy; 3 = crime; 4 = documentation; 5 = drama; ...

genre_id	genre
0	action
1	animation
2	comedy
3	crime
4	documentation
5	drama

Figura 11 - Colunas "genre_id" e "genre" em "genres.csv"

Sabendo que um filme/série podem ter um ou mais géneros (relação de muitos para muitos), procedeu-se à criação de um terceiro csv onde cada linha apresenta uma chave do filme/séria e uma chave correspondente a um género:
refTitle - refGenre
O título com o id “ts300399” apenas tem um género, que tem o id 4, enquanto que, por exemplo, o título com o id “tm82169” apresenta dois géneros de id 5 e 15, tal como demonstrado na figura seguinte:

refTitle	refGenre
ts300399	*4*
tm82169	*5*
tm82169	*15*
tm17823	*13*
tm17823	*2*
tm191099	*3*
tm191099	*5*
tm191099	*2*
tm191099	*11*
tm69975	*5*

Figura 12 - Colunas "refTitle" e "refGenre" em "title_genre.csv"

Os mesmos passos procederam-se para o atributo “production_countries”:

1. coluna original em “titles.csv”

production_countries
['US']
['US']
['US']
['US']
['US']
['GB']

2. dicionário de países

id_country	country
0	AE
1	AF
2	AL
3	AO
4	AR
5	AT
6	AU

3. id - id_country = refTitle - refCountry

refTitle	refCountry
ts300399	100
tm82169	100
tm17823	100
tm191099	100
tm69975	100
tm127384	35

Figura 13, 14 e 15 – Procedimento de tartamentod e dados “production_countries”

O dataset “titles.csv” não p\ossuía inicialmente as variáveis “id_country” e “date” associadas às variáveis imdb, por isso, com o auxílio do python, foi possível criar estas duas colunas de forma aleatória. “id_country” refere-se ao país do qual foram feitas as avaliações e “date”, à data (aaaa-mm-dd) que foram feitas as avaliações. “id_country” foi criada aleatoriamente a partir do dicionário criado para os diferentes países e “date”, a partir de números aleatórios para cada campo do seu tipo (aaaa, 2000-2024; mm, 1-12; dd, 1-28).

id_country	date
95	25/09/2014
25	03/08/2015
27	17/07/2001
23	13/06/2004
22	20/02/2009
40	24/08/2009
37	28/02/2014
72	08/12/2020

Figura 16 - Colunas "id_country" e "date" em "titles.csv".

Spoon (.ktr)

Usámos:

Funções	Tabelas/passos	Utilidade
Filter rows	“Dim_country”; “Dim_genre”; “Dim_title_country”; “Dim_title_genre”; “Dim_titles_person”; “Fact_imdb”	Ajuda a eliminar as linhas com id nulos de forma a apenas passarem linhas com chaves não nulas para as tabelas SQL
Sort rows	“Dim_titles”; “Dim_person”; “Dim_title_person”	Passo necessário para a utilização de “Unique rows”, ordena de forma ascendente uma variável à escolha (por exemplo chaves), fácil leitura
Unique rows	“Dim_titles”; “Dim_person”; “Dim_title_person”	Após ordenadas as chaves, aquelas que estiverem duplicadas são eliminadas para apenas passarem linhas com chaves únicas.

Tabela 1 - Funções, tabelas e utilidade em Spoon

Data loading:

Todos os dados foram introduzidos utilizando as ferramentas “CSV input” e “Table Output”/”Insert/Update”.

Este passo está dividido em três fases:

Tranformation 1	Nesta fase todas as tabelas SQL que não possuem chaves estrangeiras recebem dados dos devidos ficheiros CSV: “Dim_country”, “Dim_genre”, “Dim_titles” e “Dim_person”. Esta fase foi criada para ser realizada antes da segunda fase.
Transformation 2	Esta fase possui todas as tabelas SQL cujas contém chaves estrangeiras, esta fase apenas ocorre após a “Tranformation 1” ocorrer, pois as chaves estrangeiras provêm das tabelas da transformação anterior: “Dim_title_country”, “Dim_title_genre”, “Dim_title_person”, “Fact_imdb”.
Job	Esta fase foi criada para correr as transformações de forma ordenada.

