Universidade Estadual de Campinas

Departamento de Estatística

ME315 - Manipulação de Banco de Dados

Professor Benilton Carvalho

Análise de Dados IMDb: Tendências em Filmes e Séries

Grupo A

Bernardo Abib - 236053

Juan Gabriel Sotomayor Lopes - 204303

Luiz Felipe de Oliveira Barbosa Nunes - 255403

Marcel Hideki Sousa Adati - 242910

Pedro Constantino de Freitas - 253596

Yejee Kim - 204397

Campinas - SP
Novembro de 2024

Conteúdo

1 Introdução

Este relatório tem como objetivo responder a três perguntas propostas, conforme as orientações do trabalho da disciplina ME315 - Manipulação de Banco de Dados. O estudo utiliza dados de produções audiovisuais (filmes e séries), disponibilizados gratuitamente pela plataforma *Internet Movie Database* (IMDb).

A partir de uma análise detalhada do banco de dados, foram definidas três perguntas principais para serem respondidas ao longo deste relatório:

- **Pergunta 1**: Existe uma relação entre a quantidade de temporadas/episódios e a qualidade das séries?
- **Pergunta 2**: Quais são os profissionais mais populares em cada área (tanto para séries quanto para filmes) e quais são seus projetos mais conhecidos?
- **Pergunta 3**: Quais são as produções mais populares de cada década?

2 Metodologia

2.1 Unificação dos Bancos em SQLite

Inicialmente, foram disponibilizados sete bancos de dados distintos, descritos a seguir:

- **title.akas.tsv.gz** - Informações sobre nomes alternativos estrangeiros das produções;
- **title.basics.tsv.gz** - Informações básicas das produções, como o seu tipo, nome, data de produção, gênero da obra;
- **title.crew.tsv.gz** - Nome dos diretores e escritores referentes às produções;
- **title.episode.tsv.gz** - Informações referentes a episódios de séries, como a série a qual pertence, o número da temporada e episódio;
- **title.principals.tsv.gz** - Informações sobre as pessoas envolvidas nas produções, e suas funções;
- **title.ratings.tsv.gz** - Informações referentes às avaliações dos usuários às produções;
- **name.basics.tsv.gz** - Informações sobre profissionais e suas obras mais creditadas.

Com o objetivo de facilitar a manipulação dos dados de forma integrada, todos os arquivos foram carregados em um único banco de dados SQLite denominado `dados_imdb`, onde cada um dos sete arquivos se tornou uma tabela distinta. O código em R utilizado para realizar essa tarefa está descrito a seguir:

```r
# Instalação e carregamento das bibliotecas necessárias
install.packages("RSQLite")
install.packages("readr")

library(RSQLite)
library(readr)

# Carregamento dos arquivos em data frames
banco1 <- read_tsv("name.basics.tsv")
banco2 <- read_tsv("title.akas.tsv")
banco3 <- read_tsv("title.basics.tsv")
banco4 <- read_tsv("title.crew.tsv")
banco5 <- read_tsv("title.episode.tsv")
banco6 <- read_tsv("title.principals.tsv")
banco7 <- read_tsv("title.ratings.tsv")

# Criação e conexão ao banco SQLite
con <- dbConnect(SQLite(), "dados\_imdb")

# Escrita dos data frames no banco de dados como tabelas
dbWriteTable(con, "name_basics", banco1, overwrite = TRUE)
dbWriteTable(con, "title_akas", banco2, overwrite = TRUE)
dbWriteTable(con, "title_basics", banco3, overwrite = TRUE)
dbWriteTable(con, "title_crew", banco4, overwrite = TRUE)
dbWriteTable(con, "title_episode", banco5, overwrite = TRUE)
dbWriteTable(con, "title_principals", banco6, overwrite = TRUE)
dbWriteTable(con, "title_ratings", banco7, overwrite = TRUE)

# Desconexão do banco
dbDisconnect(con)
```

O banco de dados resultante, denominado `dados_imdb`, foi armazenado no Google Drive para facilitar o acesso por todos os membros do grupo. Ele pode ser acessado através deste link.

2.2 Pergunta 1

A primeira pergunta, definida na Introdução — "Há uma relação entre a quantidade de temporadas/episódios e a qualidade das séries?"— é ampla e foi abordada em diferentes ângulos pois desejou-se que ela pudesse ser respondida de diversas maneiras. As consultas apresentadas a seguir refletem a três abordagens distintas, implementados tanto em SQLite quanto em Python.

2.2.1 Manipulação em SQLite

Em manipulações preliminares do banco obtido em 2.1, observou-se uma quantidade de produções (tanto filmes quanto séries) com pouquíssimas avaliações, o que pode impactar na visualização de dados sobre as produções audiovisuais mais populares e mais relevantes para o mundo. Com isso em mente, e para responder à pergunta 1, foi aplicado um filtro arbitrário de no mínimo 10 mil votos para as séries.

Portanto, um dos métodos propostos foi quantificar a qualidade de uma série. Para tal, é necessário definir o que faz uma série ser de boa qualidade. Foi feito um modelo que leva em consideração tanto a avaliação média da mídia quanto sua popularidade. O modelo foi definido como:

$$Qualidade = Nota * ln(Popularidade)$$

Em que Nota é a média ponderada das avaliações dos usuários, calculada pelos modelos do IMDb, cujo domínio de valores é entre 1 e 10 (inclusos), enquanto a Popularidade é a quantidade de votos que a série recebeu, podendo assumir qualquer valor positivo, exceto o 0. Isso permite um peso maior para a nota média que os usuários avaliaram a série, enquanto a função logarítmica diminui a influência na quantificação da qualidade para séries mais populares, mas ainda assim dando uma certa vantagem pois a função é sempre crescente.

Tendo em vista tudo isso, a *query* a seguir foi escrita também para capturar a qualidade de cada temporada individual, portanto inclui dados como o nome da série, a temporada referente, as notas, número de votos (popularidade), as qualidades quantificadas e outras informações por curiosidade.

```
series <- dbGetquery(conn, "SELECT nome_serie, temp, qnt_episodios, duracao_media_min, nota_temp, votos_temp, inicio_temp, nota_temp * LOG(votos_temp) AS qualidade_temp, tot_temp, fim_serie - inicio_serie + 1 AS anos_serie, (CAST(inicio_serie/ 10 AS INTEGER) * 10) || 's' AS decada_inicio, nota_serie, votos_serie, nota_serie * LOG(votos_serie) AS qualidade_serie
FROM
(SELECT parentTconst AS ID, primaryTitle AS nome_serie, MAX(CAST(seasonNumber AS INTEGER)) AS tot_temp, CAST(startYear AS INTEGER) AS inicio_serie, CAST(endYear AS INTEGER) AS fim_serie, averageRating AS nota_serie, numVotes AS votos_serie
 FROM title_episode
 INNER JOIN title_basics ON title_basics.tconst = title_episode.parentTconst
 INNER JOIN title_ratings ON title_ratings.tconst = title_basics.tconst
 WHERE startYear NOT GLOB '\\N' AND endYear NOT GLOB '\\N' AND seasonNumber NOT GLOB '\\N' AND votos_serie > 10000
 GROUP BY ID
 ORDER BY -votos_serie) AS a
INNER JOIN
(SELECT parentTconst AS ID, CAST(seasonNumber AS INTEGER) AS temp, CAST(startYear AS INTEGER) AS inicio_temp, AVG(averageRating) AS nota_temp, AVG(numVotes) AS votos_temp, MAX(CAST(episodeNumber AS INTEGER)) as qnt_episodios, AVG(runtimeMinutes) AS duracao_media_min
 FROM title_episode
 INNER JOIN title_basics ON title_basics.tconst = title_episode.tconst
 INNER JOIN title_ratings ON title_ratings.tconst = title_episode.tconst
 WHERE seasonNumber NOT GLOB '\\N' AND numVotes NOT GLOB '\\N'
 GROUP BY ID, temp) AS b
ON a.ID = b.ID
```

```
ORDER BY -qualidade_serie, -qualidade_temp")
```

A *query* foi ordenada em ordem decrescente por qualidade das séries, seguida pelas temporadas de maior qualidade. A seguir o resultado das 6 primeiras linhas e algumas das colunas de interesse (não necessariamente na ordem em que estão na *query*), como exemplo:

Tabela 1: Exemplo das 6 primeiras linhas e colunas da *query* 'series'

nome_serie	temp	qnt_episodios	tot_temp	qualidade_temp	qualidade_serie
Breaking Bad	5	16	5	102.65814	138.7426
Breaking Bad	4	13	5	93.26907	138.7426
Breaking Bad	1	7	5	90.90681	138.7426
Breaking Bad	2	13	5	90.11753	138.7426
Breaking Bad	3	13	5	89.99440	138.7426
Game of Thrones	6	10	8	101.11052	134.9281

Foi definido um limite de no máximo 20 temporadas de cada série para serem analisadas, pois séries com mais que isso atrapalhavam a análise. A mesma lógica foi aplicada para o filtro de no máximo 500 episódios.

