UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciência Exatas
Programa de Pós-graduação em Estatística

Rodrigo do Vale Andrade

ENTRETENIMENTOS: Estudos da Base de Dados Audiovisuais

Belo Horizonte / MG

Julho / 2020

Rodrigo do Vale Andrade

ENTRETENIMENTOS: Estudos da Base de Dados Audiovisuais

Monografia de especialização apresentada à Faculdade de Ciências Exatas Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Especialista em Estatística.

Orientadora: Prof.ª. Lourdes C. Contreras Montenegro

Belo Horizonte / MG

Julho / 2020

Ficha catalográfica elaborada pela bibliotecária Belkiz Inez Rezende Costa CRB 6ª Região nº 1510

Andrade, Rodrigo do Vale

A553e Entretenimentos: estudos da base de dados audiovisuais / Rodrigo do Vale Andrade.— Belo Horizonte, 2020.
59 f. il.; 29 cm.

Especialização (mestrado) - Universidade Federal de Minas Gerais – Departamento de Estatística.
Orientadora: Lourdes Coral Contreras Montenegro.

1. Estatística. 2. Amostragem (Estatística). 3. Inferência (Lógica). 4. Linguagem de programação (Computadores). 5. População -Estatistica. I. Orientadora. II. Título.

CDU 519.6 (043)

Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística
Programa de Pós-Graduação / Especialização
Av. Pres. Antônio Carlos, 6627 - Pampulha
31270-901 – Belo Horizonte – MG

E-mail: pgest@ufmg.br
Tel: 3409-5923 – FAX: 3409-5924

ATA DO 212º. TRABALHO DE FIM DE CURSO DE ESPECIALIZAÇÃO EM ESTATÍSTICA DE RODRIGO DO VALE ANDRADE.

Aos trinta e um dias do mês de julho de 2020, às 09:15 horas, com utilização de recursos de videoconferência a distância, reuniram-se os professores abaixo relacionados, formando a Comissão Examinadora homologada pela Comissão do Curso de Especialização em Estatística, para julgar a apresentação do trabalho de fim de curso do aluno **Rodrigo do Vale Andrade**, intitulado: "Entretenimento: Estudo da Base de Dados Audiovisuais.", como requisito para obtenção do Grau de Especialista em Estatística. Abrindo a sessão, a Presidente da Comissão, Professora Lourdes Coral Contreras Montenegro – Orientadora, após dar conhecimento aos presentes do teor das normas regulamentares, passou a palavra ao candidato para apresentação de seu trabalho. Seguiu-se a arguição pelos examinadores com a respectiva defesa do candidato. Após a defesa, os membros da banca examinadora reuniram-se sem a presença do candidato e do público, para julgamento e expedição do resultado final. Foi atribuída a seguinte indicação: o candidato foi considerado Aprovado condicional às modificações sugeridas pela banca examinadora no prazo de 30 dias a partir da data de hoje por unanimidade. O resultado final foi comunicado publicamente ao candidato pela Presidente da Comissão. Nada mais havendo a tratar, a Presidente encerrou a reunião e lavrou a presente Ata, que será assinada por todos os membros participantes da banca examinadora. Belo Horizonte, 31 de julho de 2020.

Profa. Lourdes Coral Contreras Montenegro (Orientadora)
Departamento de Estatística / UFMG

Profa. Edna Afonso Reis
Departamento de Estatística/UFMG

Prof. Ilka Afonso Reis
Departamento de Estatística/UFMG

Rodrigo do Vale Andrade

ENTRETENIMENTOS: Estudos da Base de Dados Audiovisuais

Monografia de especialização apresentada à Faculdade de Ciências Exatas Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Especialista em Estatística.

Banca examinadora:

_____
Dra. Lourdes Coral Contreras Montenegro - UFMG (Orientadora)
Julgamento: _____

_____
Dra. Edna Afonso Reis – UFMG (Banca examinadora)
Julgamento: _____

_____
Dra. Ilka Afonso Reis – UFMG (Banca examinadora)
Julgamento: _____

Belo Horizonte, 31 de julho de 2020

RESUMO

Com o grande aumento de provedores de filmes e series de televisão, como Netflix, Amazon Prime, Google Play, HBO Go, Net Movies, Itunes Store e Microsoft Store, fora os tradicionais estúdios de cinema, é de grande interesse em descobrir os gostos mais populares das pessoas por estes tipos de entretenimentos e com isso ajuda-los na escolha baseados em suas características, como por exemplo, tipo, gênero, duração, ano, etc. Sendo assim, é importante conhecer e analisar um conjunto de dados que represente os mais diversos tipos de entretenimentos audiovisuais, televisivos e/ou cinematográficos, e suas variáveis. Desta forma, o objetivo deste trabalho é colocar em prática o aprendizado adquirido durante a pós-graduação em estatística na UFMG, na exploração, organização e seleção de dados, bem como fazer uma coleta de uma amostra para desenvolvimento dos estudos estatísticos, aprendidos durante o curso e com isso fazer análises críticas das informações quantitativas e qualitativas, interpretando adequadamente este tipo de dados. As etapas foram desde a extração e transformação dos dados, unificando-os em um único arquivo, uso da metodologia de amostragem para determinar o tamanho suficientemente grande de uma amostra e descrever os métodos estatísticos de regressão linear ou não linear que represente bem a variável de interesse (média de classificação). Porém, o propósito inicial não ofereceu bons ajustes em seus modelos com todas as suposições violadas. Logo, partimos para o estudo de variáveis qualitativas, mudando o foco do estudo para uma análise de variância (ANOVA) paramétrica e não paramétrica com o intuito de testar, se o tipo de filme tem alguma relação com o gênero. Além disso, uma análise de Kruskall Wallis foi utilizada para comparar os níveis das variáveis, gênero e tipo, em sua forma particular.

Palavras-chave: Amostra. Inferência. Linguagem de programação. População.

Abstract

With the huge increase in providers of films and television series, such as Netflix, Amazon Prime, Google Play, HBO Go, Net Movies, Itunes Store and Microsoft Store, apart from the traditional movie studios, it is of great interest to discover the most popular tastes. popular with people for this type of entertainment and thereby help them choose based on their characteristics such as type, gender, duration, year, etc. Therefore, it is important to know and analyze a set of data that represents the most diverse types of audiovisual, television and / or cinematographic entertainment, and their variables. Thus, the objective of this work is to put into practice the learning acquired during the postgraduate course in statistics at UFMG, in the exploration, organization and selection of data, as well as to collect a sample for the development of statistical studies learned during the course and thereby making critical analyzes of quantitative and qualitative information, properly interpreting this type of data. The steps were from the extraction and transformation of the data, unifying them in a single file, using the sampling methodology to determine the sufficiently large size of a sample and describing the statistical methods of linear or non-linear regression that well represents the variable of interest (average rating). However, the initial purpose did not offer good adjustments to its models with all the assumptions violated. Thus, we started to study qualitative variables, changing the focus of the study to a parametric and non-parametric analysis of variance (ANOVA) in order to test whether the type of film has any relation with the genre. In addition, a Kruskall Wallis analysis was used to compare the levels of the gender and type variables in their particular form.

Keywords: Inference. Population. Programming language. Sample.

SUMÁRIO

1 – INTRODUÇÃO

Com o grande avanço das plataformas de streaming, que marcou uma das evoluções tecnológicas mais destacadas dos últimos 10 anos, conhecer a preferência dos usuários é fundamental para oferecer a eles os melhores entretenimentos. Desta forma, é crucial fazer estudos nessas bases de dados do site IMDb.

Infelizmente, a maioria das plataformas não liberam suas bases para estudos. Em contrapartida, o site IMDb (Internet Movie Database) (IMDb, 2019) disponibiliza uma série de arquivos referentes aos gostos populares dos mais diversos tipos de entretenimentos. Tanto é que, foram elaborados alguns estudos com esta fonte de dados, como por exemplo: Sarkar (2020) e Huang (2020). Estudos estes que envolvem usar os dados para outros fins.

Então, com o intuito de estudar a base de dados do site, foram extraídos 7 (sete) arquivos no formato TSV[1] que representam dados online, contendo informações sobre música, cinema, filmes, programas de televisão, comerciais para televisão e jogos de computadores. Destes acervos foram retirados e concatenados somente os campos coerentes em nossa análise, formando assim um único arquivo para nosso estudo.

O tamanho do arquivo, população, até o presente momento era de 911.026 títulos de conteúdos votados, que são entretenimentos produzidos a partir de 1960. Com esta base, já concatenada e em um único arquivo, foram efetuados métodos de amostragem para retirada de uma amostra, já que a população é muito grande para alguns cálculos estatísticos na ferramenta escolhida, sendo assim precisaria de máquinas e softwares potentes para o processamento massivo destes dados.

O presente trabalho tem como objetivos explicar a média de avaliação dos títulos, segundo o tipo de título (curta, filme, seriado e programas de tv), separadamente, assim como, comparar a média de avaliação dos tipos de títulos e gêneros (ação, adulto, animação, aventura, biografia, comedia, crime, curtas, documentário, drama, esportes, família, fantasia, faroeste, ficção cientifica, guerra, musical, notícias, reality, romance, suspense, terror).

Utilizando Amostragem Aleatória Simples, sem reposição associada a Amostragem Estratificada Proporcional (PINHEIRO, 2017, p. 12), foi possível coletar dados que, posteriormente, foram analisados por meio de técnicas de estatísticas descritivas, para representar o conjunto de dados amostral e com isso supor um modelo de regressão linear ideal para nossa análise. Neste caso são os modelos lineares

[1] Fonte de dados salvas em arquivos tipo texto separados por tabulações.

generalizados (MLG) (COSTA, 2019. p.71), em situações em que as hipóteses do modelo de regressão linear não foram atendidas.

Como não obtivemos sucessos nos modelos acima descritos, utilizamos a técnica de Análise de Variância (ANOVA) paramétrica (RUMSEY, 2011, p. 207) para as variáveis qualitativas (gênero e tipo), com o propósito de encontrar interação entre eles.

O objetivo inicial, seria encontrar um modelo que representasse uma análise clara dos melhores títulos votados pelos usuários, porém não foi obtido um resultado coerente. A outra alternativa, variância paramétrica, também não conseguiu interação entre as variáveis gênero e tipo. Como o último recurso, utilizamos, testes não parametrizados, por exemplo, teste de Kruskal-Wallis (Teste de Kruskal-Wallis, 2020), que compara a mediana das avaliações entre os níveis da variável gênero e tipo, individualmente.

2 – OBJETIVO GERAL

Estudar a avaliação dos usuários de plataforma de streaming de entretenimento, em função do tipo e do gênero dos títulos avaliados, utilizando uma amostra dos dados disponíveis no site IMDb.

2.1 – Objetivos Específicos

1. Explorar a base de dados do site IMDb e selecionar somente os campos necessários para o trabalho*;*
2. Trabalhar os dados com a ferramenta Python para gerar um arquivo final dos campos selecionados acima;
3. Fazer os cálculos para apurar o tamanho ideal de uma amostra;
4. Gerar os arquivos de amostras;
5. Encontrar modelos que apresentem pelo menos uma variável significativa, para pelo menos uma das distribuições avaliadas, preservando a parcimônia;
6. Ajustar e testar modelos estatísticos para análises, utilizando a ferramenta R.

3 – METODOLOGIA

3.1 - Descrição dos Dados

3.1.1 - Descrição do Site

IMDb é conhecida como Internet Movie Database, uma base de dados online de informações sobre música, cinema, filmes, programas, comerciais de televisão e jogos de computador, pertencente a *Amazon.com*[2].

Qualquer usuário adulto cadastrado pode acessar ao site *https://m.imdb.com*. Ele contém uma base de dados com mais de 9 milhões de registro sobre os temas acima citados. São mais 900 mil títulos de filmes, programas de tv, jogos, comerciais e series classificadas e qualificadas. Nele, também encontram todas as características de um entretenimento. O objetivo é ajudar o usuário a procurar um título com a maior pontuação para assistir.

A base de dados está no link *https://www.imdb.com/interfaces* e pode ser baixado os 7 (sete) arquivos que estão detalhados abaixo.

3.1.2 – Os Arquivos

São 7 arquivos no formato TSV, que segundo site reviversoft.com (2019):

> "...TSV significa "Valores separados por tabulações", e estes arquivos valores guia separados são criados e usados por muitos aplicativos de planilha. O conteúdo de um arquivo de valores separados guia podem incluir textos, dados matemáticos, científicos ou estatísticos separados em linhas e colunas...".

Que representam:

Os arquivos que foram utilizados na análise são: título.akas, que contém 8 (oito) campos representados pelos títulos, conforme a origem das produções. O segundo é o título.básico, que representa o título de acordo com a categoria, tempo e ano de produção. O terceiro é, título.episódio, que contém a quantidade de episódios e temporadas para cada título. E o último arquivo é o título.classificação, que contém a classificação da IMDb e as quantidades de votos de todos os títulos.

Todos os quatro acervos contêm um uma chave de identificação (código, ex.: tt0000001), que identifica cada título votado. Por meio dela, foram efetuadas as ligações entre os arquivos mencionados anteriormente e concatenados, produzindo um único arquivo, que é explicado no próximo item.

Os outros arquivos são: nome.básico contendo 6 (seis) colunas, que são os nomes dos atores, atrizes e produções relacionados. O segundo arquivo é título.equipe,

[2] Amazon.com, Inc. é uma empresa transnacional de comércio electrónico dos Estados Unidos fundada por Jeff Bezos em julho de 1994 com sede em Seattle, estado de Washington.

que contém informações do diretores e escritores das produções associados aos códigos dos títulos. E o terceiro arquivo é o título.principal que traz informações dos elencos e das equipes associados a cada produção. Estes três, não foram utilizados em nossa análise, pois não tem representatividade em nosso estudo.

3.1.3 – Arquivo Final

Esta etapa, é a junção dos arquivos título.akas, título.classificação, título.equipe e titulo.básico pelo campo chave de identificação tconst, comum em todos os arquivos. O campo tconst representa linha a linha, de cada título votado pelos internautas.

A tabela 1, apresenta a concatenação dos registros, que foi feita através da ferramenta Python, ligando as linhas comuns dos quatro arquivos e gera uma nova linha, dentro de uma nova tabela, contendo as seguintes colunas: média classificação, número de votos, classificação de tipo de título, ano de produção, é adulto, minutos de duração, número de temporadas e episódios, gênero e originalidade. E desta tabela foi extraída uma amostra para os estudos.

Tabela 1: Recorte do arquivo final.

Média Classificação	Número de Votos	Classificação Tipo de Títulos	Ano de Produção	É Adulto	Minutos de Duração	Número de Temporadas	Número de Episódios	Gênero	Originalidade
8.4	8	1	1963	0	20	1	10	Aventura	0
5.6	10	1	2015	0	5	1	8	Animação	0
8.9	8	1	1947	1	10	4	11	Drama	0
7.2	30	1	2019	1	9	3	8	Comédia	0
7.7	34	1	1998	0	18	2	13	Curta	0
5.8	66	1	2013	0	3	2	21	Comédia	0
7.9	15	1	2008	0	16	1	25	Biografia	1
7.0	13	1	1950	0	13	1	5	Notícias	0
5.9	19	1	1972	1	12	3	7	Terror	0
7.4	37	1	2014	1	7	5	11	Animação	0
6.8	8	1	2002	1	5	11	10	Animação	1
5.9	92	1	2018	1	1	2	12	Documentário	0
5.4	17	1	1953	1	6	3	10	Comédia	0
9.3	6	1	1996	0	9	1	5	Ação	0
6.1	91	1	2001	1	24	6	5	Ficção Científica	0
8.2	8	1	2009	0	5	6	6	Curta	0
...	...	...	...	...	...	...	...	...	...
5.2	6	1	2004	1	15	2	4	Terror	0
Total de Registros									**911.026**

A média de classificação representa a média do total dos números de votos de cada internauta adulto. Ela pode variar de zero até o valor máximo dez. Já o número de votos, é o total geral das votações computadas de cada indivíduo adulto, cadastrados no site da IMDb. A classificação tipo de título, foi uma decodificação dos itens da coluna tipo de título, convertidos em valores inteiros e classificados em: curta = 1; filme = 2; tv filme = 3; tv séries = 4; tv episódios = 5; tv curta = 6; tv mini séries = 7; tv especial = 8; vídeo = 9; vídeo game = 10. Esta classificação representa o tipo de entretenimento de cada produção.

O campo ano de produção é, o ano em que cada entretenimento foi produzido. A classificação adulta é, um método de classificação que indica diretamente qual é a idade mínima recomendada, neste caso os entretenimentos são de classificação livre (0 – não é adulto) ou impróprio para menores (1 – conteúdo adulto). O conteúdo adulto pode conter cenas fortes, como violência, sexo ou imagens impróprias.

Minutos de duração é, o tempo em minutos que cada entretenimento gasta para ser assistido. No caso de séries existem o número de temporadas e episódios que foram produzidos. Originalidade é a variável que indica se o título é uma produção original (1 - original) ou não (0 - não original). E por último, o campo gênero que é uma classificação utilizada para distinguir os variados tipos de produções como: ação, suspense, drama, terror, aventura, dentre outros.