Tabela 2 - Transformação/Job criados em Spoon

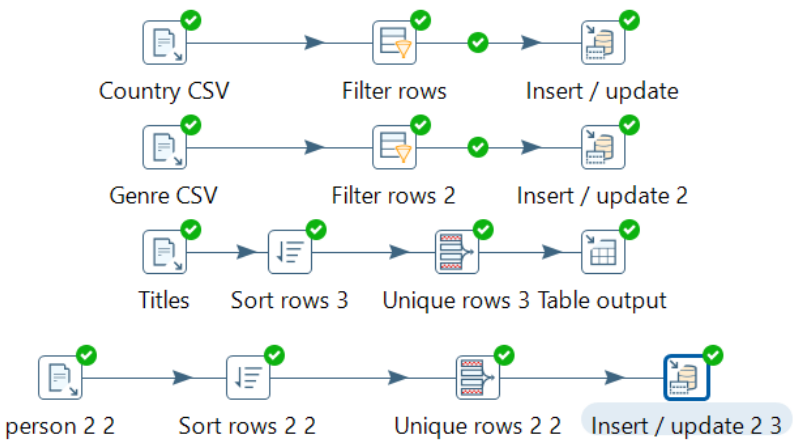

Figura 17 - Tranformation 1

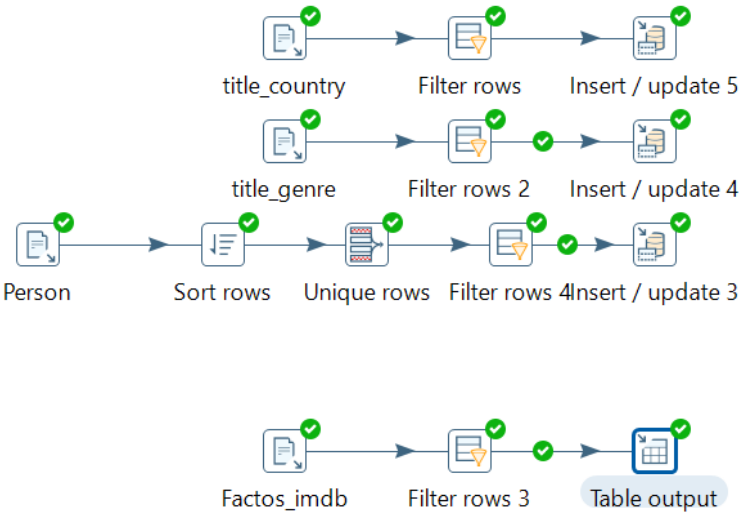

Figura 18 - Tranformation 2

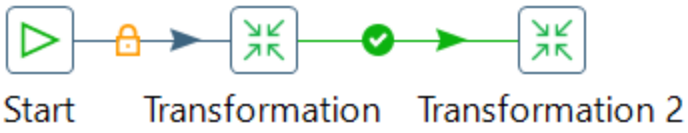

Figura 19 - Job

5. Construção do Data Warehouse:

Para a construção do Data Warehouse criámos as devidas tabelas em mySQL Workbench, otimizamos consultas e criámos views para complementar a análise realizada em Power BI.

Otimização de consultas:

```
CREATE INDEX idx_imdb_score ON Fact_imdb(imdb_score);
CREATE INDEX idx_release_year ON Dim_titles(release_year);
CREATE INDEX idx_runtime ON Dim_titles(runtime);
CREATE INDEX idx_country ON Dim_Country(country);
CREATE INDEX idx_age_certification ON Dim_titles(age_certification);
CREATE INDEX idx_type ON Dim_titles(type);
CREATE INDEX idx_seasons ON Dim_titles(seasons);
```

E criámos as seguintes views:

Média de IMDb Score por país

country	avg_imdb_score
AE	6.178571428571429
AF	6.672131147540983
AL	6.425925925925926
AO	6.529411764705882
AR	6.327586206896552
AT	6.181818181818182
AU	6.467741935483871
BD	6.5636363636363635
BE	6.375

Figura 20 - View: média de IMDb score por país

Total de votos por título

	title	total_votes
▶	The Blazing Sun	1219
	White Christmas	46586
	Dark Waters	703
	Cairo Station	4878
	Professor	312
	Saladin the Victorious	2670
	Amrapali	251
	Monty Python's Flying Circus	75654
	The Land	2732

Figura 21 – View: total de votos por título

Títulos com maior IMDb score

	title	imdb_score
▶	Crazy Delicious	10
	Khawatir	10
	#ABtalks	10
	Breaking Bad	10
	Rubaru Roshni	9
	The Chosen	9
	Kota Factory	9
	Demon Slayer: Kimetsu no Yaiba	9
	Our Planet	9