Os códigos para ver graficamente as relações entre as qualidades das séries e temporadas pelo número e quantidade de temporadas e episódios são os seguintes:

Relação entre Qualidade da temporada e Número da temporada

```
series %>%
  filter (temp > 1 & temp < 20) %>%
  ggplot(aes(x = temp, y = qualidade_temp))+
  geom_jitter(width = 1, alpha = 0.25)+
  geom_smooth(method = "lm")+
  labs(x = "Temporada", y = "Qualidade da temporada")+
  theme_bw()
```

Relação entre Qualidade da série e Quantidade de episódios

```
series %>%
  filter (temp > 1 & temp < 20) %>%
  group_by(nome_serie) %>%
  summarise(qnt_episodios = sum(qnt_episodios), qualidade_serie) %>%
  filter(qnt_episodios < 500) %>%
  ggplot(aes(x = qnt_episodios, y = qualidade_serie))+
  geom_jitter(width = 1, alpha = 0.25)+
  geom_smooth(method="lm")+
  labs(x = "Quantidade de episodios", y = "Qualidade da serie")+
  theme_bw()
```

Relação entre a Qualidade da série e a Quantidade total de temporadas

```
series %>%
  filter (tot_temp > 1 & tot_temp < 20) %>%
  ggplot(aes(x = tot_temp, y = qualidade_serie))+
  geom_jitter(width = 1, alpha = 0.25)+
  geom_smooth(method="lm")+
  labs(x = "Quantidade total de temporadas", y = "Qualidade da serie")+
  theme_bw()
```

As análises estão na seção 3.2.1.

2.2.2 Manipulação em Python

Além da abordagem do banco de dados em SQLite quantificando a qualidade, foi realizada também uma análise geral em Python, investigando a relação entre o número de episódios e as avaliações de todas as séries de 1980 a 2024. A leitura e a organização dos dados necessários para a observação foi feita da seguinte forma:

```
import polars as pl

episode = pl.read_csv("/content/drive/MyDrive/ME315py/title.episode.tsv.gz",
    separator = '\t', null_values = '\\N')

ratings = pl.read_csv("/content/drive/MyDrive/ME315py/title.ratings.tsv.gz",
    separator = '\t', null_values = '\\N')

tlbasics = pl.read_csv("/content/drive/MyDrive/ME315py/title.basics.tsv.gz",
    separator = '\t', null_values = '\\N', ignore_errors=True)

table1 = (
    tlbasics.filter([pl.col("titleType") == "tvSeries", pl.col("startYear")
   >= 1980])
    .join(ratings, on = "tconst", how = "inner")
    .join(episode, left_on="tconst", right_on="parentTconst", how = "inner")
    .drop_nulls("averageRating")
    .drop_nulls("episodeNumber")
    .drop_nulls("seasonNumber")
    .select(["primaryTitle", "startYear", "averageRating", "numVotes"])
    .group_by("primaryTitle")
    .agg(pl.all().first(), pl.len().alias("número de episódios"))
    .sort("averageRating", "número de episódios", "numVotes", descending =
   True)
)
```

Devido ao grande volume de filmes analisados, para garantir uma visualização mais clara, o resultado foi dividido em 2 gráficos (1980 a 2000 e 2000 a 2024).

```
    t8000 = (
    table1.filter([pl.col("startYear") >= 1980, pl.col("startYear") <=
   2000])
    .sort("startYear", "averageRating", "número de episódios")
)
    t0020 = (
    table1.filter(pl.col("startYear") >= 2000)
    .sort("startYear", "averageRating", "número de episódios")
)
gparte1 = (
    ggplot(t8000, aes(x = "averageRating", y = "número de episódios", color
   = "startYear"))
    + stat_smooth(method = "lm", se = False)
    + facet_wrap("startYear")
    + theme_bw()
    + labs(title = "Relação entre número de episódios e avaliação média", x
   = "Avaliação média (0-10)", y = "Número de Episódios", color = "Ano de
   Estreia")
    )
gparte1

gparte2 = (
    ggplot(t0020, aes(x = "averageRating", y = "número de episódios", color
   = "startYear"))
```

```
    + stat_smooth(method = "lm", se = False)
    + facet_wrap("startYear")
    + theme_bw()
    + labs(title = "Relação entre número de episódios e avaliação média", x
    = "Avaliação média (0-10)", y = "Número de Episódios", color = "Ano de
    Estreia")
    )
gparte2
```

2.3 Pergunta 2

A pergunta 2, definida na Introdução, "Quais são os profissionais mais populares de cada área (para séries e filmes), e seus projetos mais conhecidos", pode ser respondida pelas seguintes consultas.

2.3.1 Profissionais populares em filmes por área

Na primeira consulta, temos o código para identificar os profissionais mais populares de cada área, considerando a soma do número de votos recebidos em seus projetos:

```
profissionais_populares <- dbGetQuery(conn, "
  SELECT Nome_Profissional, Categoria, Numero_de_Projetos, Soma_de_Votos
  FROM ( SELECT np.primaryName AS Nome_Profissional, tp.category AS
    Categoria, COUNT(tp.tconst) AS Numero_de_Projetos, SUM(tr.numVotes) AS
    Soma_de_Votos, ROW_NUMBER() OVER (PARTITION BY tp.category ORDER BY SUM(
    tr.numVotes) DESC) AS rank
  FROM name_basics np
  INNER JOIN title_principals tp ON np.nconst = tp.nconst
  INNER JOIN title_basics tb ON tp.tconst = tb.tconst
  INNER JOIN title_ratings tr ON tb.tconst = tr.tconst
  WHERE tb.titleType = 'movie'
  GROUP BY Categoria, Nome_Profissional ) AS rank
  WHERE rank = 1
  ORDER BY Categoria, Soma_de_Votos DESC;
")
```

Essa consulta agrupa os profissionais por categoria, calcula o número total de projetos e a soma de votos para cada profissional, e seleciona o profissional mais popular de cada área, baseado no número de votos.

2.3.2 Profissionais populares e seus projetos mais populares

Para incluir o título do projeto mais popular de cada profissional, utilizamos o seguinte código:

```
profissionais_projetos_populares <- dbGetQuery(conn, "
  WITH ProfissionaisMaisPopulares AS ( SELECT np.primaryName AS Nome_
    Profissional, tp.category AS Categoria, np.nconst, SUM(CASE WHEN tr.
    numVotes != '\\N' THEN tr.numVotes ELSE 0 END) AS Soma_de_Votos, ROW_
    NUMBER() OVER (PARTITION BY tp.category ORDER BY SUM(CASE WHEN tr.
    numVotes != '\\N' THEN tr.numVotes ELSE 0 END) DESC) AS rank FROM name_
    basics np
  INNER JOIN title_principals tp ON np.nconst = tp.nconst
  INNER JOIN title_basics tb ON tp.tconst = tb.tconst
  INNER JOIN title_ratings tr ON tb.tconst = tr.tconst
  WHERE tb.titleType = 'movie'
  GROUP BY np.nconst, np.primaryName, tp.category),
  ProjetosMaisPopulares AS (SELECT tp.nconst, tb.primaryTitle AS Projeto_
    Mais_Popular, tr.numVotes AS Numero_de_Votos_Projeto, ROW_NUMBER() OVER (
    PARTITION BY tp.nconst ORDER BY tr.numVotes DESC) AS rank_projeto FROM
    title_principals tp
```

```
  INNER JOIN title_basics tb ON tp.tconst = tb.tconst
  INNER JOIN title_ratings tr ON tb.tconst = tr.tconst
  WHERE tb.titleType = 'movie')
  SELECT pmp.Nome_Profissional,pmp.Categoria, pmp.Soma_de_Votos, pmp.rank,
   pp.Projeto_Mais_Popular, pp.Numero_de_Votos_Projeto FROM
   ProfissionaisMaisPopulares pmp LEFT JOIN ProjetosMaisPopulares pp ON pmp.
   nconst = pp.nconst
  WHERE pmp.rank = 1 AND pp.rank_projeto = 1
  ORDER BY pmp.Categoria;
")
```

Essa consulta utiliza a função `ROW_NUMBER` para identificar o projeto com o maior número de votos para cada profissional, assim como o total de votos somados para cada profissional em sua categoria. Dessa forma, obtemos os profissionais mais populares em cada área, junto com seus projetos mais conhecidos e o número de votos desses projetos.