No arquivo final foram excluídos os valores dos campos do tipo de título, igual a vídeo game e gêneros sem descrição, já que eles não fazem parte da análise.

As variáveis a serem analisadas, são:

- Média classificação: variável quantitativa contínua, que representa a classificação média da votação dos usuários;
- Número de Votos: variável quantitativa discreta, que representa a quantidade de usuários que votaram;
- Classificação tipo de título: variável qualitativa, que representa a categoria dos títulos;
- É adulto: variável qualitativa nominal, indicador 1 ou 0 que representa a classificação dos títulos adulto ou não adulto;
- Minutos de duração: variável quantitativa discreto que representa a o minuto de cada entretenimento;
- Ano de produção: variável quantitativa discreto que representa o ano de produção dos entretenimentos;
- Originalidade: variável nominal, indicador 1 ou 0, que representa a classificação do título quando a originalidade;
- Número de temporadas e episódios: variável quantitativa discreto que representa a quantidade de temporadas e episódios de cada série;
- Gênero: variável qualitativa, que representa a classificação em relação aos gêneros dos títulos.

Para melhorar o critério de nossa análise descritiva, a amostra foi dividida em quatro categorias com os tipos de títulos semelhantes, que são eles:

Filmes: O filme é uma obra de arte realizada por meio da sucessão de imagens em vídeo e com som, voltado para exibição nos cinemas. Sua característica é um conteúdo único sem episódios e temporadas;

Curtas: Engloba os títulos do tipo curtas e curtas de tv (filmes com duração de até 30 minutos, de intenção estética, informativa, educacional ou publicitária. Geralmente exibido como complemento de um programa cinematográfico) e vídeos (curtas com produção independente);

Programas de tv: Engloba os títulos do tipo filmes para tv (é o termo utilizado na indústria cinematográfica para definir o conceito de "filme produzido para ter a sua estreia na televisão", ao contrário da maioria dos filmes, que estreiam em salas de cinema), especiais de tv (é um programa de televisão stand-alone. Disponível através do meio televisivo -noticiário, drama, comédia, variedade, cultura), em vários formatos (ao vivo, documentário, produção de estúdio, animação, filme)), e mini séries de tv (Seriado de televisão, de cunho ficcional ou documentário, exibido em um número reduzido de capítulos). Em nossa base de análise não constam para estes títulos números de temporadas e episódios;

Seriados: Seriados de tv e episódios de tv (séries de televisão, séries televisivas, séries de tv ou telessérie. É um tipo de programa televisivo ou programa online com um número pré-definido temporadas e de capítulos ou episódios).

O próximo passo é realizar um estudo para identificar qual melhor método de amostragem a ser aplicada e por meio dele calcular o tamanho de amostra a ser extraída.

3.2 - Amostragem

Segundo (PINHEIRO, 2017, p. 12) “Amostragem é o processo ou técnica de retirar uma seleção de elementos em um conjunto de dados ou universo chamado de população”. O processo resulta em metodologias que selecionam um subconjunto de dados denominado amostra, com a finalidade de conhecermos melhor as características da população.

Esta seleção utilizou o método mais adequado, de tal forma que os resultados das amostras sejam claros para avaliar as características de toda a população.

Os motivos para usar o processo de amostragem são: economia, tempo, confiabilidade dos dados e operacionalidade, ou seja:

- É mais econômico fazer o levantamento de uma parte da população;
- Fica mais rápido a pesquisa;
- Em um número reduzido de elementos pode-se dar mais atenção aos casos individuais, evitando erros nas respostas;
- E por último, é mais fácil realizar operações em pequena escala.

Ao contrário, os dificultadores para não analisar a população são:

- A necessidade de um hardware e um software potentes;

- O pacote utilizado na ferramenta R tem um limite para alguns cálculos, como por exemplo o teste de normalidade, que trabalha com no máximo 5 (cinco) mil registros;
- A agilidade e a precisão para fazer inferências com volumes de dados menores.

Neste trabalho usamos a amostragem probabilística, os quais foram estudados os processos Amostragem Aleatória Simples (AAS) com e sem reposição, Amostragem Aleatória Sistemática (AAS) e Amostragem Aleatória Estratificada (AAE), já que elas se adequam mais ao tipo de trabalho que foi descrito.

3.2.1 – Amostragem Aleatória Simples (AAS)

Amostragem aleatória simples (AAS) é uma técnica para retirar um subconjunto de elementos “amostra”, selecionados totalmente ao acaso a partir de um conjunto maior “população”, por um processo que garanta que todos os indivíduos da população tenham a mesma probabilidade de serem escolhidos. Este é o método mais simples para a seleção de uma amostra. Representado pelo tamanho *n* (número total de itens dentro do subconjunto), desenhado a partir de uma população de tamanho N. Este tipo de amostragem consiste em selecionar os itens por meio de um sorteio sem restrição.

A figura 1 mostra um exemplo de uma AAS, que é representada a partir do método de retirada de itens aleatoriamente.

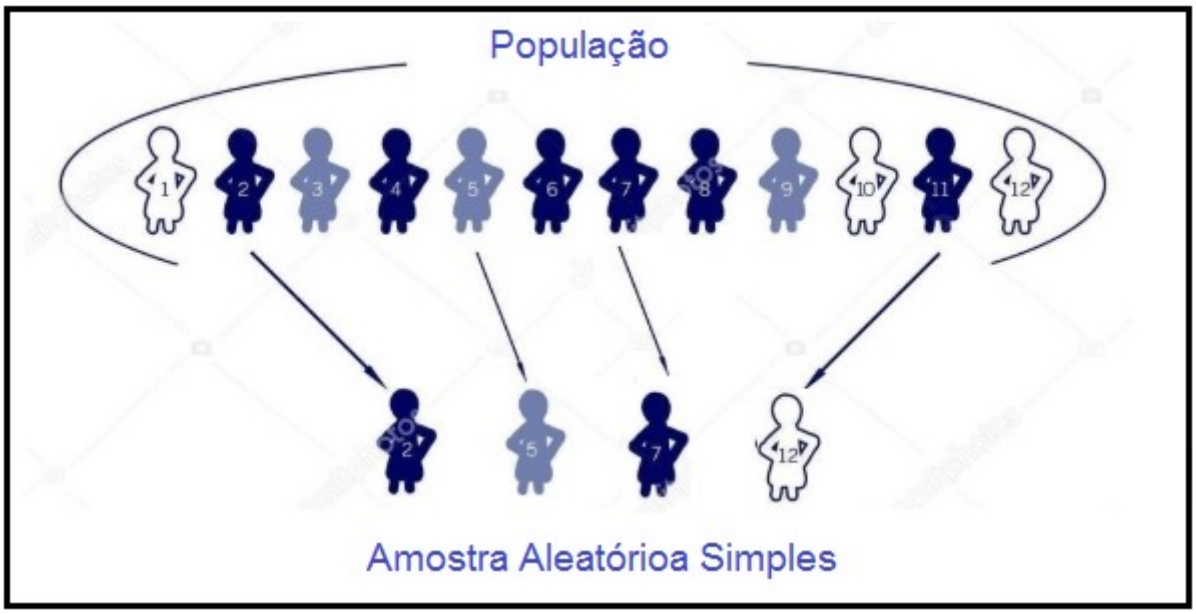


Figura 1: Exemplo de AAS.

O processo utiliza ou não de um software estatístico ou planilha de cálculos para gerar números aleatórios, sendo retirados da população, cujos valores correspondem aos números gerados pelo programa.

A amostragem aleatória simples, pode ser feita com ou sem reposição. Se a população for infinita ou muito grande, as retiradas com e sem reposição serão equivalentes. Se, no entanto, a população for finita ou pequena, é necessário fazer uma distinção entre os dois procedimentos, pois na extração com reposição as diversas retiradas se tornam independentes, mas no processo sem reposição haverá dependência entre as retiradas.

O processo da AAS com reposição (AASc) seguiu os seguintes passos:

- Seleciona uma unidade da população com equiprobabilidade;
- Reponha a unidade selecionada na população;
- Repita os Passos 1 e 2 até ter feito *n* seleções.

Para uma variável de interesse *y*, temos que os valores amostrais $x_1, x_2, \ldots, x_n$ se tornam independentes e identicamente distribuídos. Desta forma, usaremos a equação abaixo para determinar o tamanho da amostra:

$$n = \left(\frac{\sigma}{B/z_\alpha} \right)^2 = \frac{\sigma^2}{D} \text{ , tal que D = } \frac{B^2}{z_\alpha^2}$$

$\boldsymbol{\sigma^2}$: Variabilidade estimada da população, que é determinada por uma amostra piloto para estabelecer um estimador aproximada a sigma;

B: Erro máximo fixado;

$\boldsymbol{z_\alpha}$: Grau de confiança;

Já no processo da AAS sem reposição (AASs), opera de modo idêntico como da reposição, alterando-se apenas o segundo passo, sem devolver o elemento a população.
Logo, usaremos a equação abaixo para determinar o tamanho da amostra:

$$n = \left(\frac{S^2/D}{1+\frac{S^2/D}{N}} \right) = \frac{1}{D/S^2 + 1/N} \text{ , tal que D = } \frac{B^2}{z_\alpha^2}$$

$\boldsymbol{S^2}$: Variabilidade da população;
N: Tamanho da população;
B: Erro máximo fixado;
$\boldsymbol{z_\alpha}$: Grau de confiança.

Após os cálculos, teremos dois tamanhos de amostras com características e técnicas distintas. É importante descobrir qual é o melhor método a ser utilizado. Em geral, o critério mais adotado em amostragem é o Erro Quadrático, quando os estimadores não são viesados. Entretanto, Kish (1965) propôs uma medida chamado de efeito do planejamento amostral EPA (design effect, deff).

EPA foi desenvolvido por Kish (1965), que é uma técnica para comparar ganhos ou perda de precisão, sob diferentes planos amostrais no estágio de planejamento da pesquisa. Ele equivale a razão entre a variância de um estimador verdadeiro e a variância do estimador induzida pelo plano de amostragem aleatória simples. Ou seja, para um valor de $\overline{y}$ temos o cálculo da variância entre valor *Var AASs* ($\overline{y}$) pela *Var AASc* ($\overline{y}$). Essa razão é representada pela seguinte expressão:

$$EPA = \frac{Var_{\mathrm{AASs}}[\overline{y}]}{Var_{\mathrm{AASc}}[\overline{y}]} = \frac{(1-f)S^2/n}{\sigma^2/n} = \frac{N-n}{N-1}$$

Os valores do EPA conduzem a importância de se considerar o verdadeiro plano amostral ao se estimar as variâncias associadas às estimativas dos parâmetros.

A partir dos valores encontrados podem-se tirar as conclusões:

EPA < 1 variância sob AAS superestimada;
EPA = 1 não há diferença entre as estimativas de variância;
EPA > 1 variância sob AAS subestimada.

3.2.2 – Amostragem Aleatória Sistemática (AAS)

É um processo de amostragem, no qual o critério de probabilidade é estabelecido na escolha do primeiro item de forma aleatória (seguindo a AASs) dentro dos elementos ordenados da população e posteriormente seleciona um intervalo de amostragem *K,* elementos para retirada do resto da amostra. O propósito é cobrir a população em toda sua extensão a fim de obter um modelo sistemático simples e uniforme.

A amostra aleatória sistemática é um método de amostragem muito simples e que só requer a seleção do primeiro indivíduo aleatório, pois o restante foi feito pela divisão (*K* = *N/n*), para achar os próximos elementos a serem retirados, como mostra a figura 2:

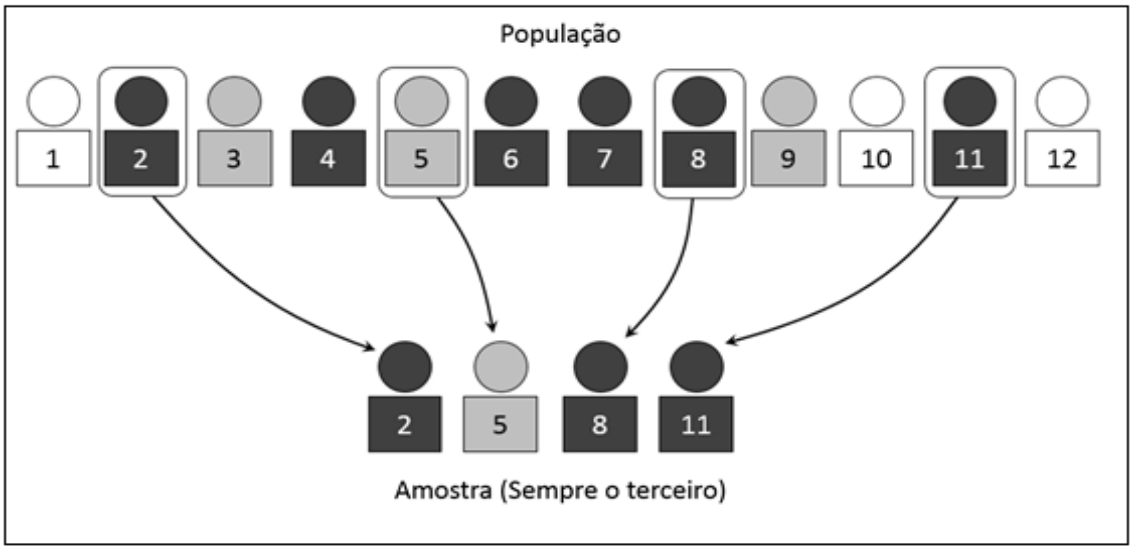


Figura 2: Método pelo sorteio sistemático.

O exemplo na figura 2, mostra uma população de 12 e uma amostra de 4 elementos, logo k = 12/4 = 3. As amostras serão retiradas a cada intervalo de k = 3, ou seja, intervalos de comprimento selecionando valores a cada 3 em 3 posições da lista da população até chegar o seu total.

Algumas vantagens são:

- A sistematização proporciona uma boa estimativa da média e do total, devido à distribuição uniforme da amostra em toda população;
- Maior rapidez e menor custo;
- O deslocamento entre as unidades é mais fácil;
- Obtém boas propriedades de representatividade, similar a amostragem aleatória simples;
- Pode garantir uma seleção perfeitamente equivalente a população.

As fórmulas para o cálculo do tamanho amostral são as mesmas do módulo de AASc.

$$n = \left(\frac{\sigma}{B/z_\alpha} \right)^2 = \frac{\sigma^2}{D}, \text{ tal que D} = \frac{B^2}{z_\alpha^2}$$

$\boldsymbol{\sigma^2}$: Variabilidade estimada da população, que é determinado por uma amostra piloto para estabelecer um estimador razoável a sigma;
B: Erro máximo fixado;
$\boldsymbol{z_\alpha}$: Grau de confiança;

3.2.3 – Amostragem Aleatória Estratificada (AAE)

Esta técnica, consiste em dividir toda a população em diferentes subgrupos ou estratos, de maneira que um indivíduo pode fazer parte apenas de um único estrato. Após as camadas serem definidas, para criar uma amostra, selecionam-se

indivíduos utilizando uma ou mais técnicas de amostragem em todos ou em cada um dos estratos, figura 3.

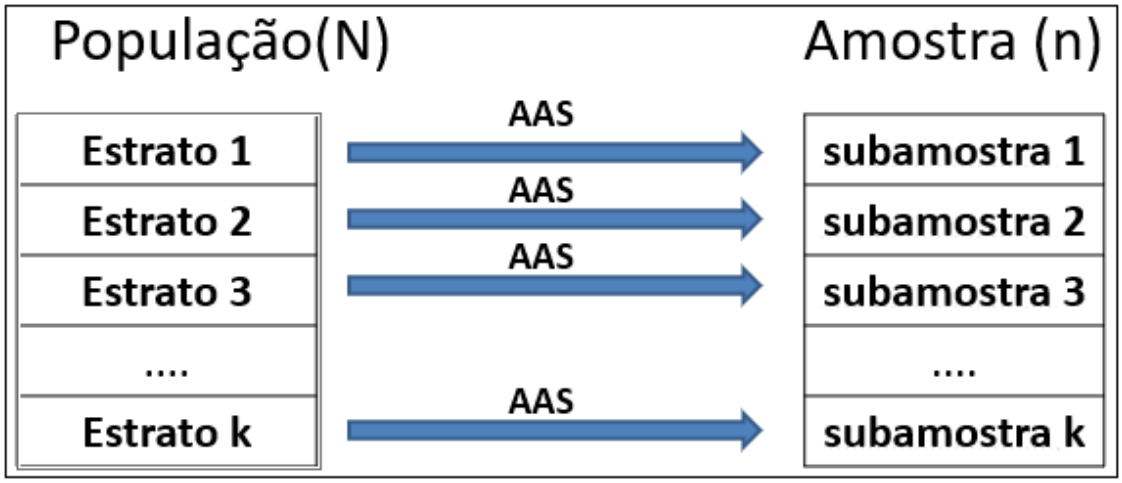


Figura 3: Divisão por estrato.