Figura 22 - View: títulos com maior IMDb score

IMDb scores por ano de lançamento

release_year	avg_imdb_score	min_imdb_score	max_imdb_score
1954	7.5	7	8
1956	7	7	7
1958	8	8	8
1962	7	7	7
1963	8	8	8
1966	7	7	7
1969	8	7	9
1970	8	8	8
1971	6.333333333333333	6	7

Figura 23 - View: IMDb score por ano de lançamento

Título por género com IMDb score

genre	num_titles	avg_imdb_score
crime	945	6.652822151224707
drama	2999	6.633310946944258
romance	1045	6.438159156279962
comedy	2399	6.437762825904121
music	254	6.596837944664031
thriller	1195	6.372582001682086
action	1113	6.443037974683544
history	275	7.141818181818182
war	151	7.152317880794702

Figura 24 - View: Título por género com IMDb score

IMDb scores por certificado de idade

age_certification	avg_imdb_score
NULL	6.260162601626016
TV-14	7.206971677559912
R	6.28952772073922
PG	6.308333333333334
TV-Y	6.653061224489796
TV-MA	7.001028806584362
PG-13	6.481651376146789
TV-PG	7.010416666666667
TV-G	6.487804878048781

Figura 25 - View: IMDb scores por certificado de idade

6. Análise e BI

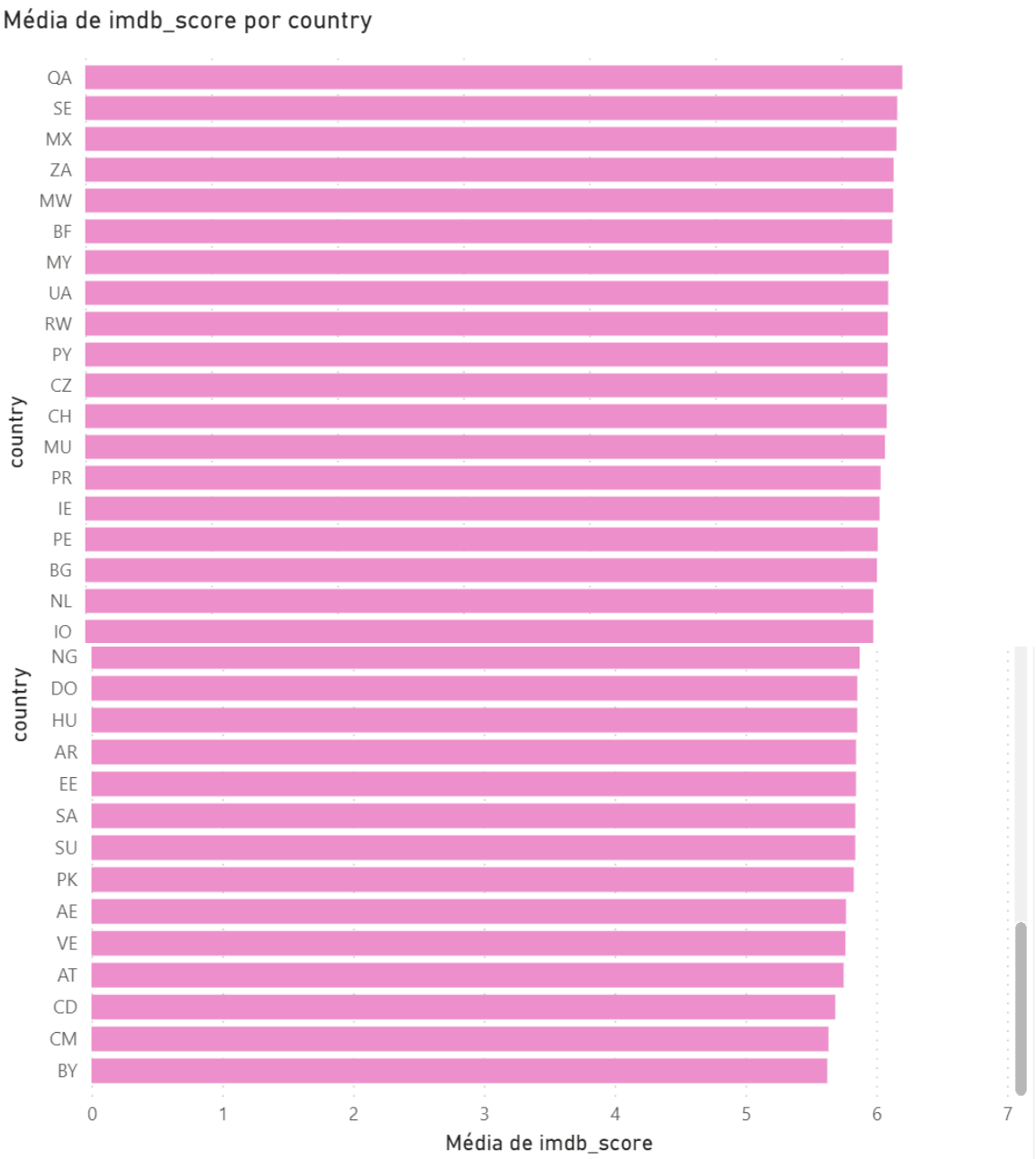

Figura 26 - Média de score do imdb por país

Com este gráfico conseguimos observar que a média de score no imdb é superior no Qatar enquanto é inferior no Burundi

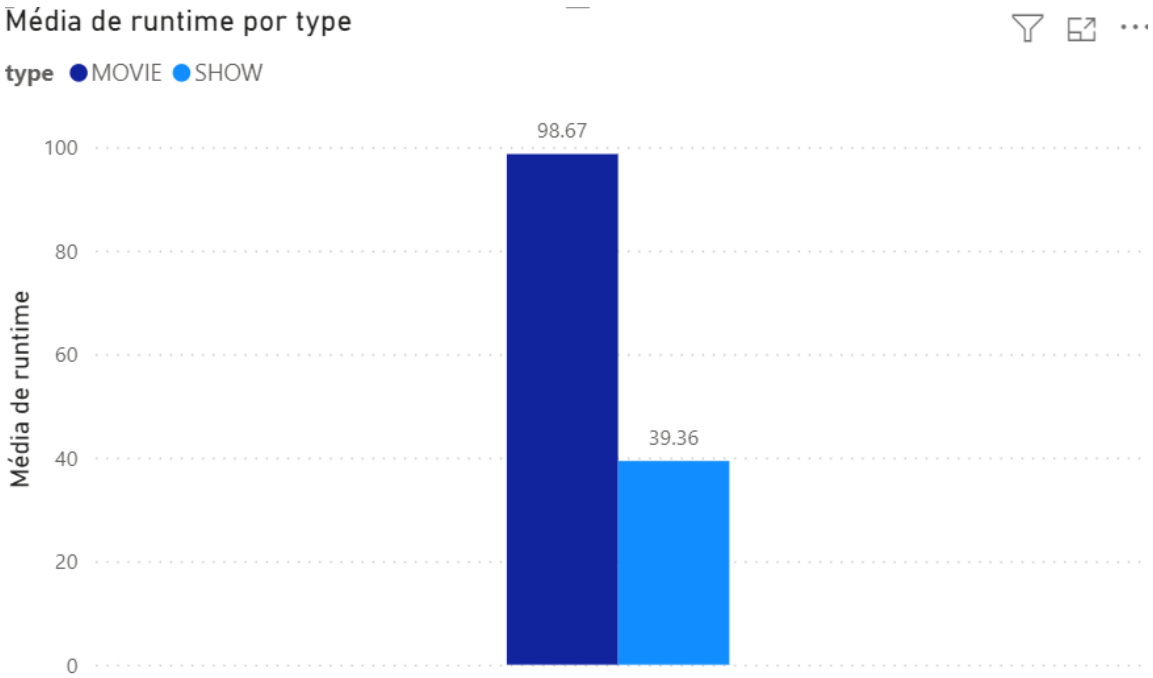

Figura 27 - Média de runtime por tipo (show ou filme)

Neste gráfico conseguimos observar que a média de duração de um filme é de 98.67 minutos e a média de duração de uma série é de 39.36 minutos.