As categorias dos profissionais foram traduzidas livremente do inglês para o português na tabela final para melhor compreensão.

2.4 Pergunta 3

A pergunta 3, definida na Introdução, "Quais as produções mais populares por década?", pode ser respondida com algumas simples *queries*.

Para definir popularidade, foi utilizado a mesma definição proposta nas outras perguntas, na qual popularidade é a quantidade de votos que a mídia recebeu. Além disso, são mostrados apenas os três resultados mais populares por década.

2.4.1 Séries mais populares por década

Para as séries mais populares por década, a *query* 'series', também feita em 2.2.1, pode ser reutilizada e trabalhada em cima.

O código para averiguar as três séries mais populares por década:

```
series %>%
  group_by(decada_inicio, nome_serie) %>%
  summarise(ano_de_lancamento = min(inicio_temp), nota_serie = max(nota_
   serie), votos_serie = max(votos_serie)) %>%
  group_by(decada_inicio) %>%
  slice_max(order_by = votos_serie, n = 3)
```

O código agrupa as séries por nome e década de início. Em seguida, pega o ano de lançamento da primeira temporada, a nota da série e os votos. Tudo isso é agrupado novamente pela década e ordenado pelas três séries mais populares por década.

2.4.2 Filmes mais populares por década

Para identificar os três filmes mais populares de cada década utilizamos a seguinte *query*:

```
filmes_populares_top3 <- dbGetQuery(conn, "
  SELECT startYear / 10 * 10 AS Decada, primaryTitle AS `Nome do Filme`,
   startYear AS `Ano de Lancamento`, averageRating AS `Media das Avaliacoes
   `, numVotes AS `Quantidade de Votos`
  FROM (SELECT startYear, primaryTitle, averageRating, numVotes, ROW_NUMBER
   () OVER (PARTITION BY startYear / 10 * 10 ORDER BY numVotes DESC) AS rn
  FROM title_basics AS tb
  INNER JOIN title_ratings AS tr ON tb.tconst = tr.tconst
  WHERE tb.titleType = 'movie' AND tb.startYear != '\\N') AS ranked
  WHERE rn <= 3
  ORDER BY Decada ASC, numVotes DESC;
```

```
")
```

Essa consulta organiza os filmes por década, calculando a década a partir do ano de lançamento com a operação `startYear / 10 * 10`. Dentro de cada década, utilizamos a função `ROW_NUMBER()` para atribuir uma classificação aos filmes com base na quantidade de votos (`numVotes`) em ordem decrescente. Dessa forma, são selecionados apenas os três filmes mais votados de cada década. No resultado final, a ordenação é feita primeiramente pela década e, em seguida, pela quantidade de votos, destacando os filmes mais populares de cada período.

2.4.3 Episódios mais populares por década

O código para identificar os três episódios de séries mais populares por década:

```
episodios <- dbGetQuery(conn, "SELECT nome_serie, nome_ep, temp, ep, decada_
    inicio, nota_ep, votos_ep, inicio_temp
     FROM
     (SELECT parentTconst AS ID, primaryTitle as nome_ep, CAST(seasonNumber
    AS INTEGER) AS temp, CAST(episodeNumber AS INTEGER) AS ep,
     (CAST(CAST(startYear AS INTEGER)/ 10 AS INTEGER) * 10) || 's' AS decada_
    inicio, CAST(startYear AS INTEGER) AS inicio_temp,averageRating AS nota_
    ep, numVotes as votos_ep
     FROM title_episode
     INNER JOIN title_basics ON title_basics.tconst = title_episode.tconst
     INNER JOIN title_ratings ON title_ratings.tconst = title_episode.tconst
     WHERE seasonNumber NOT GLOB '\\N' AND numVotes NOT GLOB '\\N' AND
    startYear NOT GLOB '\\N') AS a
     INNER JOIN
     (SELECT parentTconst AS ID, primaryTitle as nome_serie
     FROM title_episode
     INNER JOIN title_basics ON title_basics.tconst = title_episode.
    parentTconst
     INNER JOIN title_ratings ON title_ratings.tconst = title_episode.tconst
     GROUP BY parentTconst) AS b
     ON a.ID = b.ID")
```

A *query* acima é semelhante à *query* feita em 2.2.1, com a diferença de que ela pega informações sobre os episódios das séries.

```
episodios %>%
  group_by(decada_inicio, nome_serie, nome_ep) %>%
  summarise(ano_de_lancamento = min(inicio_temp), nota_ep = max(nota_ep),
    votos_ep = max(votos_ep)) %>%
  group_by(decada_inicio) %>%
  slice_max(order_by = votos_ep, n = 3)
```

O código acima é semelhante ao feito em 2.4.1.

3 Análise e Resultados

3.1 Análise Descritiva

A seguir, apresentamos uma análise descritiva dos dados dos filmes, incluindo estatísticas das avaliações, popularidade dos gêneros, distribuição de durações e tipos de produção.

Tabela 2: Estatísticas descritivas das avaliações dos filmes

Média das Avaliações	Mediana das Avaliações	Número de Filmes	Mínima Avaliação	Máxima Avaliação
6.18	6.3	315252	1	10

A Tabela 2 apresenta as estatísticas descritivas das avaliações dos filmes. A média das avaliações é de 6.18, com uma mediana de 6.3, indicando uma leve inclinação positiva nos escores. O filme mais bem avaliado possui nota 10, enquanto o menos avaliado tem nota 1. Com um total de 315,252 filmes avaliados.

Tabela 3: Gêneros de filmes mais populares

Gênero	Número de Filmes	Média das Avaliações
Drama	73882	6.32
Documentary	36741	7.20
Comedy	27991	5.74
Romance	19459	6.10
Thriller	14314	5.57
Horror	10702	4.88
Drama, Romance	7193	6.18
Family	5300	6.27
Crime, Drama	5156	6.01
Action	4435	5.75

Na Tabela 3, observamos os dez gêneros mais populares em termos de quantidade de filmes e suas médias de avaliação. O gênero **Drama** lidera com 73,882 filmes e uma média de 6.32 nas avaliações. O **Documentário** apresenta a maior média de avaliação, com 7.20, apesar de ter metade do volume de produções em comparação com o drama.

Tabela 4: Distribuição de filmes por intervalo de duração

Intervalo de Duração	Número de Filmes
1h - 1h30	185437
Até 1h	181356
Mais de 2h30	68102

A Tabela 4 mostra a distribuição dos filmes conforme o intervalo de duração. A maior parte dos filmes tem entre 1h e 1h30 de duração, com um total de 185,437 filmes, enquanto filmes mais curtos (até 1h) são quase tão frequentes, com 181,356 produções. Enquanto, filmes com mais de 2h30 são menos comuns, totalizando 68,102.

Tabela 5: Distribuição de tipos de produção

Tipo de Produção	Número de Produções
tvEpisode	8470179
short	1014777
movie	691015
video	297390
tvSeries	269362
tvMovie	148626
tvMiniSeries	56428
tvSpecial	49184
videoGame	39735
tvShort	10392
tvPilot	1

A Tabela 5 detalha os diferentes tipos de produção. O tipo **tvEpisode** é o mais frequente, com mais de 8 milhões de registros, seguido por **short** e **movie**, com 1,014,777 e 691,015 produções, respectivamente. Esses números destacam a popularidade de episódios de séries de televisão e curtas-metragens, além dos filmes tradicionais.

Além disso, observamos que a categoria `tvPilot` possui apenas uma produção listada, com o título "TV Pilot". Esse "TV Pilot", datado de 1991, parece ser uma entrada genérica ou um *placeholder* no banco de dados, especialmente porque o campo de gênero está marcado como "\N", indicando um valor ausente ou não especificado. Portanto, essa entrada não representa um programa específico, mas sim o conceito geral de um piloto de TV, utilizado para introduzir e testar novos programas.