Os estratos são grupos homogêneos de itens, que por sua vez, são heterogêneos entre diferentes grupos. Se a condição comentada anteriormente for cumprida de forma correta, o uso da amostragem aleatória estratificada reduz o erro amostral melhorando a precisão dos resultados ao realizar um estudo sobre a amostra.

Utilizou-se a variável tipo de título para estratificar o valor total de *n* e dividi-la em diferentes quantidades para os subgrupos: curta, filme, tv filme, tv séries, tv episódios, tv curta, tv mini séries, tv especial e vídeo.

A divisão é de modo proporcional e não em estratos uniformes, ou seja, a amostra deverá obter subgrupos que tenham as mesmas proporções observadas na população.

$W_h = \left(\frac{Nh}{N}\right)$ onde N_h é o tamanho da população de cada estrato e N o tamanho da população.

Assim, pode-se calcular a amostra para cada estrato (n_h) através da expressão:

$$n_h = w_h n$$

Mas, antes é calculado o tamanho da amostra, utilizando a equação abaixo:

$$n = \frac{1}{D}\sum^{H} Wh\sigma^2 h = \frac{\sigma^2}{D}, \text{ tal que D} = \frac{B^2}{z_\alpha^2}$$

D: Resultado do erro máximo sobre o grau de confiança elevado a segunda potência.

$\boldsymbol{Wh}$**:** Peso da proporção do extrato *h*;

$\boldsymbol{\sigma^2}$**:** Variabilidade estimada da população, que é determinado por uma amostra piloto para estabelecer um estimador razoável a sigma;
h: Extrato;
B: Erro máximo fixado;
$\boldsymbol{Z_\alpha}$**:** Grau de confiança;

Para finalizar foi utilizado novamente o teste EPA para comparar AASs com o AAE.

$$EPA = \frac{Var_{\mathrm{AASs}}[\overline{y}]}{Var_{\mathrm{AASc}}[\overline{y}]}$$

No próximo capítulo, calculamos o tamanho ideal de uma amostra e apresentamos as análises descritivas da variável dependente média de classificação com as covariáveis, número de votos, classificação tipo de título, é adulto, minutos de duração, ano de produção, originalidade, número de temporadas, número de episódios e gênero, a fim de ajudar a descobrir o melhor modelo de regressão linear.

3.3 - Ajuste de modelos

Uma importante ferramenta na análise de dados, são os modelos de regressão estatísticos. Estes têm como principal objetivo, explicar por meio de uma equação devidamente estabelecida a variação e comportamento dos dados, e por isso são aplicáveis em diversas áreas de pesquisa.

Apesar dessa abrangência, os modelos de regressão linear, possuem suposições que necessitam ser validadas para que os resultados obtidos sejam válidos. Em muitos casos, quando esses pressupostos não são atendidos, faz-se necessário a utilização de modelos mais complexos que sejam capazes de acomodar as particularidades do problema estudado.

O modelo de regressão normal linear é em um primeiro momento o mais simples entre todos, no entanto, esse supõe, além da linearidade entre a resposta e as covariáveis, distribuição normal para os resíduos do modelo. No caso do conjunto de dados estas suposições não foram satisfeitas.

Nesta perspectiva, os modelos lineares generalizados surgem como um vantajoso método de estudo nas situações em que as hipóteses do modelo de regressão linear normal não foram atendidas.

3.3.1 – Modelos Lineares Generalizados

Um Modelo Linear Generalizado (MLG) (COSTA, 2019. p.71), é composto por três partes, sendo o componente aleatório, o componente sistemático e a função de ligação. O componente aleatório, envolve a distribuição que foi atribuída à variável dependente do modelo (variável resposta), o componente sistemático, é o preditor

linear que construímos com as covariáveis inseridas no modelo e a função de ligação, tem o papel de relacionar a variável resposta com o preditor linear.

1. Componente Aleatório: A variável resposta do modelo é a média de avaliação dos títulos. Tendo em vista as características desta (numérica, contínua e com intervalo de variação nos reais positivos até 10) a distribuição Gamma parece adequada.
2. Componente Sistemático: inicialmente o preditor linear é composto por:
    - Número de Votos: variável quantitativa discreta, que representa a quantidade de usuários que votaram;
    - É adulto: variável qualitativa nominal, indicadora 1 ou 0 que representa a classificação dos títulos adulto ou não adulto;
    - Minutos de duração: variável quantitativa discreta que representa a duração do título em minutos;
    - Ano de Produção: variável quantitativa discreto que representa o ano de produção dos entretenimentos;
    - Gênero: variável qualitativa nominal, que representa a classificação em relação aos gêneros dos títulos;
    - Originalidade: Indicadora 1 ou 0 que representa a classificação dos títulos como original ou não original.

A variável “número de votos” foi usada na construção da variável resposta, portanto não seria adequado utilizá-la diretamente no preditor linear, sendo assim, esta variável foi utilizada como uma ponderação no modelo.

3. Função de ligação: A distribuição Gama (atribuída à resposta) tem como ligação canônica a função inversa. Porém, pode se adequar também a outras como a logaritmo. Dessa forma, foram testadas a função de ligação inversa e a logaritmo.

Cada um dos modelos ajustados assume a estrutura a seguir.

1. ---Componente Aleatório:

$$Y_i = \ Média\ de\ Classificação_i$$

2. --- Componente Sistemático:

$$\eta_i = \ \beta_{0w} + \beta_{1w} * Minutosdeduração + \beta_{2w} * Classificaçãolivre + \ \beta_{3w} * Tituloorginal + \beta_{4w} * G.Adulto + \beta_{5w} G.Animação + \beta_{6w} G.Aventura + \beta_{7w} * G.Biograbia + \beta_{8w} * G.Comedia + \beta_{9w} * G.Crime + \beta_{10w} * G.Curtas + \beta_{11w} * G.Documentário + \beta_{12w} * G.Drama + \beta_{13w} * G.Esportes + \beta_{14w} * G.Família + \beta_{15w} * G.Fantasia + \beta_{16w} * G.FAroeste + \beta_{17w} * G.Ficcçaocientifica + \beta_{18w} * G.Musical + \beta_{19w} * G.Noticias + \beta_{20w} * G.Reality + \beta_{21w} * G.romance + \beta_{22w} * G.terror$$

3. --- Função de Ligação:

$Log(Y_i) = \eta_i$ ou $Y_i^{-1} = \eta_i$

3.3.2 – Variância Paramétrica e Não Paramétrica

Para tal objetivo, tentamos usar a técnica de Análise de Variância ANOVA paramétrica (RUMSEY, 2011, p. 153), considerando o estudo do tipo experimento fatorial com dois fatores (gênero e tipo). No entanto, é importante destacar aqui que as suposições necessárias para a aplicação desta são:

1. Observações são independentes e normalmente distribuídas;
2. Variância constante para cada combinação de níveis dos fatores.

A estrutura de definição do modelo pretendido é dada por:

$$Y_{ij} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \epsilon_{ij}$$

Y_{ij}: Variável resposta média de classificação.
μ: Efeito geral da média;
τ_i: Efeito do fator “Tipo”, *i*=1,2,3,4;
β_j: Efeito do fator “Gênero”, *j* =1, …, 22;
ϵ_{ij}: erro aleatório.

3.3.3 – Ajuste da Base

Alguns valores inconsistentes foram excluídos da base a fim de obter melhores resultados, que são eles:

1. Tipo de título igual a vídeo game;
2. Minutos de filme igual a 0;
3. Quantidade de votos abaixo de 6;
4. Números de episódios e temporadas iguais a 0.

4 – APLICAÇÃO

4.1 – Métodos de Amostragem

Após serem definidas as equações na seção 3.2, dos métodos de amostragem aleatório simples, com e sem reposição, amostragem aleatório sistemática e estratificada, as fórmulas foram utilizadas abaixo para encontrar o tamanho da amostra de cada método. Com os valores apresentados, nós os comparamos utilizando a procedimento EPA, a fim de selecionar a melhor amostragem.

- Amostragem aleatória simples com reposição:

Nível de confiança de 95%, temos um α = 0,05, $z_{2,5}$ = +1,96 e uma tolerância B = 0,05 N = 911.026, σ^2 = 1,39594.

n = $\left(\frac{1,39594*1,96}{0,05}\right)^2 \cong 2995$. O tamanho da amostra é equivalente a 2.995 registros.

- Amostragem aleatória simples sem reposição:

D = $\left(\frac{0,05}{1,96}\right)^2 \cong 0,000651$

$n = \left(\frac{1}{\frac{D+1}{S^2+N}}\right)\left(\frac{1}{\frac{0,000651+1}{1,960637+911.026}}\right) \cong 3.002$. O tamanho da amostra é equivalente a 3.002 registros.

Utilizamos o efeito do plano amostral para comparar os métodos de AAS com e sem reposição, e assim garantir o melhor dos dois a ser aplicado para cálculo do tamanho da amostra.

EPA = $\left(\frac{N-n}{N-1}\right) \Leftrightarrow \left(\frac{911.026-3002}{911.026-1}\right)$ = 0,9967. O método AAS sem reposição é o melhor dentre os dois, já que ele tem o resultado abaixo de 1 e é o mais simples de se aplicar.

- Amostragem aleatória sistemática:

Como a fórmula é igual ao processo AAS com reposição, então ficou definido que a amostragem AAS sem reposição, ainda é a melhor escolha.

- Amostragem aleatória estratificada:

D = $\left(\frac{0,05}{1,96}\right)^2 \cong 0,000651$

$$n = \left(\frac{1}{D}\sum wh\sigma h^2\right) \Leftrightarrow \left(\frac{1}{0{,}000651} * 1{,}99 * 0{,}122 + 1{,}76 * 0{,}246 + 1{,}79 * 0{,}043 + 2{,}08 * 0{,}066 + 1{,}46 * 0{,}449 + 1{,}69 * 0{,}003 + 1{,}46 * 0{,}009 + 2{,}19 * 0{,}008 + 2{,}36 * 0{,}054\right) \cong 2627$$

$$EPA = \left(\frac{AAE}{AASs}\right) \Leftrightarrow \left(\frac{2.627}{2.984}\right) = 0{,}88$$

O método AAE foi escolhido, já que apresenta valor EPA menor, o mais simples de se aplicar e com a estratificação como erro amostral reduzido.

Após a obtenção do tamanho da amostra em 2.627 itens pelo método AASs, foi calculado o peso de cada estrato (*wh)* e distribuído a amostra proporcionalmente usando a fórmula abaixo:

$Wh = \frac{Nh}{N}$ onde *Wh* é o peso de cada estrato;

exemplo: calculando o peso do tipo de título, tabela 2:

exemplo: *wh* (*filme*) = $\left(\frac{223.869}{911.026}\right) \cong 0{,}122.$

$W_1 = \left(\frac{111.302}{911.026}\right) \cong 0{,}122$ até $W_9 = \left(\frac{49.000}{911.026}\right) \cong 0{,}0538$

exemplo: *nh* (*filme*) = 0,122 * 2.627 = 320.

Os cálculos são mostrados na tabela abaixo “quadro de divisão dos extratos”:

Tabela 2: Especificações dos extratos da amostra.

Tipo do Título	Qtd.	Wh	Qtd. Extratos	Tamanho da Amostra
Curta	111302	0,1222	1º	321
Filme	223869	0,2457	2º	646
Episódio de tv	409298	0,4493	3º	1180
Mini Séries de tv	8291	0,0091	4º	24
Filmes de tv	39423	0,0433	5º	114
Séries de tv	60085	0,066	6º	173
Curtas de tv	2721	0,003	7º	8
Especiais de tv	7037	0,0077	8º	20
Vídeos	49000	0,0538	9º	141
Total (N)	**911026**	**1**		
Total (n)			**9**	**2.627**

A amostragem estratificada proporcional, produz um erro amostral menor ou igual a Amostra Aleatória Simples. A igualdade, ocorre quando as médias ou as proporções analisadas são iguais em todos os níveis dos estratos. Diferentemente no nosso arquivo, visto no gráfico acima, que mostra diferentes tamanhos em cada

grupo, por isso a estratificação produz mais assertividade maior, já que pega amostras proporcionais à população.

A conclusão foi, que o método mais eficaz para este estudo, é a combinação de Amostragem Aleatória Simples sem reposição associada a Amostragem Estratificada Proporcional, pelas várias vantagens descritas nos itens anteriores.

4.2 – Análises Descritivas da Amostra

Foram aplicadas técnicas descritivas (gráficos, descrição tabular e descrição paramétricas), para representar o conjunto de dados amostral, cujo objetivo básico é analisar as variáveis permitindo que se tenha uma visão global desses valores, ressaltando as tendências observadas, isoladamente, ou em comparação com outras variáveis.

4.2.1 – Filmes

A figura 4 apresenta a distribuição de frequência da avaliação dos títulos classificados como filme.

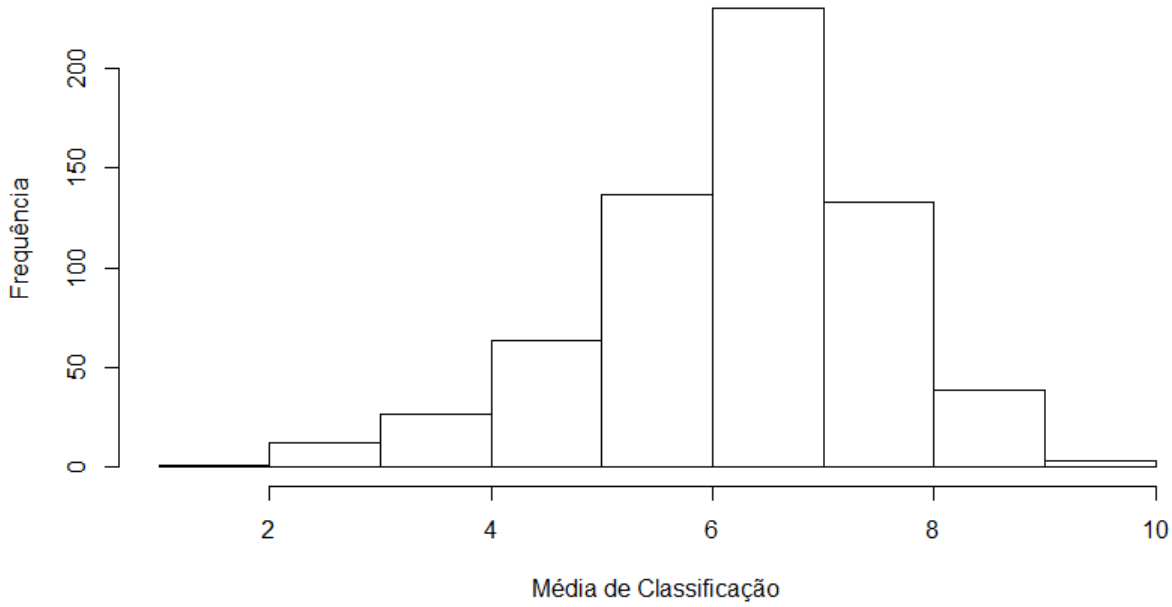


Figura 4: Histograma da média de classificação dos títulos classificados como filme.

Pela análise da figura 4 podemos notar que, a distribuição da média de classificação para a categoria filmes apresenta assimetria negativa e maior ocorrência entre os valores 5 e 8.

Nas estatísticas descritivas, mostradas na tabela 3, vemos que a variável dependente média de classificação tem o valor médio das notas dos usuários de 6,263 com desvio padrão de 1,281. A medida Q1 (primeiro quartil) indica que 25% dos usuários tem nota inferior a 5,600 e pelo Q3 (terceiro quartil) temos que 75% deles tem a média de notas até 7,100. Os valores da moda e da mediana são

superiores ao valor da média, indicando assimetria negativa na distribuição dos dados. Isto pode ser evidenciado no histograma apresentados na figura 4.

A covariável, número de votos tem média 3.387,400 votos com desvio padrão de 21.125,860 votos. A medida Q1, indica que 25% dos títulos tem a quantidade de votos inferior a 19,000 votos e pelo Q3 temos que 75% deles têm até a 310,000 votos. Já a covariável minutos de duração tem média de 94,390 minutos com desvio padrão de 23,299 minutos. A medida Q,1 indica que 25% dos títulos têm duração inferior a 81,250 minutos e pelo Q3 temos que 75% deles tem duração de até 105,000 minutos.

A medida de coeficiente de variação, indica que a variável número de votos, apresenta maior coeficiente de variação (623%) o que indica que esta variável é mais dispersa em relação a sua média, enquanto a média de classificação é a mais homogênea dentre as três.

Tabela 3: Medidas descritivas das variáveis quantitativas do tipo filmes.