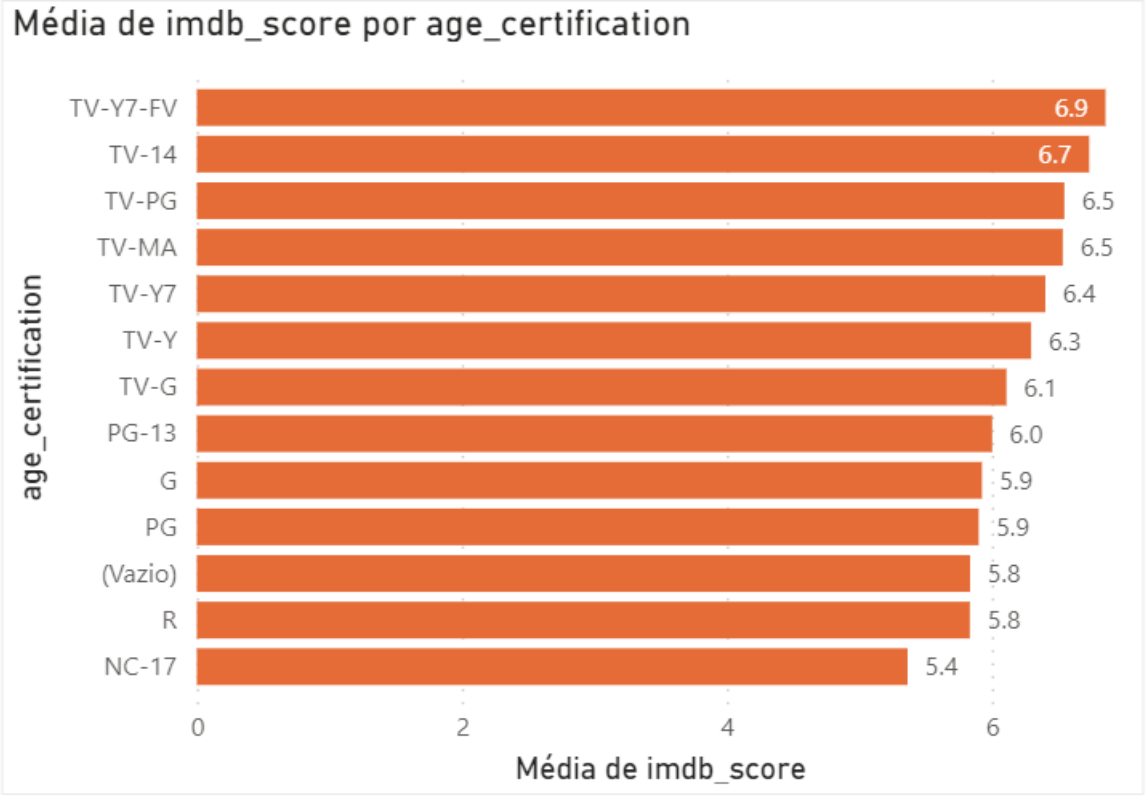

Figura 28 - Média de IMDb score por certificação de idade

Legenda:
TV-Y7-FV - O programa contém "fantasia violenta" e pode ser mais intenso do que outro programa classificado como TV-Y7.
TV-14 - Este tipo de programa contém material que os pais podem não considerar adequados para crianças menores de 14 anos.
TV-PG - Este tipo de programa contém material que pode não ser adequado para crianças mais novas.
TV-MA - Este tipo de programa é específico para adultos e por isso desaconselhado a menores de 18.
TV-Y7- Este tipo de programa é desaconselhado para menores de 7 anos.
TV-Y - Este tipo de programas são específicos para um público mais novo, incluindo crianças dos 2 aos 6 anos.
TV-G - Este tipo de programa não tem uma idade aconselhada ou restrita, sendo por isso, permitido a crianças mais novas.
PG-13 - Este tipo de programa é desaconselhado a menores de 13 anos.
G- Todas as idades são permitidas
PG- Este tipo de programas contém alguns tipos de materiais que não são aconselhados a crianças.

R - Este tipo de programa contém algum tipo de material adulto, sendo permitido a crianças menores de 18 mas acompanhadas por um adulto.
NC-17 - Ninguém com 17 anos ou mais novo deve ver este tipo de programa. Ou seja, tem de se ter pelo menos 18 anos para ver.

Neste gráfico podemos ver que os programas classificados como TV-Y7-FV são os que têm melhor avaliação (6.9) e os com menor avaliação (5.4) são os programas classificados como NC-17.