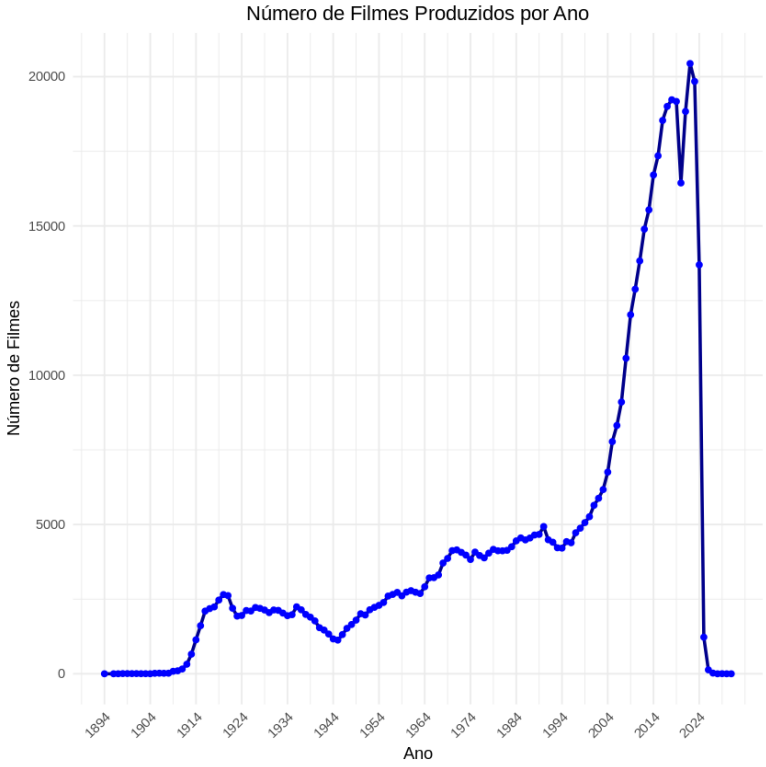

Figura 1: Número de Filmes Produzidos por Ano

Na Figura 1, observamos o crescimento do número de filmes produzidos ao longo dos anos, com um aumento significativo a partir dos anos 2000, possivelmente impulsionado pela globalização da indústria cinematográfica e pelo advento de plataformas de streaming.

Além disso, a queda no final do gráfico é completamente normal, pois se refere ao período atual. Muitos

filmes ainda estão em processo de produção ou serão lançados em breve, o que explica o número reduzido de produções registradas.

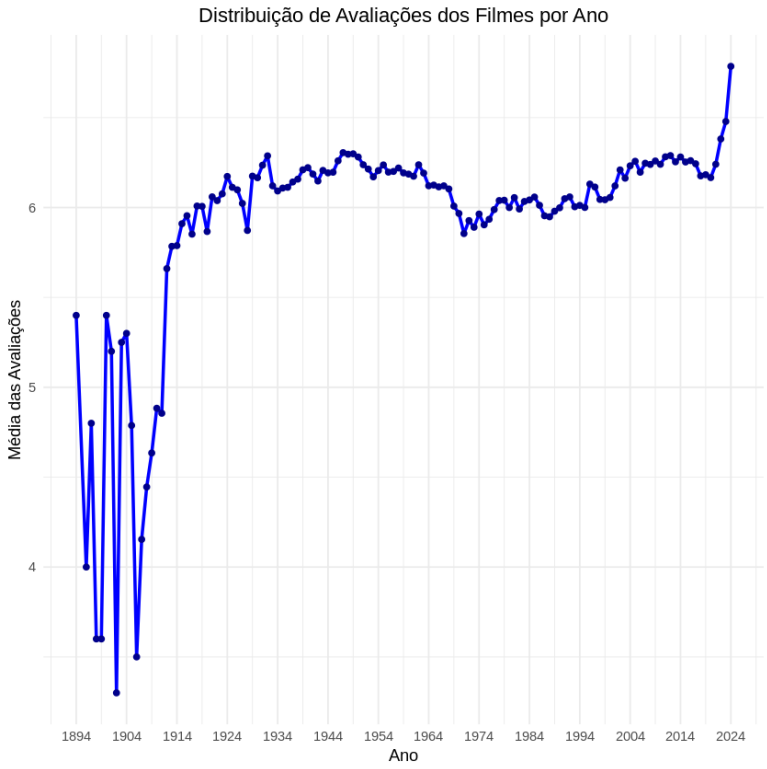

Figura 2: Distribuição de Avaliações dos Filmes por Ano

A Figura 2 exibe a média das avaliações dos filmes por ano. Observamos uma leve tendência de estabilidade, a partir de 1914, na qualidade média ao longo do tempo, e com algumas variações durante o período inicial de produção cinematográfica.

3.2 Pergunta 1

A pergunta 1 foi definida na Introdução, "Há uma relação entre quantidade de temporadas/episódios com a qualidade da série?".

3.2.1 Quantificando a qualidade

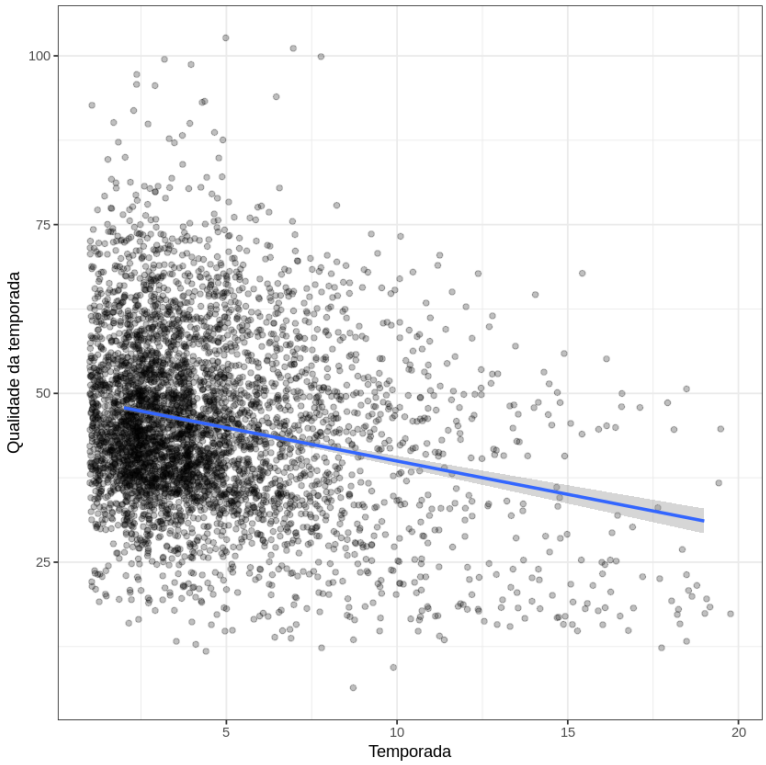

Figura 3: Relação entre Qualidade da temporada e Número da temporada

O gráfico 3 acima mostra a relação entre a qualidade de cada temporada analisada e o número da temporada. Percebe-se que há um grande aglomerado de temporadas de número baixo (até 5, aproximadamente). Isso acontece porque há um maior número dessas temporadas. Mas o mais interesse de se notar é que, **quanto maior o número da temporada** (digamos, por exemplo, temporada 15), **menor é a qualidade dela quando comparada com temporadas iniciais das séries**. Isso fornece uma dica sobre o comportamento dessa relação estudada.

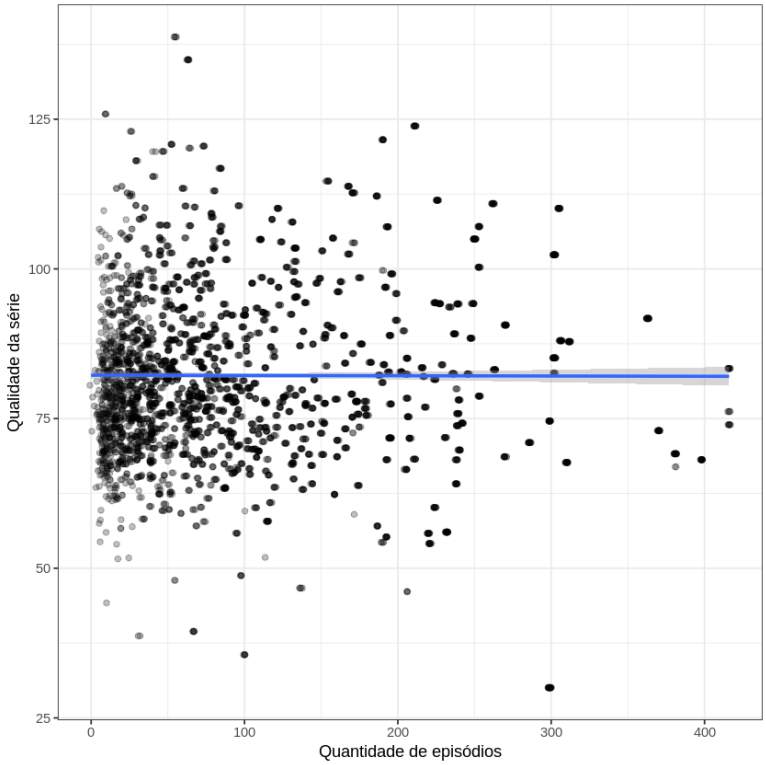

Figura 4: Relação entre Qualidade da Série e Quantidade de episódios

Na figura 4 acima, ao contrário do gráfico da figura 3, mostra a relação entre a qualidade das séries analisadas pela quantidade de episódios. Em contraste com o caso anterior, aqui há uma relação neutra entre as duas variáveis, ou seja, **a quantidade de episódios não parece influenciar a qualidade das séries**.