Variáveis	n	Média	Mediana	Moda	Variância	Desvio Padrão	Coeficiente de Variação	Mínimo	Máximo	1º Quartil	3º Quartil
Média de Classificaçao	646	6,263	6,400	6,900	1,641	1,281	20,453	1,300	9,500	5,600	7,100
Número de Votos	646	3387,400	66,500	6,000	446301902,000	21125,860	623,660	6,000	352401,000	19,000	310,000
Minutos de Duração	646	94,390	91,000	90,000	539,631	23,229	24,610	30,000	199,000	81,250	105,000

A figura 5 apresenta a dispersão das médias de classificação com o número de votos.

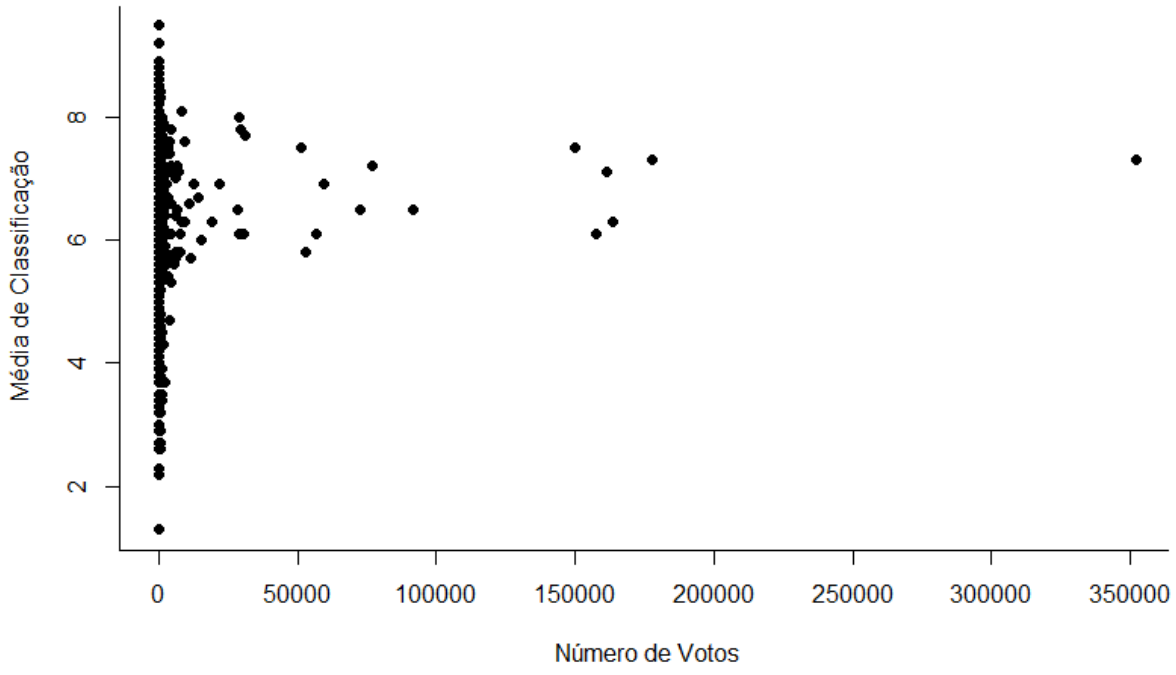


Figura 5: Gráfico de dispersão relacionando a média de classificação com o número de votos.

Na figura 5, inicialmente, não é possível tirar muitas conclusões, pois 75% dos dados obtiveram 273 votos ou menos e isso dificulta a visualização, tendo em vista que o máximo de votos foi 412382. Para melhorar a visualização faremos uma

ampliação do gráfico retirando os quatro títulos com mais de 100000 votos, sendo que todos os quatro obtiveram média maior do que 6.

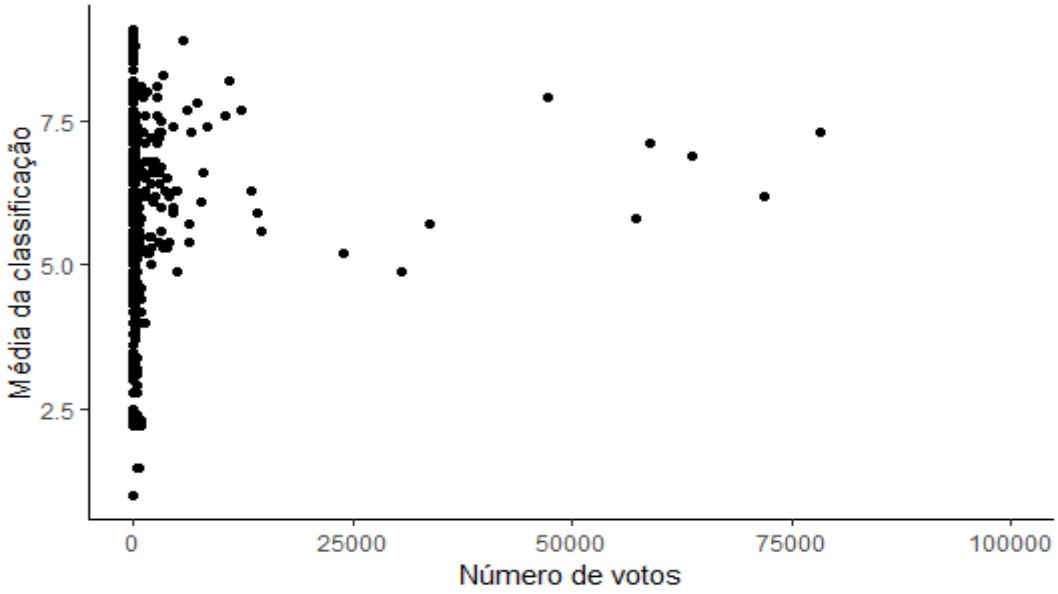


Figura 6: Gráfico de dispersão relacionando a média de classificação com o número de votos (recorte 1).

Ainda assim, a visualização da dispersão dos pontos no gráfico da figura 6 é inconclusiva, já que poucos pontos apresentam número de pontos maior que 25000.

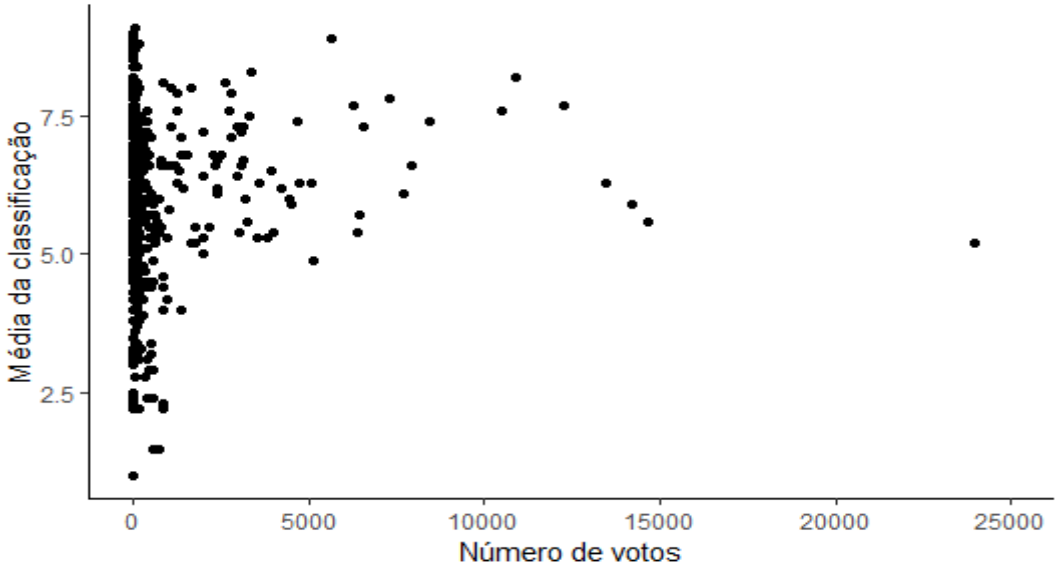


Figura 7: Gráfico de dispersão relacionando a média de classificação com o número de votos (recorte 2).

Por intermédio da figura 7, ainda é difícil enxergar com clareza a distribuição dos dados, podemos observar apenas que aparentemente, os títulos com mais votos tem média acima de 5. No entanto, é possível observar que os títulos que obtiveram número de votos maiores apresentam médias de classificação maiores que 5 no geral. Também se nota que não mudaram em relação as figuras anteriores: Ou seja, não há correlação entre avaliação e número de votos.

A figura 8, apresenta o Boxplot das médias de classificação com a classificação indicada.

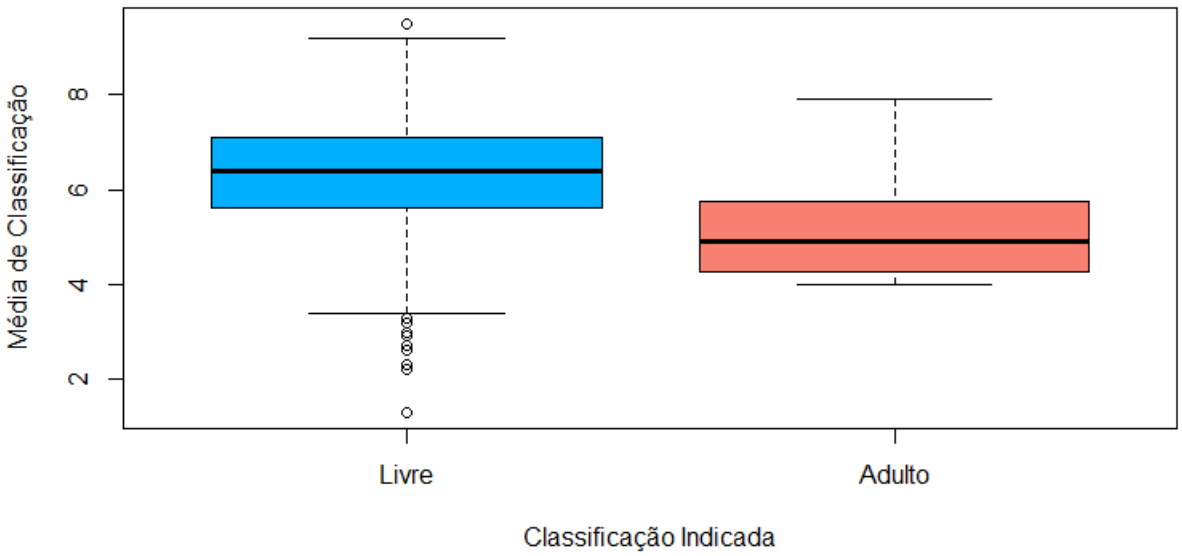


Figura 8: Boxplot relacionando a média de classificação e a classificação indicativa.

Por meio da figura 8, é possível observar que dentre os 263 filmes de classificação adultos e os 383 filmes livres, as medianas da classificação livre e adulto estão próximas (6,3 e 5,2), a distribuição da nota de avalição é parecida, e os filmes de classificação indicativa livre tem os 3 títulos com as menores médias de avaliação.

A figura 9, apresenta a dispersão das médias de classificação com os minutos de duração.

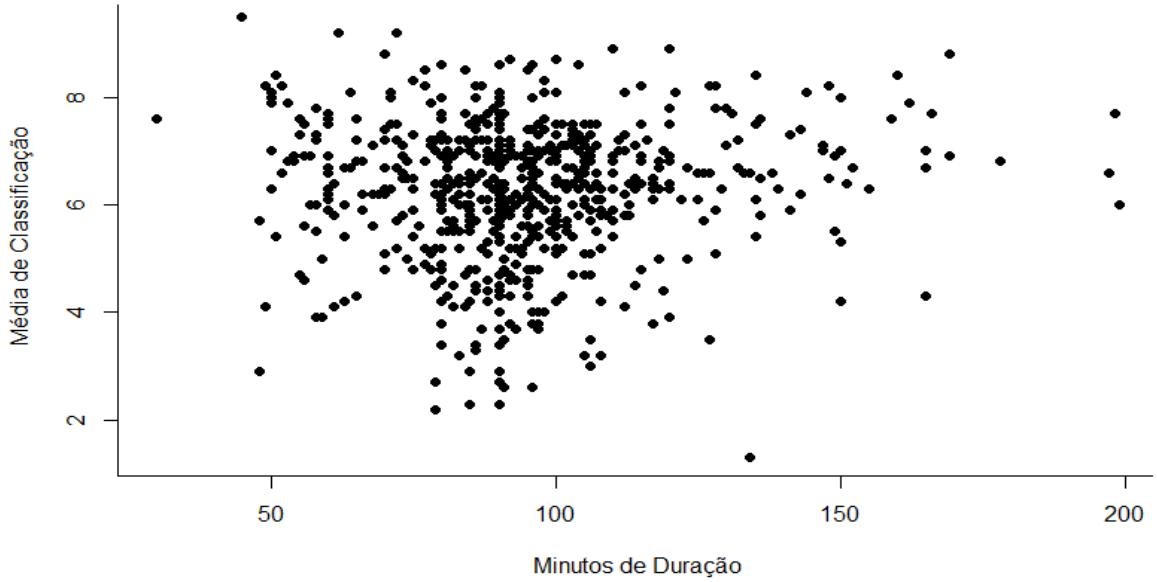


Figura 9: Gráfico de dispersão relacionando a média de classificação e minutos de duração do título.

Na figura 9, podemos ver que a maioria dos filmes têm por volta de 100 minutos, e que a média da classificação não parece seguir nenhum padrão ao longo do aumento ou diminuição do tempo de duração dos filmes.

A Figura 10, apresenta a dispersão das médias de classificação com os anos de produção.

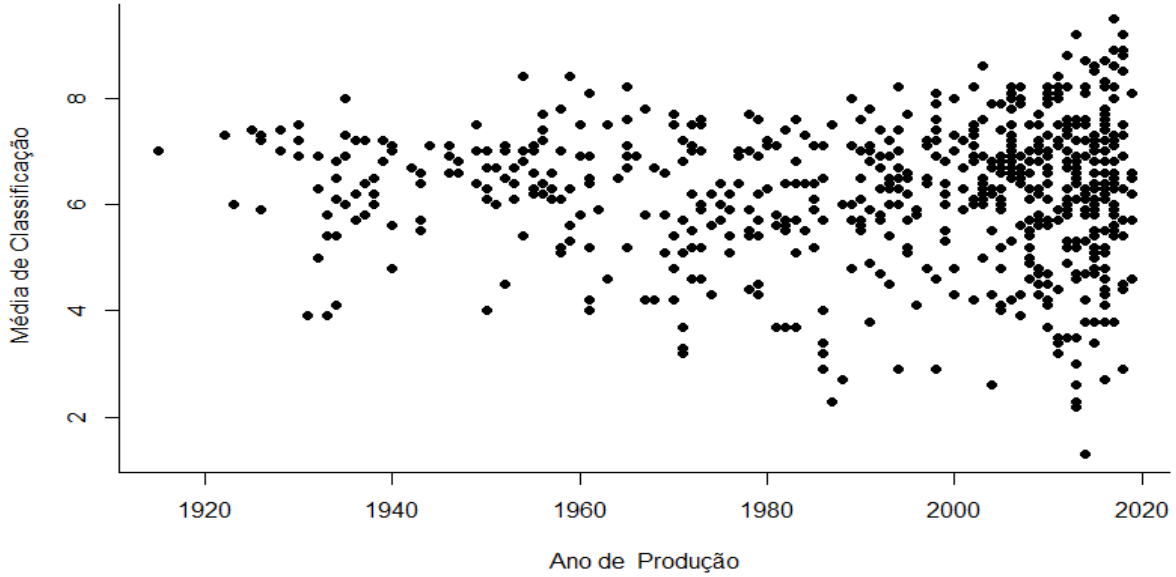


Figura 10: Gráfico de dispersão relacionando a média de classificação e ano de produção do título.

Analisando a figura 10, não é possível observar nenhuma tendência da média ao longo dos anos. É possível observar apenas um aumento na quantidade de filmes avaliados ao longo dos anos, sendo ainda mais evidente esse aumento a partir do ano 2000.

A figura 11, apresenta o boxplot das médias de classificação com a originalidade.