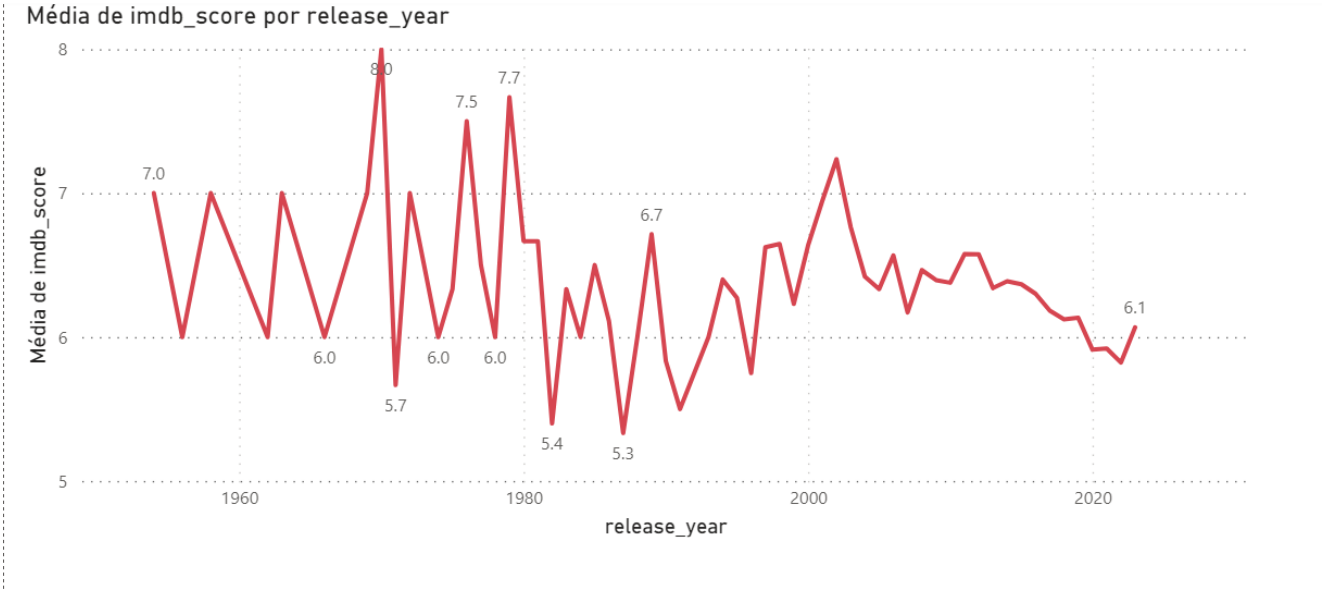

Figura 29 - Médi de IMDb score por ano de lançamento

Neste gráfico podemos observar que a média das avaliações feitas no imdb variam bastante conforme os anos em que os filmes ou séries foram lançados. Ainda assim, verificamos que em 1970 verificou-se a maior média de avaliações no imdb (8.0) e em 1987 a menor média de avaliações no imdb (5.3). Mais recentemente, em 2023 verificou-se uma média de avaliações de 6.1.