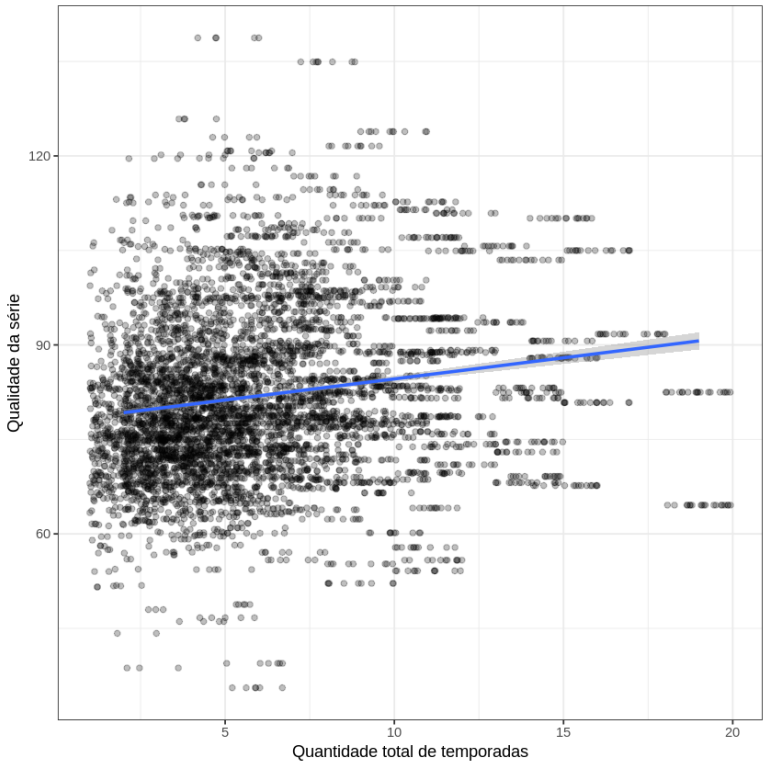

Figura 5: Relação entre a Qualidade da série e a Quantidade total de temporadas

Agora, na figura 5 acima, há a relação entre a qualidade da série e o número total de temporadas. Do contrário do que se esperava, há uma relação positiva entre as duas variáveis, ou seja, **quanto maior o número de temporadas de uma série, maior sua qualidade geral**.

Tendo em vista esses três gráficos, como pode temporadas avançadas influenciarem negativamente em suas qualidades individuais, mas o elevado número de episódios não influencia a qualidade da série, e quanto mais temporadas, melhor a série? Isso pode ser explicado por algumas razões:

- Pela diferença na quantidade de episódios por temporada nas diferentes séries (temporadas podem variar muito na quantidade de episódios);
- Pelo fenômeno de que uma série é avaliada positivamente, mesmo que tenha temporadas consideradas ruins - os usuários tendem a avaliar a série de maneira geral pelos bons momentos dela.

Pode-se concluir que a quantidade de episódios em uma série **não** afeta diretamente sua qualidade. No entanto, observa-se que **séries com um maior número de temporadas tendem a ser mais bem avaliadas**. Ainda assim, **é notável que as avaliações das temporadas individuais diminuem à medida que a série avança**.

3.2.2 Relação entre Número de episódios e Avaliação média (Python)

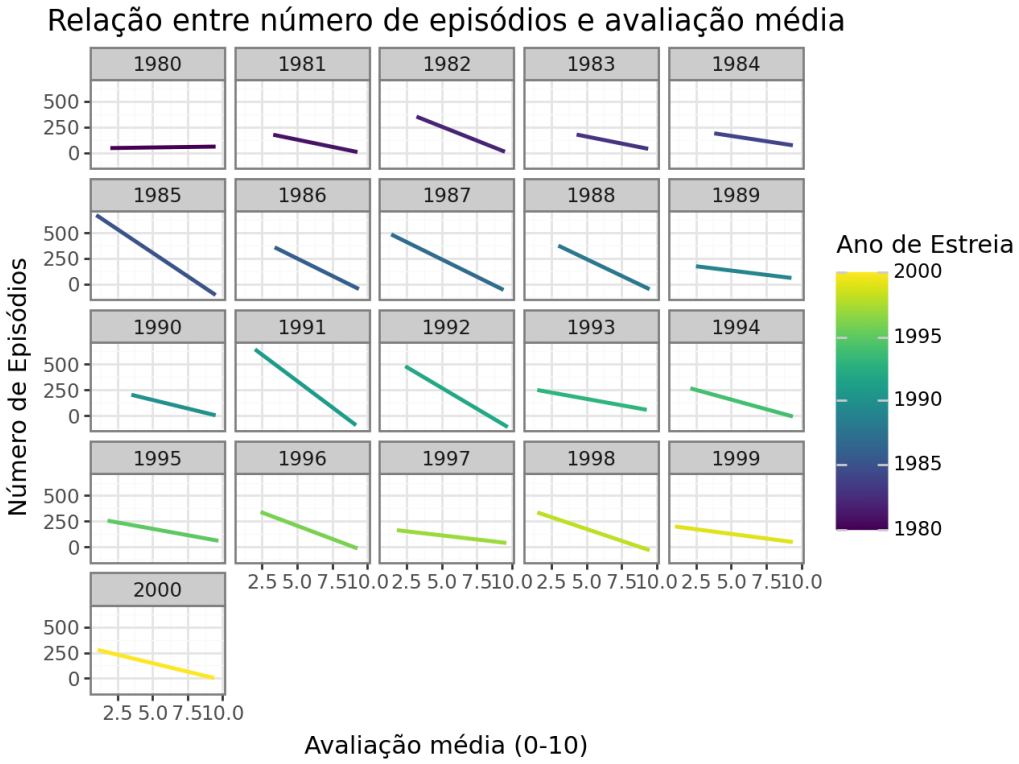

Figura 6: Séries entre 1980 - 2000

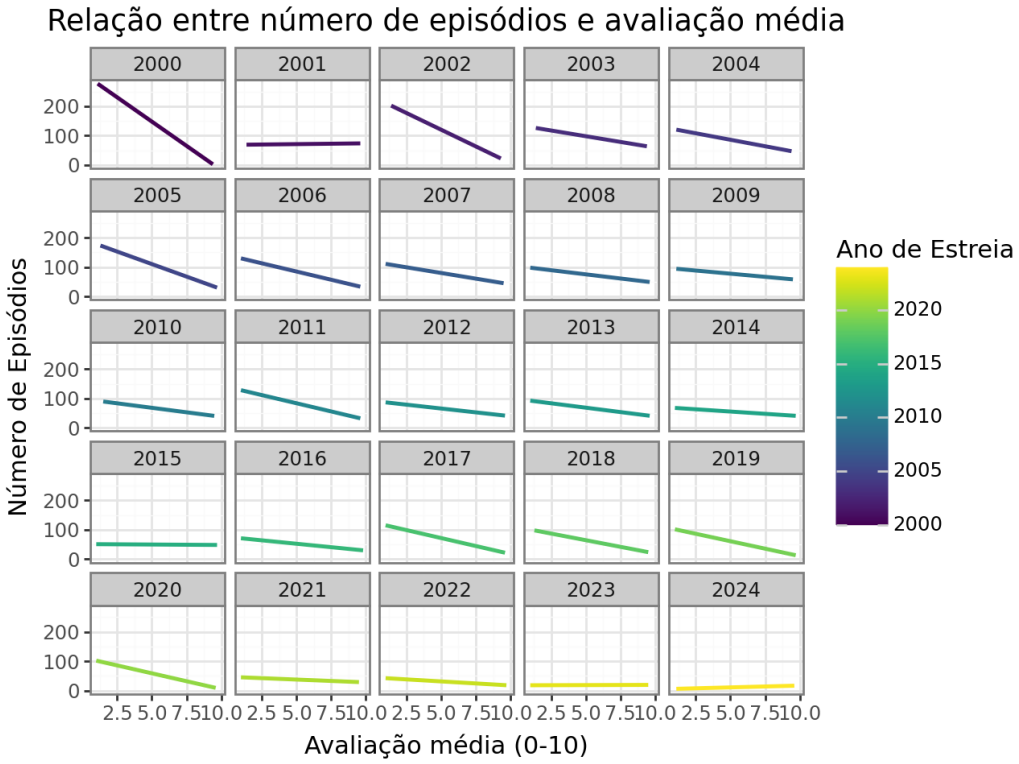

Figura 7: Séries entre 2000 - 2024

A partir da análise das figuras 6 e 7, observa-se que, de maneira geral, **quanto menor o número de episódios, melhores tendem a ser as avaliações, embora haja algumas exceções**.