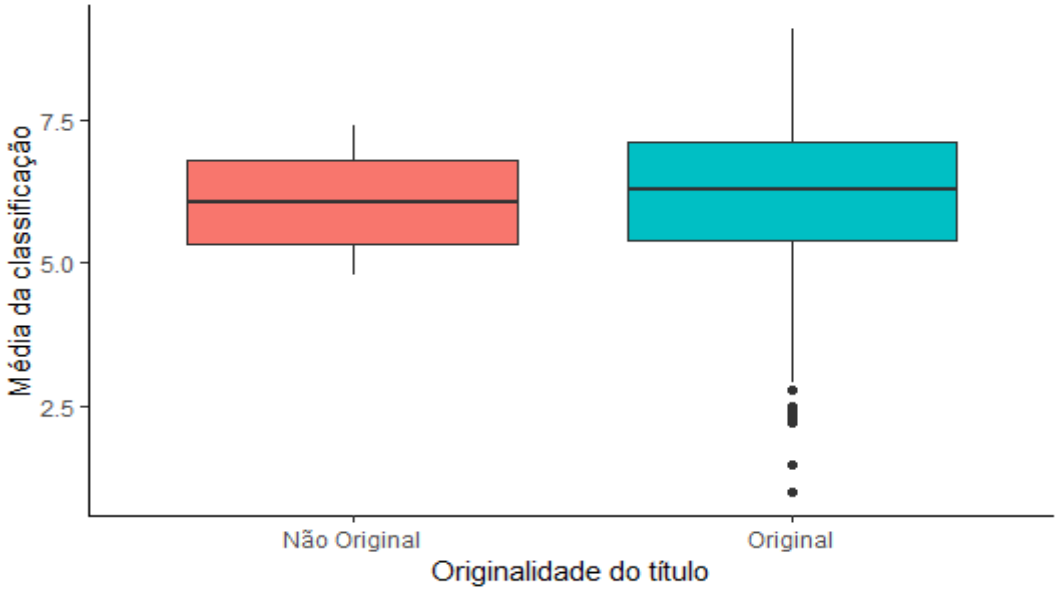


Figura 11: Boxplot relacionando a média de classificação e a originalidade do título.
Pela análise da figura 11, podemos notar que as medianas das duas categorias são parecidas (6,05 e 6,3) e a dispersão entre os títulos originais é maior. No entanto, é importante destacar que os títulos originais aparecem em maior quantidade nos

dados, uma vez que 4 títulos são classificados como não originais e 642 como originais.

A figura 12, apresenta a Boxplot das médias de classificação com os gêneros.

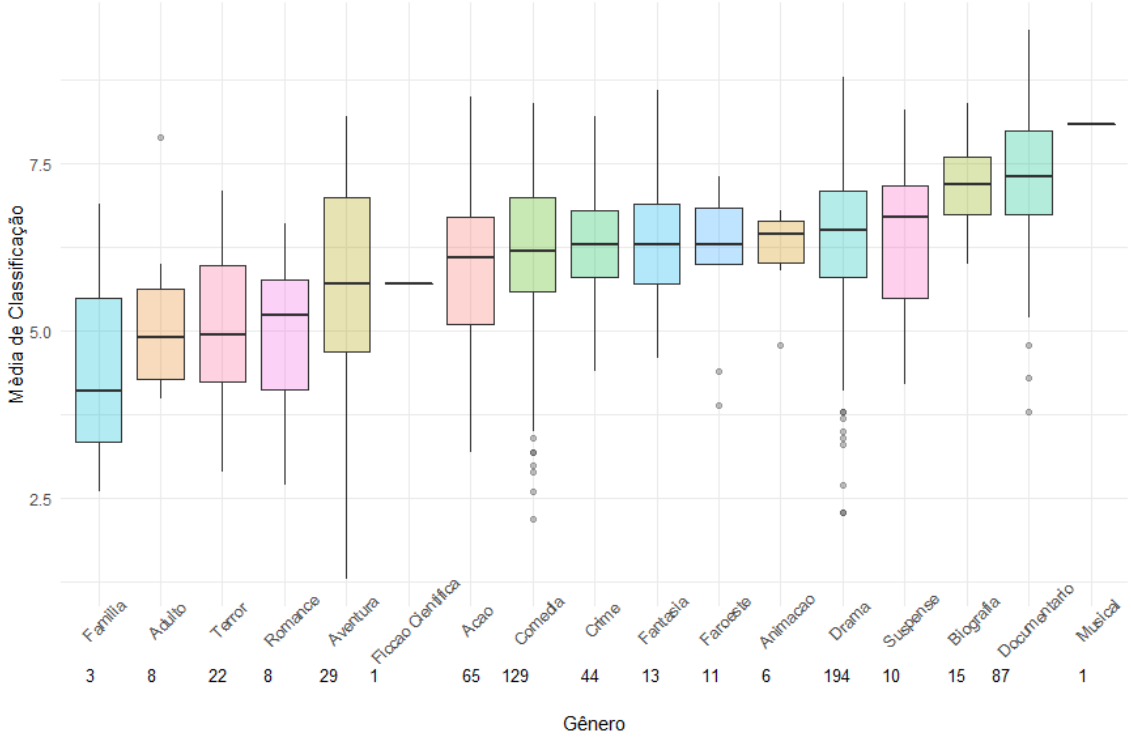


Figura 12: Boxplot relacionando a média da classificação e o gênero do título, com o valor da quantidade em cada grupo.

Pela análise da figura 12, podemos destacar alguns pontos: visualmente o gênero Documentário e família têm maior e menor mediana, respectivamente na amostra, a análise da dispersão em cada nível do variável gênero fica comprometida, devido ao desbalanceamento da quantidade de ocorrência de cada nível (por exemplo: musical e ficção científica aparecem somente uma vez, enquanto drama possui 194 ocorrências).

4.2.2 – Curtas

Baseado na análise descritiva, feita para o tipo de filme na seção 4.2.1, seguimos o mesmo critério para os títulos que se enquadram na classe curtas. A primeira variável a ser analisada, é a variável média classificação que foi representada pela figura 13, que apresenta a distribuição de frequência da avaliação dos títulos classificados como curtas.

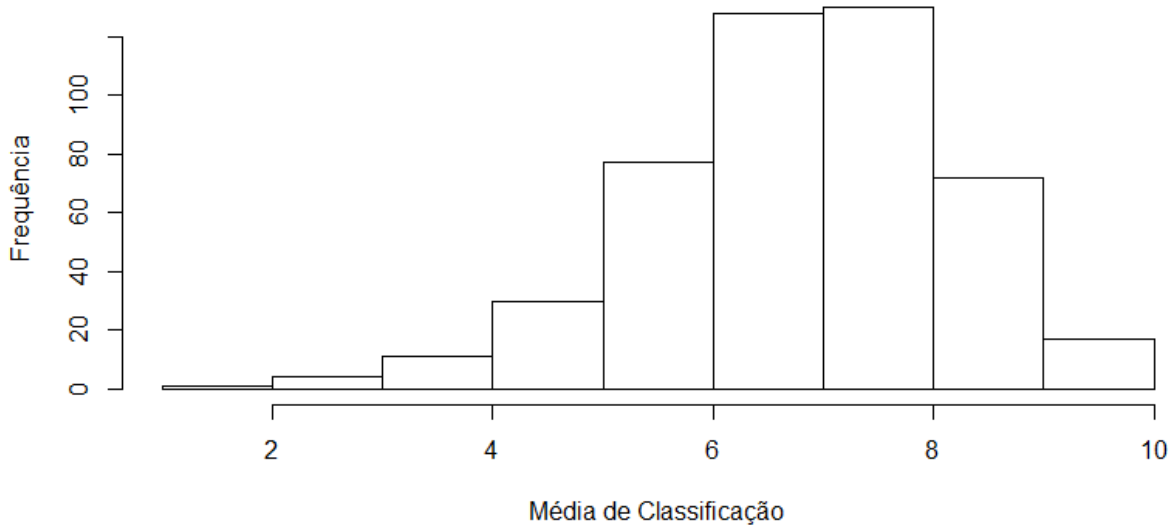


Figura 13: Histograma da média de classificação dos títulos.

Pela análise da figura 13, podemos notar que a distribuição da média de classificação para a categoria filmes, se apresenta assimétrica com concentração à direita, com maior ocorrência entre os valores 6 e 8.

Nas estatísticas descritivas mostradas na tabela 4, vemos que a variável média de classificação tem o valor médio das notas dos usuários de 6,847 com desvio padrão de 1,377. A medida Q1 (primeiro quartil) indica que 25% dos usuários têm nota inferior a 6,000 e pelo Q3 (terceiro quartil) temos que 75% deles têm a média de notas até 7,800. Os valores da moda e da mediana são maiores do que a média, indicando assimetria negativa na distribuição dos dados. Isto pode ser observado no histograma apresentados na figura 13.

A variável número de votos tem valor médio de 86,620 votos com desvio padrão de 355,668 votos. A medida Q1 indica que 25% dos usuários têm a quantidade de votos inferior a 8,000 e pelo Q3 temos que 75% deles têm até a 4866,000 votos. Já a variável minuto de duração tem a média de duração de 25,870 minutos com desvio padrão de 32,610 minutos. A medida Q1 indica que 25% tem a duração inferior a 7,000 minutos e pelo Q3 temos que 75% deles têm duração de até 27,000 minutos.

O coeficiente de variação da Média de classificação (20,1%), indica que esta variável possui os dados mais homogêneos, enquanto a variável Número de votos tem maior dispersão em relação à média do que as demais.

Tabela 4: Medidas descritivas das variáveis quantitativas do tipo curtas.

Variáveis	n	Média	Mediana	Moda	Variância	Desvio Padrão	Coeficiente de Variação	Mínimo	Máximo	1º Quartil	3º Quartil
Média de Classificaçao	470	6,847	7,000	7,000	1,898	1,377	20,111	1,400	10,000	6,000	7,800
Número de Votos	470	86,620	15,000	6,000	126499,800	355,668	410,607	6,000	4866,000	8,000	4866,000
Minutos de Duração	470	25,870	15,000	15,000	1073,284	32,610	126,053	1,000	256,000	7,000	27,000

A Figura 14, apresenta a dispersão das médias de classificação com o número de votos.

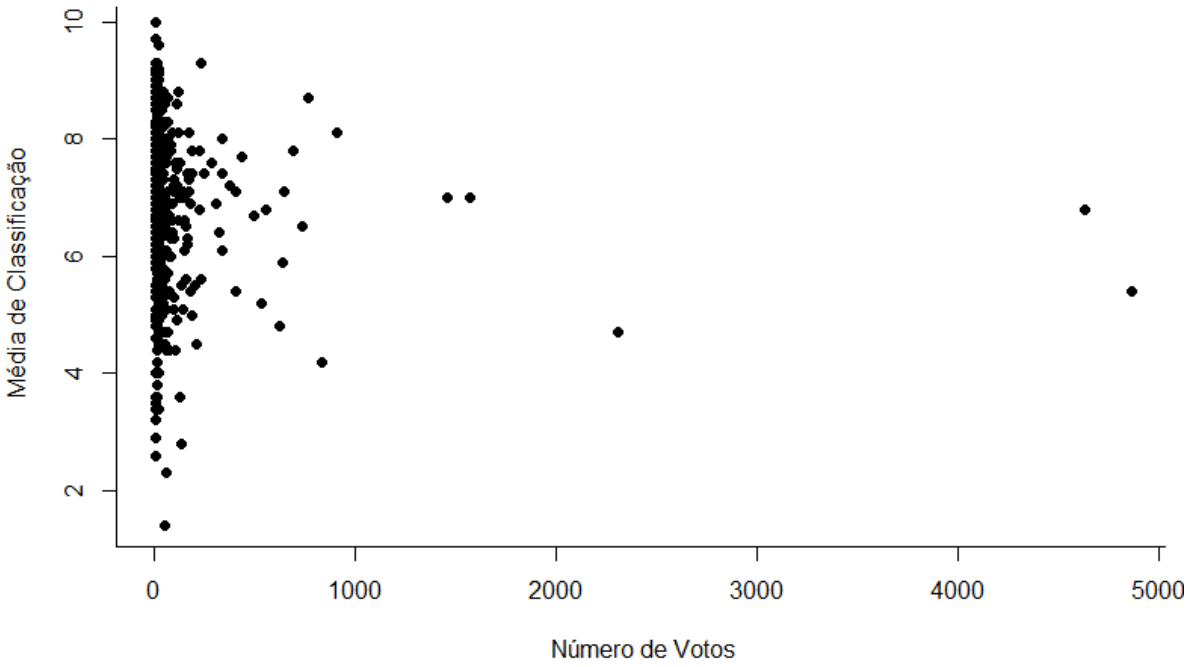


Figura 14: Gráfico de dispersão relacionando a média de classificação com o número de votos do título.

Na figura 14, não é possível tirar muitas conclusões, pois há alguns títulos que possuem muito mais votos que a maioria, e isto, dificulta a visualização. Para melhorar essa questão faremos um gráfico retirando os títulos que obtiveram mais de 1000 votos. Dentre estes, um deles obteve média 3,1 e todo o restante obteve média entre 6,8 e 8,2.

A figura 15, apresenta a dispersão das médias de classificação com o número de votos com valores acima de 1000 retirados.

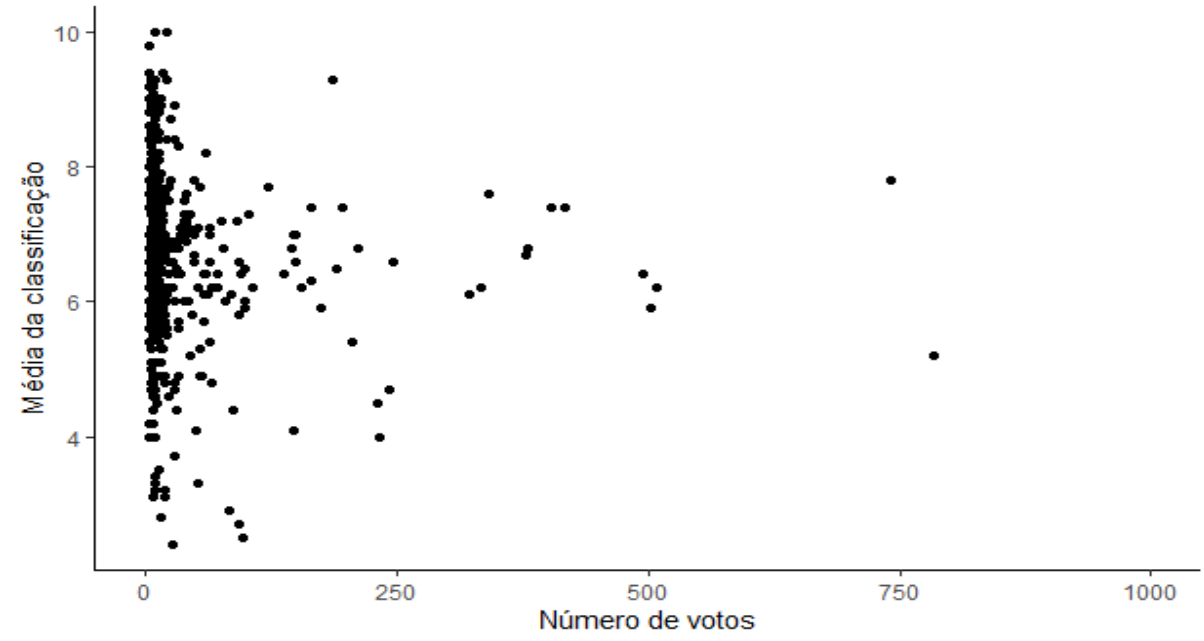


Figura 15: Gráfico de dispersão relacionando a média de classificação com o número de votos (Recorte).

A partir da figura 15, é possível observar que a maioria dos títulos tiveram poucos votos (75% abaixo de 29 votos). Visualmente não é possível enxergar um padrão evidente com o aumento do número de votos.

A figura 16, apresenta Boxplot das médias de classificação com a classificação indicativa.

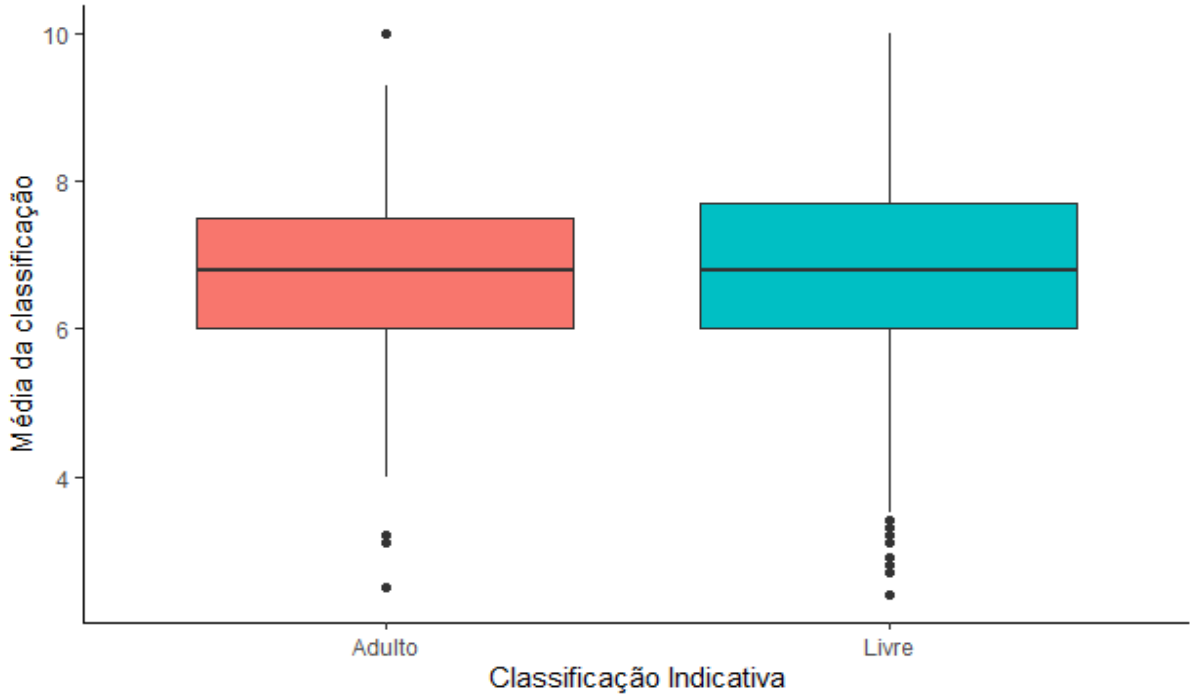


Figura 16: Boxplot relacionando a média de classificação e a classificação Indicativa do título.