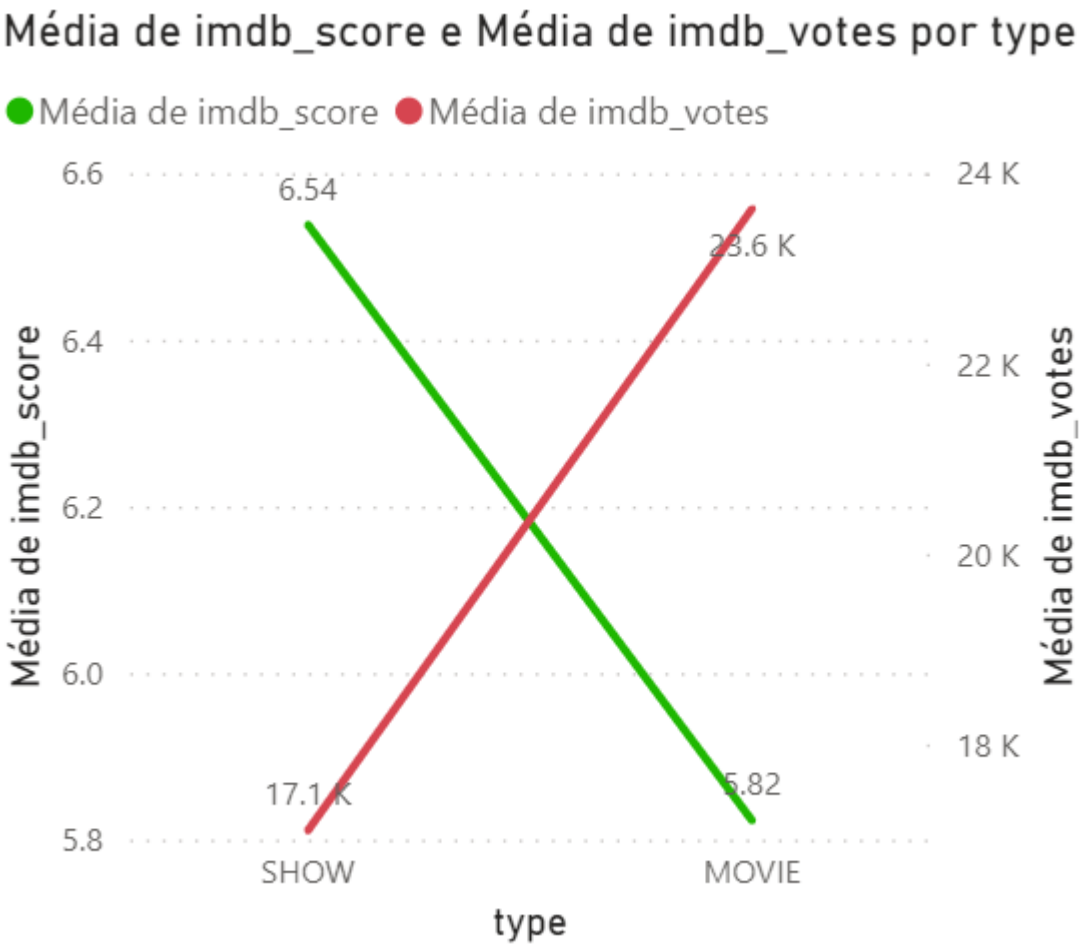

Figura 30 - Média de imdb score e votes por tipo

Neste gráfico podemos observar que existe uma relação inversa entre a média das avaliações obtidas e da média do número de avaliações feitas no imdb por filmes e séries. Assim, verificamos que enquanto nas séries a média das avaliações é de 6.54 e houve em média 17 mil avaliações, nos filmes a média das avaliações é de 5.82 e houve em média 24 mil avaliações. Ou seja, as séries em geral são melhor avaliadas mas têm um número menor de avaliações e os filmes têm uma avaliação mais baixa mas um maior número de avaliações.

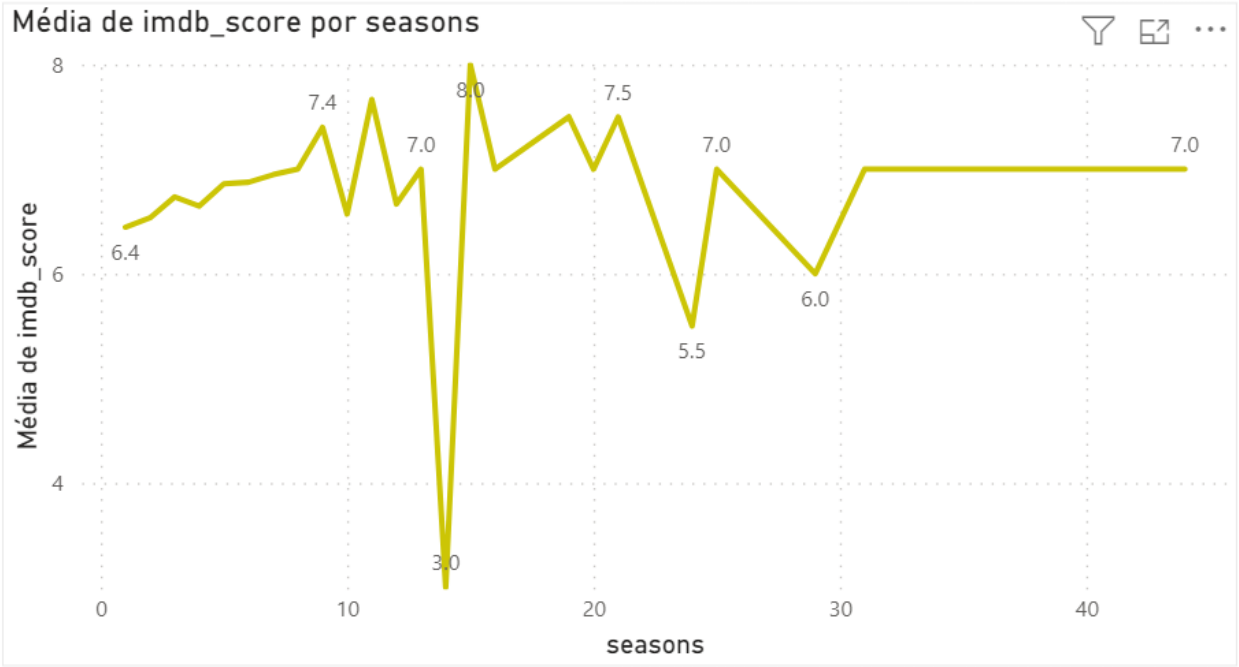

Figura 31 - Média de imdb_score por temporadas

Neste gráfico verificámos que as temporadas 14 em geral têm pior avaliação (3.0) e as temporadas 15 são as que têm melhor avaliação em média (8.0). As primeiras temporadas têm em média 6.4 de avaliação e as temporadas a partir da 31 têm avaliação de 7.0 em média.