No entanto, o que chama mais atenção é que, na Figura 7, o número de episódios é significativamente inferior ao da Figura 6. Ou seja, ao longo dos anos, **as séries têm reduzido a quantidade de temporadas e episódios**, especialmente nos anos mais recentes.

3.2.3 Análise de regressão (Python)

Na figura 8 a seguir, observa-se uma possível relação negativa entre as duas variáveis (nota média da temporada e o número da temporada), sugerindo um possível declínio da média das notas conforme o número de temporadas aumenta:

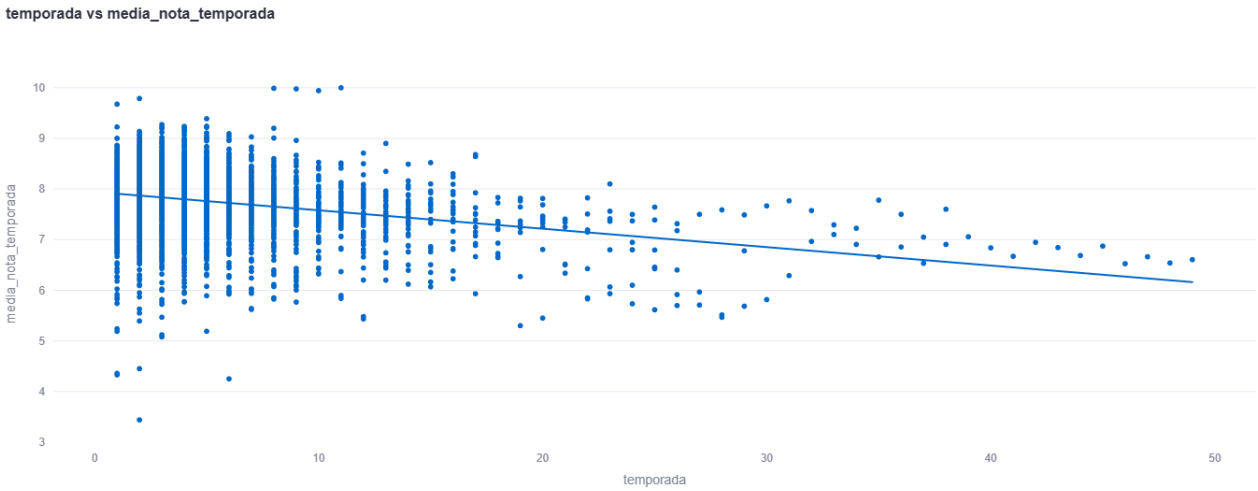

Figura 8: Relação entre Nota média e Número da temporada

A tabela 6 de coeficientes abaixo confirmou essa hipótese, possuindo o coeficiente angular negativo e o p-valor significativo (muito próximo de zero):

Tabela 6: Tabela de coeficientes

Preditor	**Coeficiente**	**Erro padrão**	**T-valor**	**P(>|t|)**
const	7.9429	0.0126	630.7295	0.0000
temporada	-0.0364	0.0019	-19.5077	0.0000

Para alcançar esses resultados, foram tomadas as seguintes medidas: 1. Foram consideradas apenas séries que possuem mais de uma temporada 2. Estabeleceu-se um número mínimo de votos que a série deve ter recebido para ser considerada.

Com o segundo critério, foi observado um padrão interessante: conforme foi aumentado o número mínimo de votos a serem considerados, houve também um aumento no coeficiente de determinação do modelo, como é vista na figura 9 a seguir.

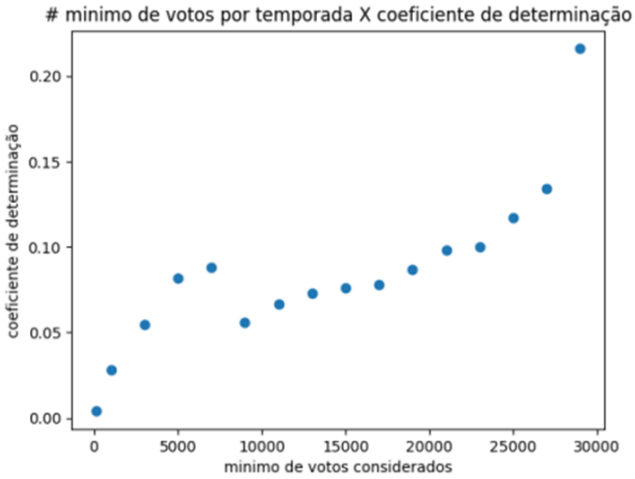

Figura 9: Coeficiente de determinação por número de votos

Esse possível padrão sugere que **quanto mais popular a série, maior parece ser o impacto que o número de temporadas exerce na nota média**. Visto que popularidade de uma série é um fator importante para a produção de mais temporadas, e que o mínimo de votos considerados é para toda a série – e não apenas para a temporada – é possível sugerir duas hipóteses interessantes:

- Séries mais populares tem uma maior tendência a produzirem mais temporadas puramente pelo engajamento; nem sempre conseguindo manter o nível;
- Séries mais populares criam maior expectativa no público, que pode fazer com que este seja mais crítico quanto a novas temporadas.

3.3 Pergunta 2

A análise da **Pergunta 2** visa identificar os profissionais mais populares em diversas áreas de atuação no cinema (como ator, atriz, diretor, entre outros) e destacar os projetos mais conhecidos de cada um. O critério de popularidade utilizado foi a soma dos votos recebidos nas produções em que esses profissionais participaram, o que indica o impacto de suas contribuições na percepção do público.

3.3.1 Tabela dos Profissionais mais Populares de cada Área

A Tabela 7 apresenta os profissionais mais populares de cada área, com informações sobre a quantidade total de projetos em que trabalharam e a soma de votos recebidos em suas produções. Já a Tabela 8 fornece, além dos profissionais mais populares, o projeto de maior destaque de cada um, com seus respectivos números de votos.

Tabela 7: Profissionais mais Populares por Área

Nome Profissional	Categoria	Número de Projetos	Soma de Votos
Samuel L. Jackson	ator	112	22871745
Scarlett Johansson	atriz	61	24488256
George W. Bush	imagens de arquivo	122	584745
Stan Freberg	som de arquivo	1	143777
Sarah Finn	diretor de elenco	125	36850115
Roger Deakins	diretor de fotografia	63	14738498
Hans Zimmer	compositor	153	37635920
Christopher Nolan	diretor	12	16498701
Lee Smith	editor	34	18092635
Kevin Feige	produtor	34	24843316
Nathan Crowley	designer de produção	19	13302662
Johnny Knoxville	ele-mesmo	10	511304
Christopher Nolan	roteirista	12	16995035

Tabela 8: Profissionais mais Populares por Área e seus Projetos mais conhecidos

Nome Profissional	Categoria	Projeto Mais Popular	Número de Votos
Samuel L. Jackson	ator	Pulp Fiction	2256119
Scarlett Johansson	atriz	The Avengers	1476184
George W. Bush	imagens de arquivo	Bowling for Columbine	149453
Stan Freberg	som de arquivo	Lady and the Tramp	150788
Sarah Finn	diretor de elenco	The Avengers	1476184
Roger Deakins	diretor de fotografia	The Shawshank Redemption	2937574
Hans Zimmer	compositor	The Dark Knight	2917884
Christopher Nolan	diretor	The Dark Knight	2917884
Lee Smith	editor	The Dark Knight	2917884
Kevin Feige	produtor	The Avengers	1476184
Nathan Crowley	designer de produção	The Dark Knight	2917884
Johnny Knoxville	ele-mesmo	Men in Black II	409403
Christopher Nolan	roteirista	The Dark Knight	2917884

3.3.2 Diversidade de contribuições e reconhecimento na Indústria Cinematográfica

É interessante notar que **George W. Bush** aparece na lista dos profissionais mais populares na categoria *imagens de arquivo*, com 122 aparições e uma soma de votos de 584.745. Esse destaque se deve ao fato de que, como ex-presidente dos Estados Unidos, George W. Bush é frequentemente retratado em documentários, programas de televisão e filmes de arquivo. Esse tipo de material utiliza imagens e vídeos históricos de figuras públicas, especialmente líderes políticos, para contextualizar eventos e narrativas de importância nacional e mundial. Dessa forma, sua presença reflete não uma atuação direta no cinema, mas a alta popularidade de conteúdos nos quais ele é mencionado ou aparece em filmagens de arquivo.