Na figura 16, podemos ver que distribuição da média de classificação é semelhante nos dois casos, sendo que em ambos a mediana é igual a 6.8. Além disso, temos que na amostra há 332 títulos de classificação livre e 138 de classificação adulta.

A figura 17, apresenta a dispersão das médias de classificação com os minutos de duração.

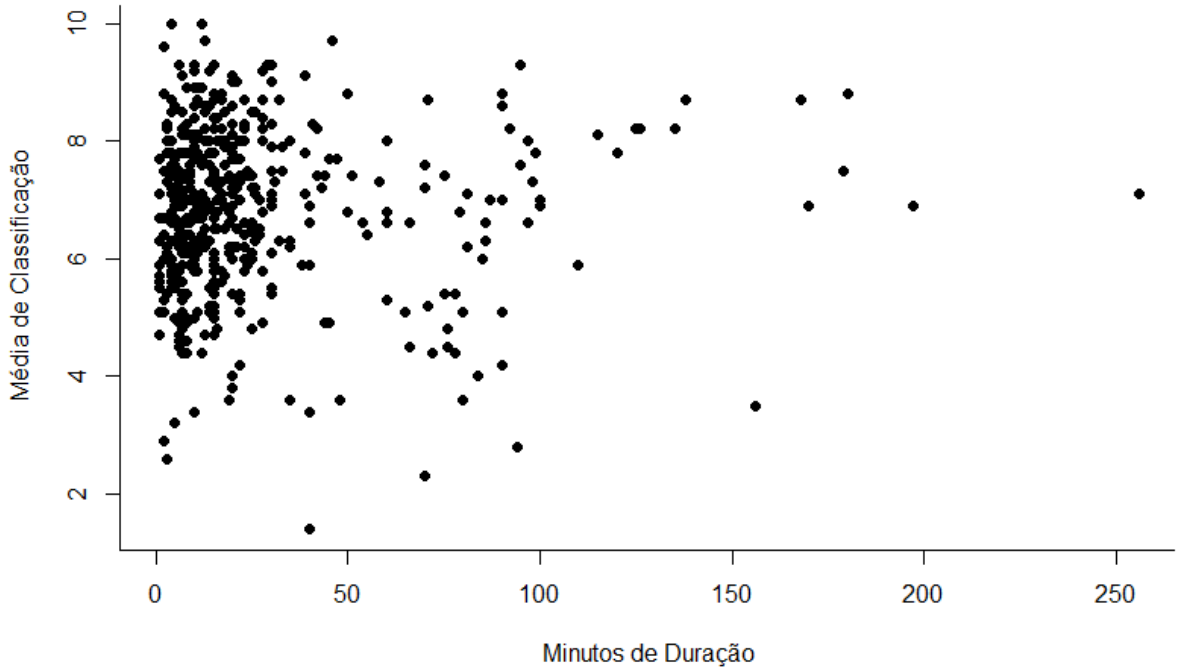


Figura 17: Gráfico de dispersão relacionando a média de classificação e minutos de duração do título.

Na figura 17, é possível enxergar que os dados estão acumulados dentro do tempo de 50 minutos, mas existem vários valores acima dele, mostrando que existem dados inconsistentes, pois grandes tempos de duração para um curta, que é definido justamente pelo tempo reduzido.

A figura 18, apresenta a dispersão das médias de classificação com o ano de produção.

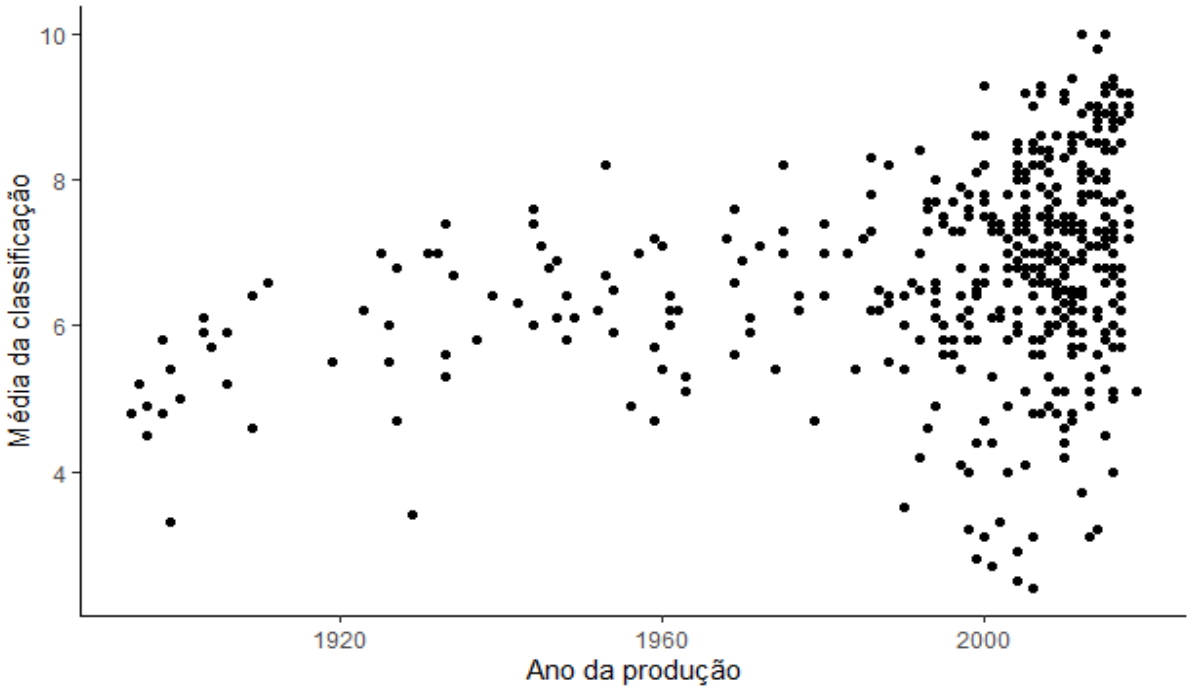


Figura 18: Gráfico de dispersão relacionando a média de classificação e o ano da produção do título.

Pela figura 18 é possível observar que ao longo dos anos, a quantidade de títulos avaliados aumenta. Além disso, a média de classificação revela visualmente uma tendência crescente assim como a dispersão.

A figura 19, apresenta o Boxplot das médias de classificação com a originalidade.

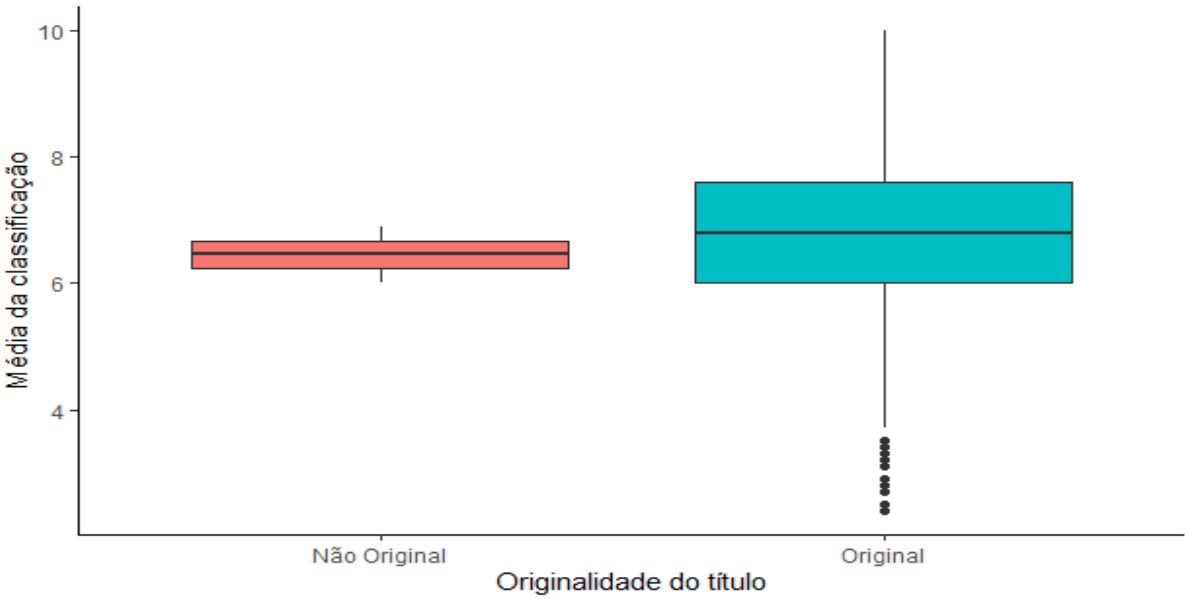


Figura 19: Boxplot que relaciona a média da classificação com a originalidade do título.

Na figura 19, podemos notar que visualmente a mediana das duas categorias são parecidas e a dispersão entre os títulos originais é maior. No entanto, é importante destacar que os títulos originais aparecem em maior quantidade nos dados, uma vez que 2 títulos são classificados como não originais e 468 como originais.

A figura 20, apresenta o Boxplot das médias de classificação com a variável gênero.

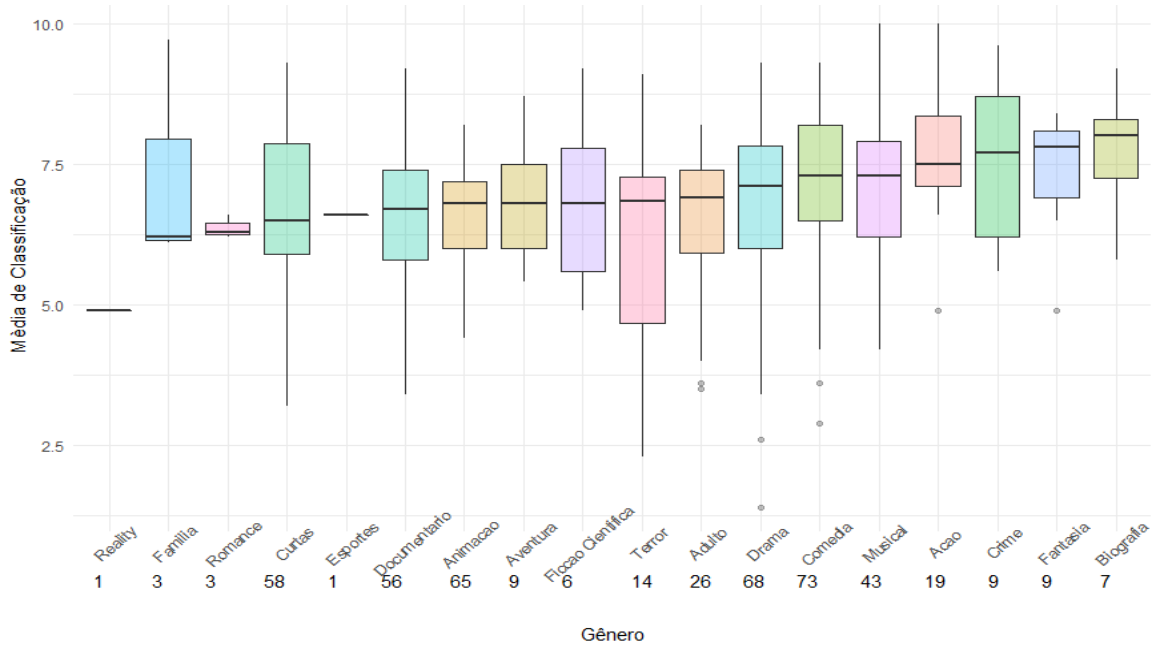


Figura 20: Gráfico de dispersão que relaciona a média de classificação com o gênero do título, com o valor da quantidade em cada grupo.

Mediante a análise da figura 20, podemos destacar que: visualmente o gênero biografia e reality têm maior e menor mediana, respectivamente na amostra, a análise da dispersão em cada nível da variável gênero fica comprometida, devido ao desbalanceamento da quantidade de ocorrência de cada nível (por exemplo: reality e esportes aparecem somente uma vez, enquanto comédia possui 73 ocorrências).

4.2.3 – Programas de tv

Continuando a análise, agora para títulos do tipo programas de tv, a figura 21, apresenta a distribuição de frequência da avaliação dos títulos classificados como programas de tv.

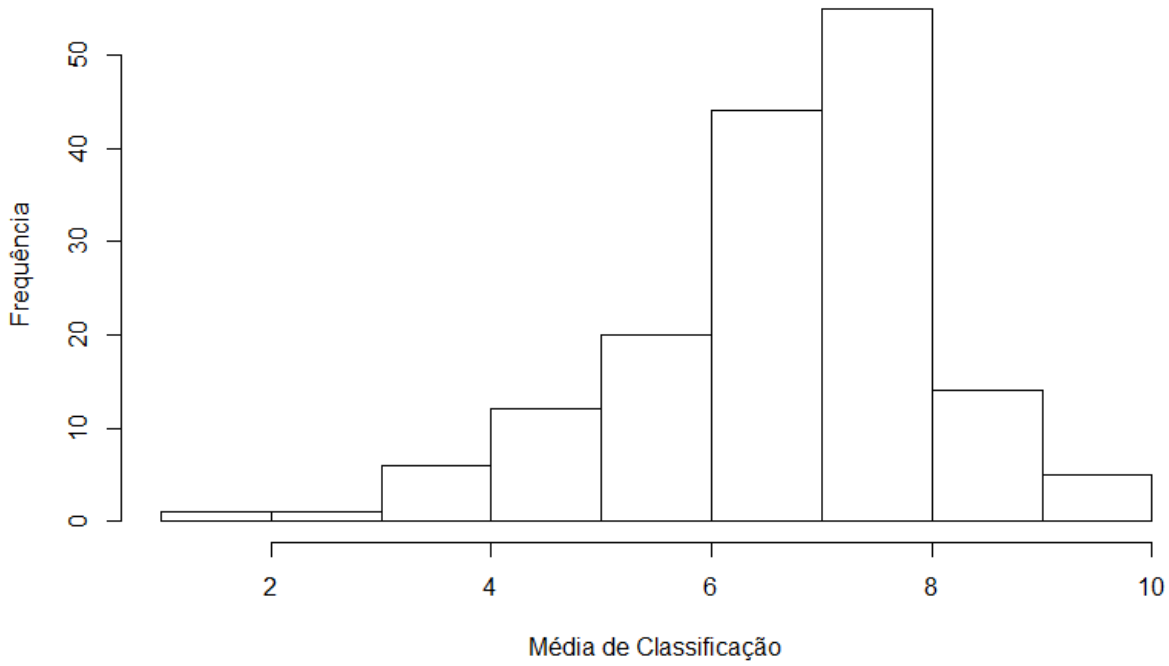


Figura 21: Histograma média de classificação do título.

Pela figura 21, podemos ver que a distribuição das médias é assimétrica negativa, com maior ocorrência de valores entre 6 e 8.

Nas estatísticas descritivas mostradas da tabela 5, vemos que a variável média de classificação tem o valor médio das notas dos usuários de 6,691 com desvio padrão de 1,402. A medida Q,1 indica que 25% dos usuários têm nota inferior a 6,025 e pelo Q3 temos que 75% deles têm a média de notas até 7,600.

A variável número de votos tem o valor médio de 257,400 votos com desvio padrão de 1.044,860 votos. A medida Q1, indica que 25% dos usuários têm menos de 15,000 votos e pelo Q3 temos que 75% deles tem quantidade inferior a 144,200

votos. Já a variável minuto de duração tem a média de duração de 94,790 minutos com desvio padrão de 56,904 minutos. A medida Q1, indica que 25% tem a duração inferior a 60,000 minutos e pelo Q3, temos que 75% deles têm duração de até 98,750 minutos.

O coeficiente de variação indica que a variável número de votos, tem maior heterogeneidade dentre as analisadas.

Tabela 5: Medidas descritivas das variáveis quantitativas do tipo programas de tv.

Variáveis	n	Média	Mediana	Moda	Variância	Desvio Padrão	Coeficiente de Variação	Mínimo	Máximo	1º Quartil	3º Quartil
Média de Classificaçao	158	6,691	6,900	7,500	1,965	1,402	20,954	1,200	9,400	6,025	7,600
Número de Votos	158	257,400	37,500	5,000	1091732,000	1044,860	405,929	6,000	10363,000	15,000	144,200
Minutos de Duração	158	94,790	90,000	90,000	3238,128	56,904	60,032	2,000	600,000	60,000	98,750

A figura 22 apresenta a dispersão das médias de classificação com o número de votos.