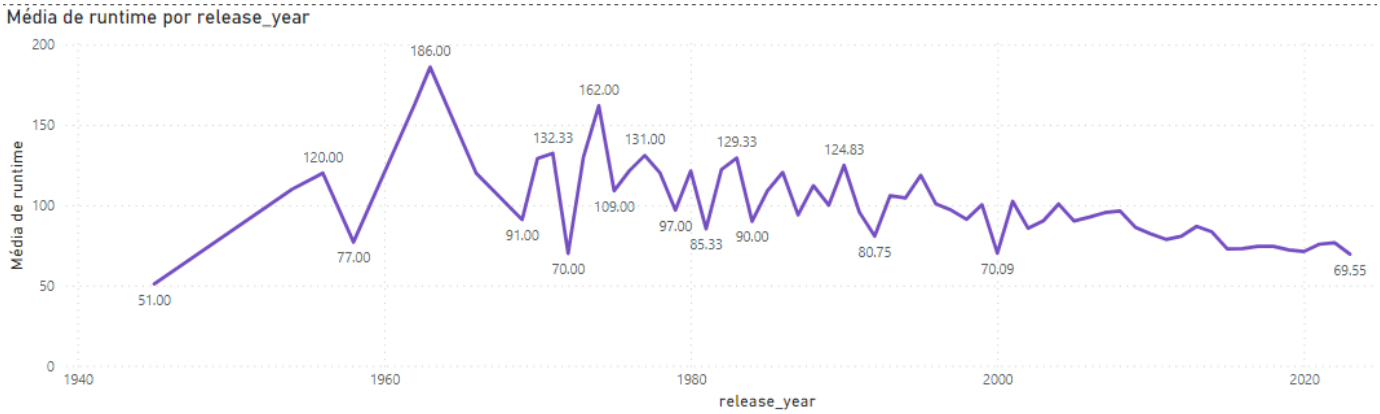

Figura 32 - Média de runtime por ano de lançamento

No gráfico 26 verificámos que em média os filmes e séries têm mais tempo de duração no ano de 1963 com 186 minutos. E o ano com séries e filmes com menos tempo de duração foi 1945 com 51 minutos. Em 2023 a média de tempo de duração é de 69 minutos.

Resultados

Os gráficos revelam que a média de score no IMDb é mais alta no Qatar e mais baixa no Burundi, e que filmes (98.67 minutos) geralmente têm duração maior que séries (39.36 minutos). Programas TV-Y7-FV são os mais bem avaliados (6.9) e NC-17 os menos (5.4). As avaliações médias variam por ano, com 1970 tendo a maior média (8.0) e 1987 a menor (5.3), e uma média recente de 6.1 em 2023. Séries têm melhores avaliações médias (6.54) que filmes (5.82), mas recebem menos avaliações. Temporadas 14 de séries são as piores avaliadas (3.0), enquanto as temporadas 15 são as melhores (8.0). Em termos de duração, 1963 teve a maior média (186 minutos) e 1945 a menor (51 minutos), com 2023 registando uma média de 69 minutos.

Concluímos assim que:

1. Programas para crianças com elementos de fantasia violenta (TV-Y7-FV) são bem recebidos, enquanto conteúdos adultos restritos (NC-17) têm avaliações mais baixas, possivelmente devido à natureza controversa desses programas.

2. A qualidade de filmes e séries tem variado ao longo das décadas, em anos como 1970 a serem bem avaliados, enquanto em anos como 1987 registam médias baixas. Isto pode refletir mudanças nas tendências de produção e consumo de multimédia.

3. As séries são geralmente melhor avaliadas, embora recebam menos avaliações do que os filmes, sugerindo um público com preferência específica para séries.

4. A qualidade das séries pode variar drasticamente entre temporadas, com algumas séries a melhorarem significativamente após várias temporadas.

5. A duração média de filmes e séries tem variado, refletindo mudanças nas preferências e padrões de produção ao longo do tempo.
Estas conclusões indicam tendências importantes na indústria de entretenimento, mostrando como a qualidade, duração e receção crítica

pode variar amplamente por região, tipo de conteúdo, classificação indicativa, época e formato.

Referências

Enrique, D. (março de 2023). *Netflix Movies and TV Shows*. Obtido de Kaggle: https://www.kaggle.com/datasets/dgoenrique/netflix-movies-and-tv-shows