Além disso, outros pontos interessantes da análise dos profissionais mais populares:

- **Samuel L. Jackson** e **Scarlett Johansson** são os atores e atrizes mais populares. A popularidade desses atores é impulsionada por suas participações no Universo Cinematográfico Marvel, que atrai uma vasta base de fãs e gera altos volumes de votos.
- **Hans Zimmer**, como compositor, destaca-se pela maior soma de votos em seus projetos. Suas composições são marcantes em produções populares de diversos gêneros, demonstrando a relevância da música para a experiência cinematográfica.
- **Christopher Nolan** aparece tanto como **diretor** quanto como **roteirista**, reforçando sua reputação de autor em Hollywood. Filmes dirigidos e escritos por Nolan, como *Batman: O Cavaleiro das Trevas*, *Interstellar*, *Oppenheimer*, entre outros.
- A presença de profissionais como **Sarah Finn** (diretora de elenco) e **Roger Deakins** (diretor de fotografia) ilustra que a popularidade não se limita a atores e diretores. Profissionais técnicos também recebem reconhecimento significativo, especialmente quando colaboram em filmes de grande visibilidade. Sarah Finn, por exemplo, é responsável pelo elenco de diversos filmes da Marvel, como *Vingadores: Ultimato*, enquanto Roger Deakins é amplamente aclamado por seu trabalho visual em produções como *Blade Runner 2049* e *007 - Operação Skyfall*.

3.4 Pergunta 3

A análise da **Pergunta 3** busca identificar as produções audiovisuais mais populares (ou seja, pela quantidade de votos recebidos). Busca-se identificar as top três séries, filmes e episódios mais populares por década.

Tabela 9: Top 3 Séries Mais Populares por Década

Década	Nome da Série	Ano de Lançamento	Média das Avaliações	Quantidade de Votos
1950	The Twilight Zone	1960	9.1	95501
	I Love Lucy	1952	8.5	29305
	Alfred Hitchcock Presents	1956	8.5	19568
1960	Star Trek	1967	8.4	94453
	Monty Python's Flying Circus	1970	8.8	79490
	Scooby Doo, Where Are You!	1970	7.9	42645
1970	Fawlty Towers	1975	8.8	101281
	M*A*S*H	1973	8.5	64926
	Columbo	1972	8.3	43186
1980	Seinfeld	1990	8.9	357441
	Star Trek: The Next Generation	1988	8.7	138829
	Married... with Children	1987	8.1	112412
1990	Friends	1995	8.9	1106980
	The Sopranos	1999	9.2	488273
	The X-Files	1994	8.6	254317
2000	Breaking Bad	2008	9.5	2201125
	The Big Bang Theory	2008	8.1	883185
	Dexter	2006	8.6	789989
2010	Game of Thrones	2011	9.2	2341026
	Stranger Things	2016	8.7	1374734
	The Walking Dead	2010	8.1	1103593
2020	The Queen's Gambit	2020	8.5	575649
	Squid Game	2021	8.0	565523
	Loki	2021	8.2	422663

A tabela 9 acima exibe as top três séries mais populares por década. É evidente que, ao comparar a quantidade de votos (popularidade), as séries das últimas décadas (2000 a 2020) são as mais populares.

Se destacam aqui as mais populares dos anos 2010, **Game of Thrones**, **Stranger Things** e **The Walking Dead**, em que todas tiveram mais de um milhão de votos - **Game of Thrones** ainda se destaca entre os três, com uma popularidade que atingiu mais de 2,34 milhões de votos, sendo uma produção audiovisual de extrema importância e que marcou essa década, revolucionando a indústria audiovisual e que se assegurou como um divisor de águas na história recente da televisão.**Friends**, de 1995, e **Breaking Bad**, de 2008 são as mais antigas da lista que atingiram mais de um milhão de votos, e **Breaking Bad** leva mais um destaque por ser a série com a maior média de avaliações, sendo sua nota **9,5**. **I Love Lucy** é a série mais antiga que aparece na lista, de 1952, enquanto as mais recentes são **Squid Game** (conhecida no Brasil como "*Round Six*") e **Loki**, ambas de 2021.

Tabela 10: Top 3 Filmes Mais Populares por Década

Década	Nome do Filme	Ano de Lançamento	Média das Avaliações	Quantidade de Votos
1890	The Corbett-Fitzsimmons Fight	1897	5.2	528
	Miss Jerry	1894	5.4	212
	Jeffries-Sharkey Contest	1899	3.9	78
1900	The Story of the Kelly Gang	1906	6.0	928
	The Life and Passion of Jesus Christ	1903	6.5	700
	Westinghouse Works	1904	5.3	360
1910	The Birth of a Nation	1915	6.1	26626
	Intolerance	1916	7.7	16946
	Broken Blossoms	1919	7.2	11203
1920	Metropolis	1927	8.3	187954
	The Kid	1921	8.2	136720
	The Gold Rush	1925	8.1	120062
1930	The Wizard of Oz	1939	8.1	434076
	Gone with the Wind	1939	8.2	337952
	Modern Times	1936	8.5	262930
1940	Casablanca	1942	8.5	613091
	It's a Wonderful Life	1946	8.6	505051
	Citizen Kane	1941	8.3	470622
1950	12 Angry Men	1957	9.0	882392
	Rear Window	1954	8.5	529760
	Vertigo	1958	8.3	433433
1960	The Good, the Bad and the Ugly	1966	8.8	822930
	2001: A Space Odyssey	1968	8.3	729877
	Psycho	1960	8.5	728902
1970	The Godfather	1972	9.2	2047441
	Star Wars: Episode IV - A New Hope	1977	8.6	1469127
	The Godfather Part II	1974	9.0	1384296
1980	Star Wars: Episode V - The Empire Strikes Back	1980	8.7	1399518
	Back to the Future	1985	8.5	1330462
	Star Wars: Episode VI - Return of the Jedi	1983	8.3	1136148
1990	The Shawshank Redemption	1994	9.3	2937574
	Fight Club	1999	8.8	2369019
	Forrest Gump	1994	8.8	2297181
2000	The Dark Knight	2008	9.0	2917884
	The Lord of the Rings: The Fellowship of the Ring	2001	8.9	2038846
	The Lord of the Rings: The Return of the King	2003	9.0	2009855
2010	Inception	2010	8.8	2590880
	Interstellar	2014	8.7	2154744
	The Dark Knight Rises	2012	8.4	1854052
2020	Spider-Man: No Way Home	2021	8.2	903750
	Dune	2021	8.0	901216
	The Batman	2022	7.8	810322

A tabela 10 exibe os top três filmes mais populares por década registrados. É interessante de notar que há filmes que datam desde o século XIX, em contraste com as séries, que só começaram a surgir a partir dos anos 1950, como notado na tabela 9, acima.

Os filmes mais populares da lista se concentram nos anos de 1970 a 2010, com todos eles acima dos um milhão de votos. É também perceptível o legado de **Christopher Nolan** na história do cinema, uma vez que os três filmes mais populares dos anos 2010 foram dirigidos por ele, como notado em Profissionais mais Populares por Área. Percebe-se também que franquias famosas são capazes de capturar toda a popularidade de uma década, como ocorreu nas décadas de 1970, 1980 e 2000 - em todas elas, as franquias **The Godfather** (O Poderoso Chefão), **Star Wars** e **The Lord of the Rings** (O Senhor dos Anéis) trouxeram dois de seus títulos como os mais populares de suas respectivas décadas.

O filme mais popular e mais bem avaliado de todos os tempos é **The Shawshank Redemption**, de 1994, com quase 3 milhões de votos e nota **9.3** - já o menos popular e pior avaliado é **Jeffries-Sharkey Contest**, de 1899 e com apenas 78 votos e nota **3.9**.