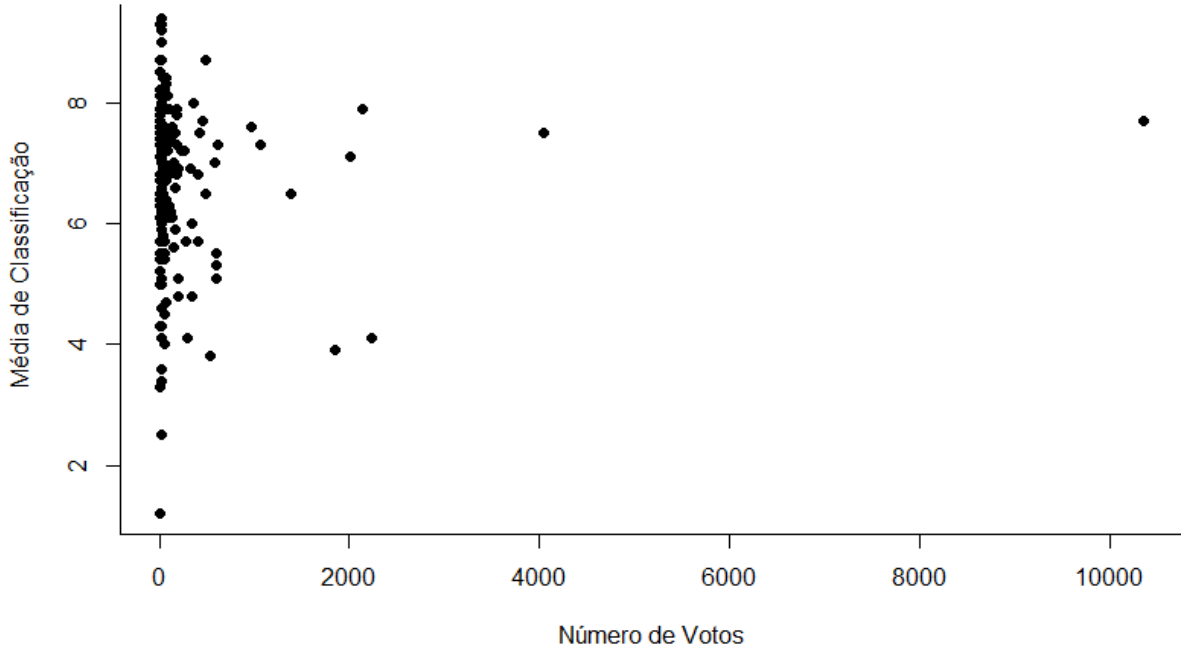


Figura 22: Gráfico de dispersão relacionando a média de classificação e o número de votos do título.

Inicialmente, não é possível tirar muitas conclusões através da figura 22, pois há um título que possui muito mais votos que a maioria, e isto, dificulta a visualização. Para melhorar a visualização faremos um gráfico retirando o título que tem 10363 votar e média de classificação igual a 7,7.

A figura 23, apresenta a dispersão das médias de classificação com o número de votos com recorte de valores acima de 6 mil.

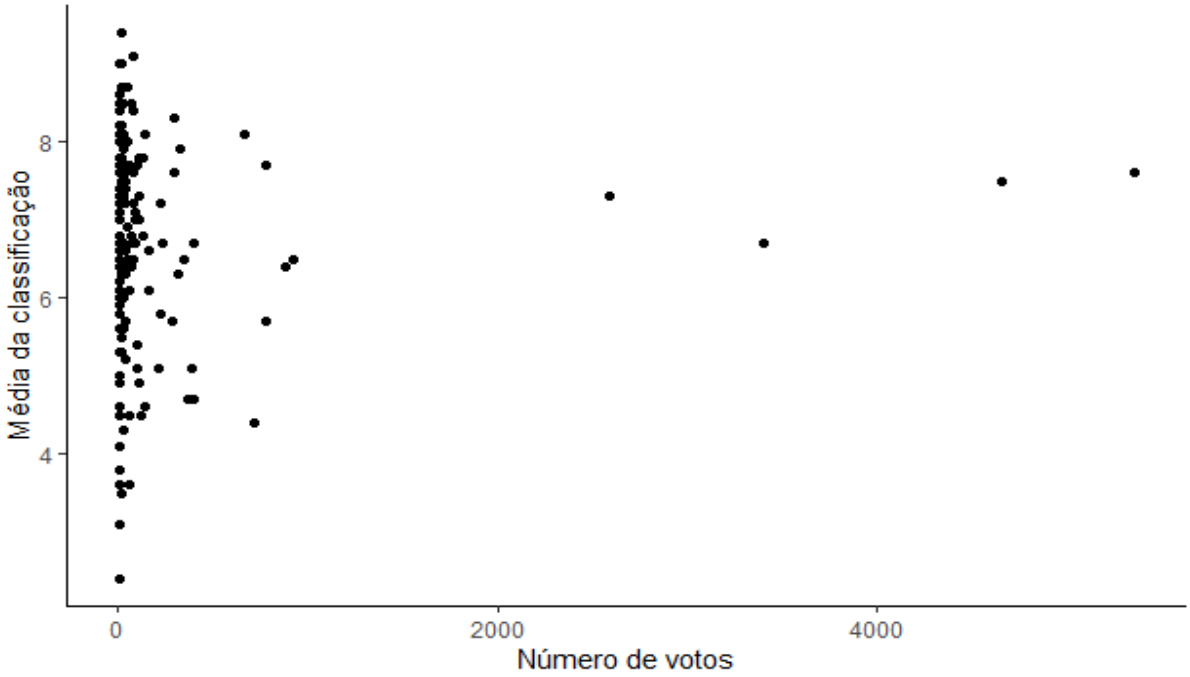


Figura 23: Gráfico de dispersão relacionando a média de classificação e o número de votos (Recorte).

Através da figura 23, é possível observar 4 pontos muito discrepantes em relação aos demais. Faremos então outro recorte, removendo estes 4 pontos, sabendo que todos possuem média acima de 6.

A figura 24, apresenta a dispersão das médias de classificação com o número de votos com recorte de valores acima de 1 mil.

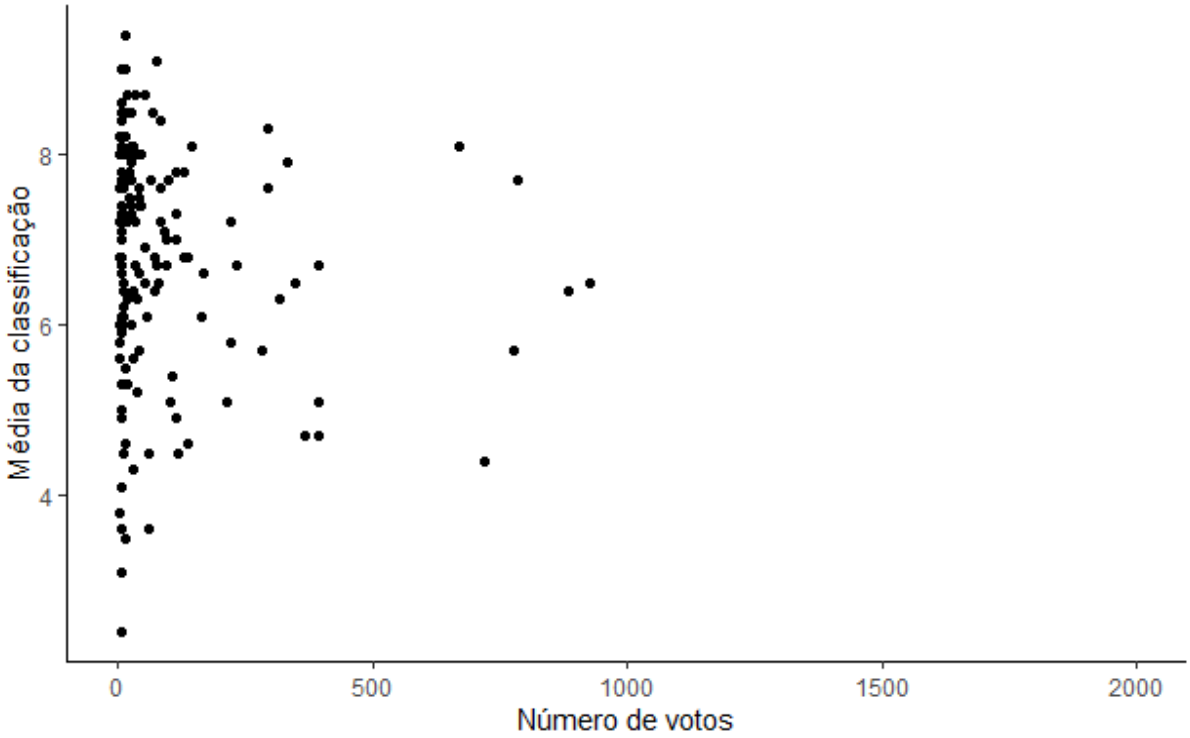


Figura 24: Gráfico de dispersão relacionando a média de classificação e o número de votos (Recorte2).

Na figura 24, podemos ver que na maioria dos títulos há poucos votos (75% abaixo de 102).

A figura 25, apresenta o Boxplot das médias de classificação com a classificação indicativa.

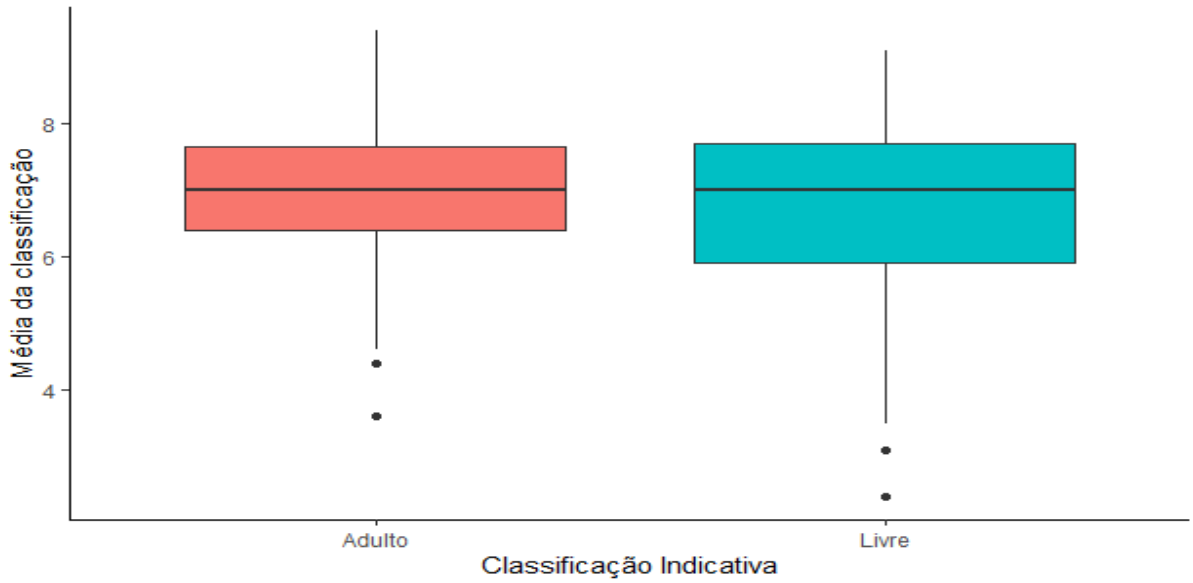


Figura 25: Boxplot relacionando a média da classificação com a classificação Indicativa do título.

Na figura 25, podemos ver que os títulos de classificação livre (332) têm maior dispersão do que os adultos (138), sendo que a mediana dos dois grupos é parecida.

A figura 26, apresenta a dispersão das médias de classificação com os minutos de duração.

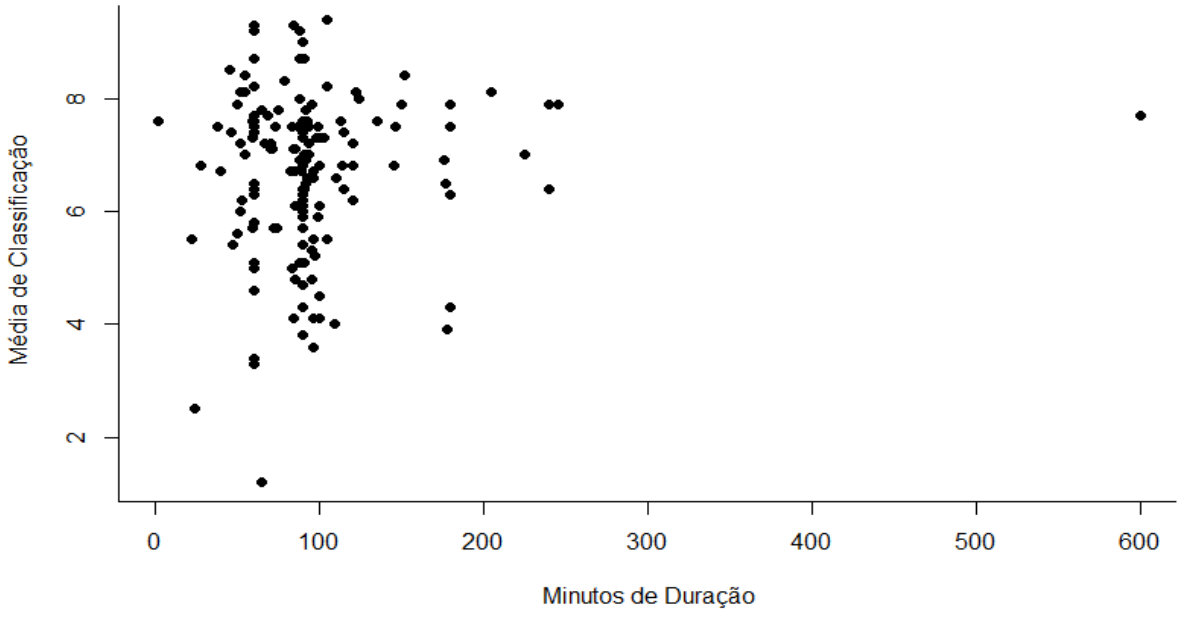


Figura 26: Gráfico de dispersão relacionando a média de classificação e o número de duração.

Através da figura 25, podemos ver que na categoria programas de tv, ainda é possível observar programas próximos a 0 minutos de duração. A maiorias dos títulos tem por volta de 100 minutos e a média não apresenta nenhum padrão com o aumento ou diminuição do tempo de duração.

A figura 27, apresenta a dispersão das médias de classificação com o ano de produção.

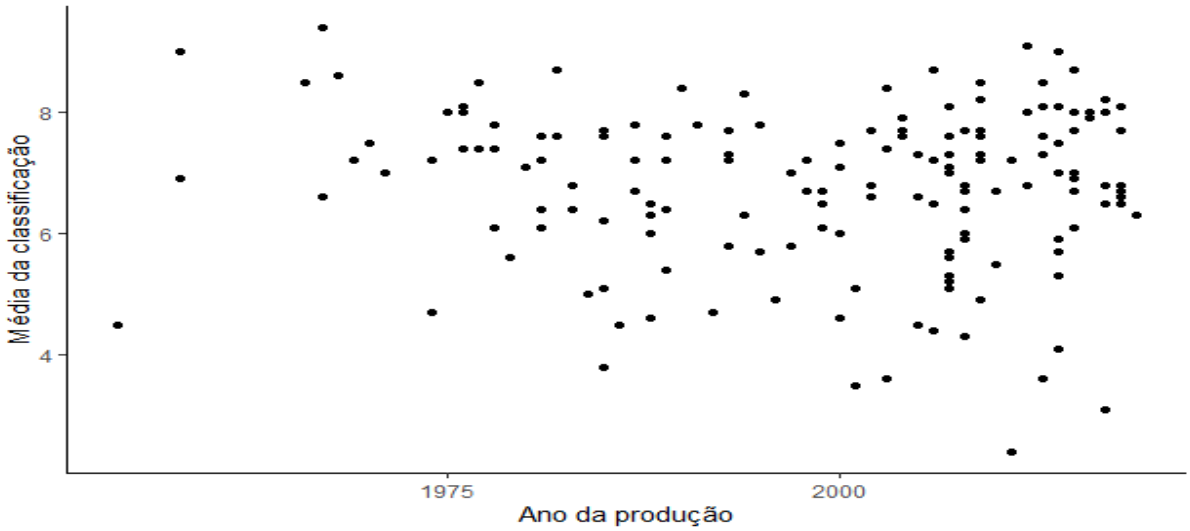


Figura 27: Gráfico de dispersão relacionando a média de classificação e o ano de produção do título.

Na figura 27, não é possível observar nenhuma tendência da média ao longo dos anos, é possível observar apenas um aumento na quantidade de títulos avaliados.

A figura 28, apresenta o Boxplot das médias de classificação com a originalidade.