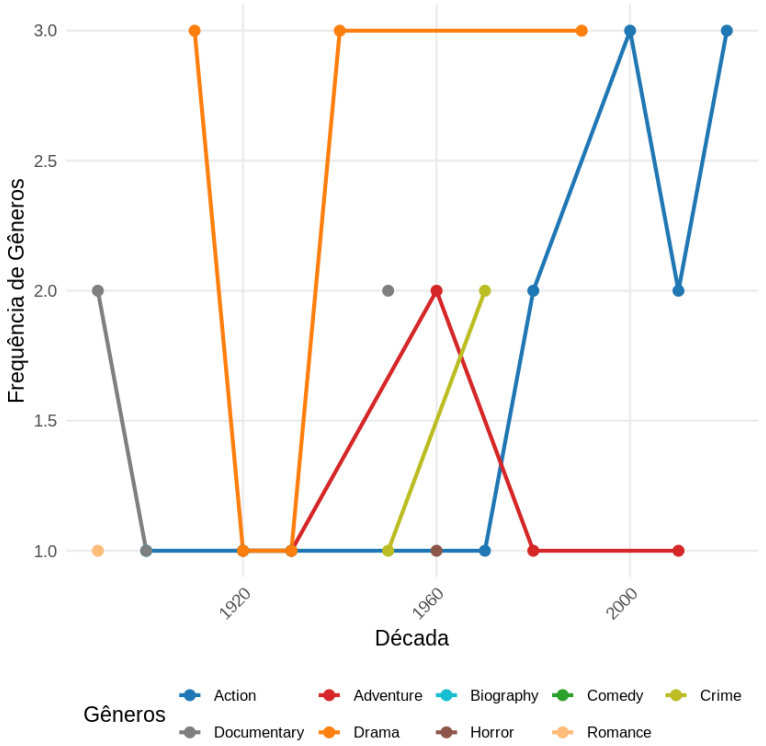

Figura 10: Evolução da Frequência de Gêneros por Década nos Top 3 Filmes Mais Populares

Com base no gráfico 10, podemos identificar as seguintes tendências ao longo das décadas:

- O gênero **Drama** mostrou-se consistentemente popular em quase todas as décadas.
- Gêneros como **Ação** e **Aventura** começaram a ganhar destaque a partir das décadas mais recentes, refletindo mudanças nas preferências do público e avanços nas técnicas de produção cinematográfica.
- No início da produção cinematográfica, os gêneros predominantes eram **Documentário** e **Romance**, possivelmente devido às limitações técnicas da época e ao foco em histórias mais simples e narrativas baseadas na realidade.

Observação: Para a construção deste gráfico, foi considerado apenas o *primeiro gênero* listado para cada filme, uma vez que muitos filmes possuem múltiplos gêneros, como no exemplo *Documentary, News, Sport.*

Tabela 11: Top 3 Episódios Mais Populares por Década

Década	Nome da Série	Nome do Episódio	Média das Avaliações	Quantidade de Votos
1940	The Lone Ranger	Enter the Lone Ranger	7.7	289
		The Lone Ranger Fights On	7.8	199
		The Lone Ranger's Triumph	7.6	169
1950	The Twilight Zone	Time Enough at Last	8.9	8007
		Where Is Everybody?	7.9	7375
		Walking Distance	8.2	6257
1960	Star Trek	The Cage	7.6	7321
		The City on the Edge of Forever	9.2	6753
		Where No Man Has Gone Before	7.7	6584
1970	Columbo	Murder by the Book	7.7	5401
		Death Lends a Hand	7.7	3888
		Any Old Port in a Storm	8.3	3868
1980	Twin Peaks	Northwest Passage	8.9	17229
	The Simpsons	Simpsons Roasting on an Open Fire	8.1	9056
	Dekalog	Dekalog, jeden	8.5	7663
1990	One Piece	I'm Luffy! The Man Who Will Become the Pirate King!	8.4	28332
	Friends	The One Where Everybody Finds Out	9.7	13464
	One Piece	The Great Swordsman Appears! Pirate Hunter, Roronoa Zoro	8.3	13388
2000	Breaking Bad	Pilot	9.0	46309
	One Piece	Nakama no Itami wa Waga Itami: Zoro Kesshi no Tatakai	9.7	39247
	Breaking Bad	Crazy Handful of Nothin'	9.3	36225
2010	Game of Thrones	The Iron Throne	4.0	266481
		Battle of the Bastards	9.9	231396
		The Long Night	7.5	228163
2020	The Last of Us	Long, Long Time	8.1	227226
	The Last of Us	When You're Lost in the Darkness	9.1	111453
	Attack on Titan	Assault	9.7	104863

A tabela 11 acima mostra os episódios mais populares por década. É interessante observar que das nove décadas presentes na lista, em cinco delas os episódios mais populares pertenciam a uma única série - 1940 a **The Lone Ranger**, 1950 a **The Twilight Zone**, 1960 a **Star Trek**, 1970 a **Columbo** e 2010 a **Game of Thrones**. Novamente, como notado na análise das Top 3 Séries Mais Populares por Década, os episódios mais populares da história pertencem à série mais popular da história, **Game of Thrones**, com todos acima dos 228 mil votos cada - o destaque vai para o episódio final da série, **The Iron Throne**, com mais de 266 mil votos e uma avaliação média de **4.0**.

Apesar da lista conter os episódios mais populares por década, o episódio menos popular dentre esses é **The Lone Ranger's Triumph** (1940), com apens 169 votos - por outro lado, todos os episódios mais populares dos anos 2020 ultrapassaram os 100 mil votos, com destaque a dois pertencentes a **The Last of Us**, **When You're Lost in the Darkness** e **Long, Long Time**. A série com o episódio mais popular é **One Piece**, da década de 2000, com uma avaliação média de **9.7**.

4 Conclusão

Após as análises, as três perguntas por fim foram respondidas. Através da manipulação do banco de dados, e usando a quantidade de votos como critério de popularidade, identificamos as produções mais populares de cada década, e os profissionais mais populares de cada setor. Também mostramos, através de uma análise de regressão linear, a relação entre número de temporadas e a qualidade das séries.

Apesar de que, conforme avançam as temporadas de uma série elas tendem a ser pior avalidas, séries com mais quantidade de temporadas são melhores avaliadas. Percebe-se também uma tendêcia na redução na quantidade de temporadas e episódios nos anos mais recentes. Além disso, nota-se que quanto mais popular a série, maior é o impacto que o número de temporadas exerce na sua nota média.

Os três filmes mais populares, dentre os filmes mais populares de todas as décadas, são dos anos 1990, sendo eles **The Shawshank Redemption**, **Fight Club** e **Forrest Gump**. Os três episódios mais populares, dentre os episódios mais populares de todas as décadas, pertencem à década de 2010, sendo eles três episódios da série Game of Thrones, **The Iron Throne**, **Battle of the Bastards** e **The Long Night**. As três séries mais populares são dos anos 2000 e 2010, sendo **Breaking Bad**, **Game of Thrones** e **Stanger Things**.

Dentre os profissionais mais populares, se destaca **Christopher Nolan** por ser o diretor e roteirista mais famoso. Já o compositor mais popular é **Hans Zimmer**, que é o que mais possui projetos creditados e compôs a trilha de **The Dark Knight**, que é o projeto mais popular dentre os profissionais mais famosos.

5 Anexo dos códigos

Banco unificado em SQLite

Código das seções 2.2.1, 2.4.1 e 2.4.3

Código das seções 2.3.1, 2.3.2, 2.4.2, 3.1, 3.3 e 3.4

6 Bibliografia

Referências

[1] Universidade Estadual de Campinas (UNICAMP), Material das aulas - Benilton. Disponível em: https://me315-unicamp.github.io/material/.

[2] IMDb Datasets, Conjunto de dados do IMDb. Disponível em: https://datasets.imdbws.com/.

[3] IMDb Developers, Descrição dos arquivos do banco de dados IMDb. Disponível em: https://developer.imdb.com/non-commercial-datasets/.

[4] DataCamp, SQLite in R: Working with Databases in R. Disponível em: https://www.datacamp.com/tutorial/sqlite-in-r.

[5] Polars, A fast DataFrame library for Rust and Python. Disponível em: https://pola.rs/.

[6] Python Software Foundation, SQLite in Python - Official Documentation. Disponível em: https://docs.python.org/3/library/sqlite3.html.