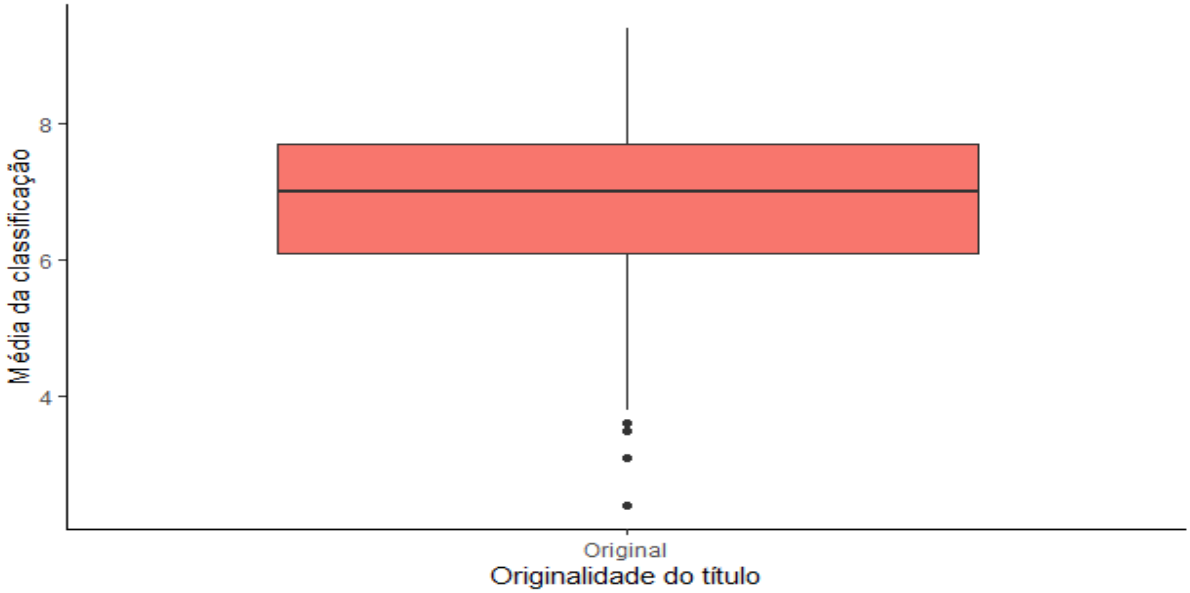


Figura 28: Boxplot relacionando a média de classificação e a originalidade do título.

Por meio da figura 28, podemos observar que a amostra não apresenta nenhum título classificado como não original, dessa forma o Boxplot é uma outra forma de visualizar a distribuição da média de classificação dos programas de tv, transmitindo as mesmas informações que o histograma mostrado na figura 21.

A figura 29, apresenta o Boxplot das médias de classificação com a variável gênero.

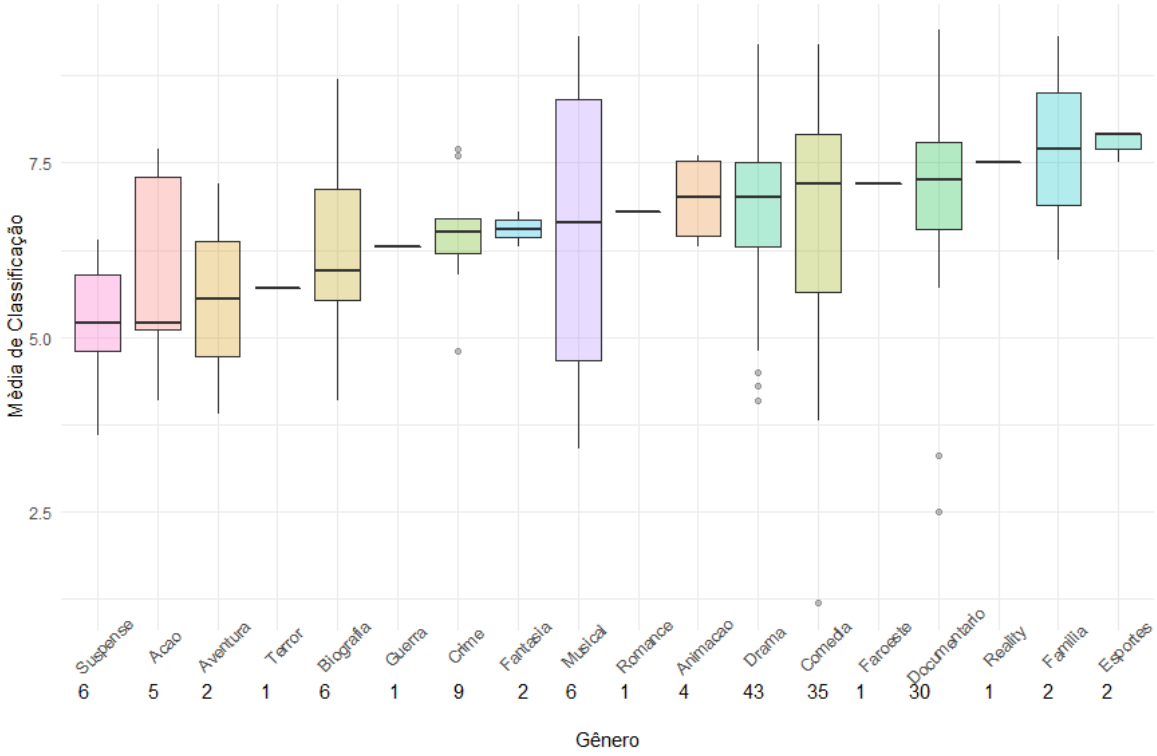


Figura 29: Boxplot relacionando a média da classificação com o gênero do título, com o valor da quantidade em cada grupo.

Através da análise da figura 29, podemos destacar que: visualmente os gêneros esportes têm maior mediana e os gêneros suspense e ação apresentam a menor mediana na amostra.

A análise da dispersão em cada nível da variável gênero fica comprometida, devido ao desbalanceamento da quantidade de ocorrência de cada nível (por exemplo: terror, guerra, romance, faroeste e reality aparecem somente uma vez, enquanto drama possui 43 ocorrências).

4.2.4 – Seriados

Para finalizar, a última análise que se refere a títulos do tipo seriados, a figura 30, apresenta a distribuição de frequência da avaliação dos títulos classificados como seriados.

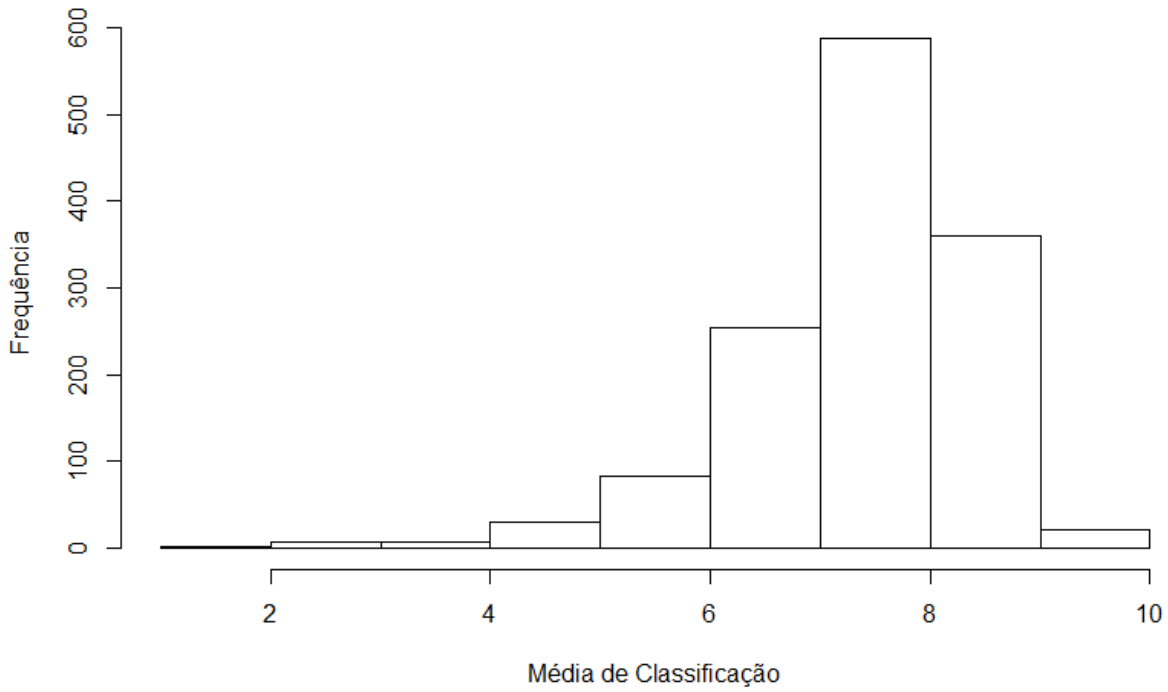


Figura 30: Histograma da média de classificação do título.

Através da figura 30, podemos ver que a distribuição das médias é assimétrica com concentração a direita, com maior ocorrência de valores entre 6 e 9.

Nas estatísticas descritivas mostradas da tabela 6, vemos que a variável média de classificação tem o valor médio da média das notas dos usuários de 7,423 com desvio padrão de 1,069. A medida Q1, indica que 25% dos usuários têm nota inferior a 7,000 e pelo Q3 temos que 75% deles têm a média de notas até 8,100. Os valores da moda e da mediana são maiores que o da média indicando assimetria negativa na distribuição dos dados.

A variável número de votos tem o valor médio de 354,400 votos com desvio padrão de 2.139,020 votos. A medida Q1, indica que 25% dos usuários têm a quantidade inferior a 6,000 votos e pelo Q3, temos que 75% deles têm até 106,000 votos. Já a variável minuto de duração tem a média de duração de 38,700 minutos com desvio padrão de 27,172 minutos. A medida Q1, indica que 25% tem a duração igual a 23,000 minutos e pelo Q3, temos que 75% deles tem duração de até 45,000 minutos.

O coeficiente de variação da média de classificação (14,4%), mostra que dentre as analisadas ela é a variável mais homogênea, enquanto a variável minuto de episódio tem maior dispersão em relação à média.

Tabela 6: Medidas descritivas das variáveis quantitativas do tipo seriados.

Variáveis	n	Média	Mediana	Moda	Variância	Desvio Padrão	Coeficiente de Variação	Mínimo	Máximo	1º Quartil	3º Quartil
Média de Classificaçao	1353	7,423	7,600	7,700	1,144	1,069	14,401	1,000	9,900	7,000	8,100
Número de Votos	1353	354,400	30,000	8,000	4575409,000	2139,020	603,561	6,000	39241,000	6,000	106,000
Minutos de Duração	1353	38,700	30,000	30,000	738,370	27,172	70,212	1,000	720,000	23,000	45,000
Número de Temporadas	1353	3,415	2,000	1,000	24,092	4,908	143,719	1,000	52,000	1,000	4,000
Minutos de Episódios	1353	21,050	7,000	1,000	31251,920	176,781	839,815	1,000	4901,000	2,000	14,000

A figura 31, apresenta a dispersão das médias de classificação com o número de votos.

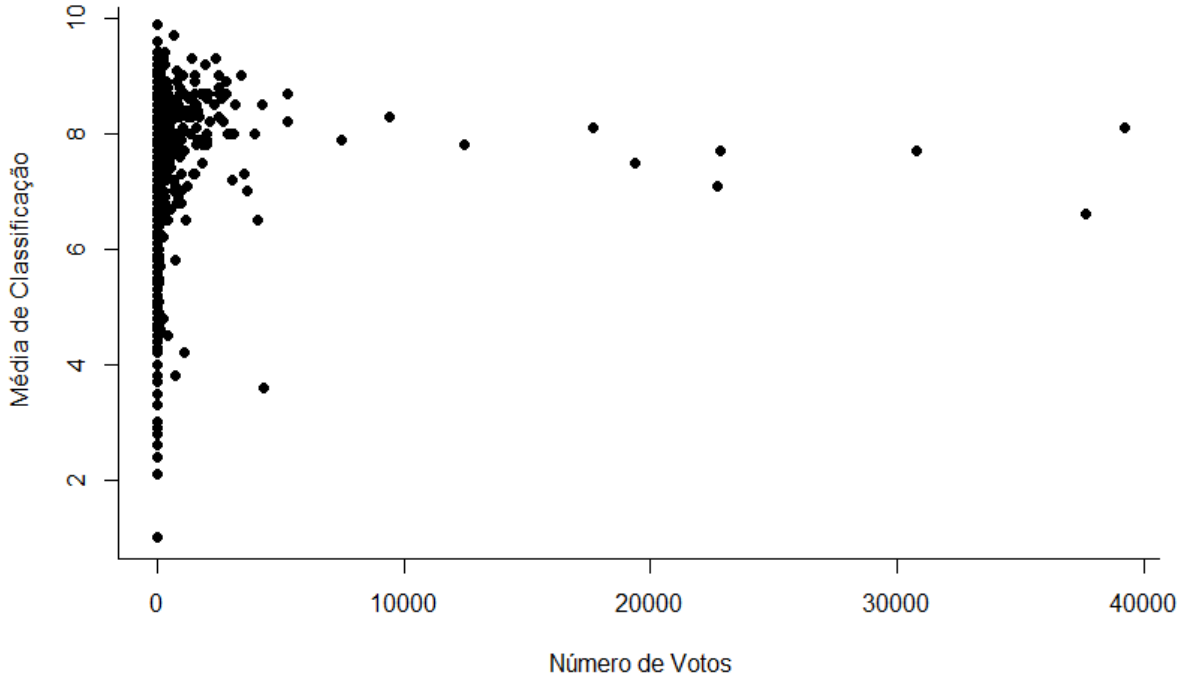


Figura 31 Gráfico de dispersão relacionando a média de classificação e o número de votos do título.

Inicialmente, não é possível tirar muitas conclusões através da figura 31, pois há dois títulos que possuem muito mais votos que a maioria, e isto, dificulta a visualização. Para melhorar a visualização faremos um recorte do gráfico retirando os dois títulos com maior número de votos. Ambos com média maior que 9.

A figura 32 apresenta a dispersão das médias de classificação com o número de votos com recorte de valores acima de 7.500.

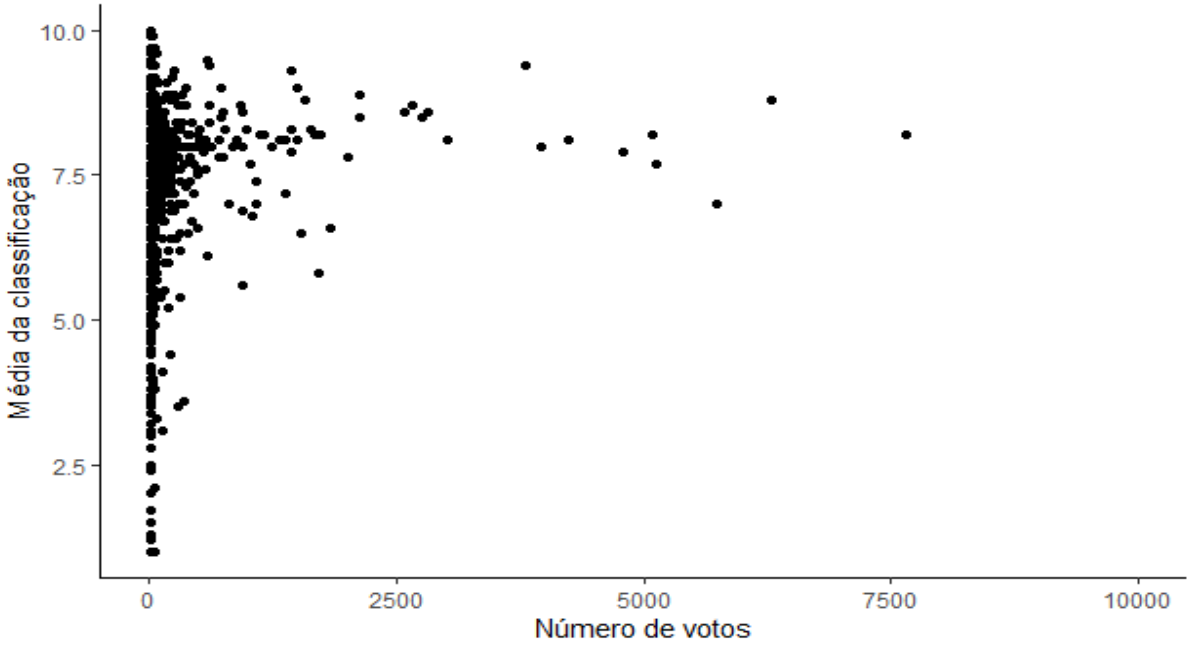


Figura 32: Gráfico de dispersão relacionando a média de classificação e o número de votos (recorte).

Através da figura 32, podemos observar que a maioria dos títulos avaliados possuem poucos votos (75% abaixo de 58), podemos ver também que os títulos com mais votos possuem média acima de 5.

A figura 33, apresenta o Boxplot das médias de classificação com a classificação indicativa.

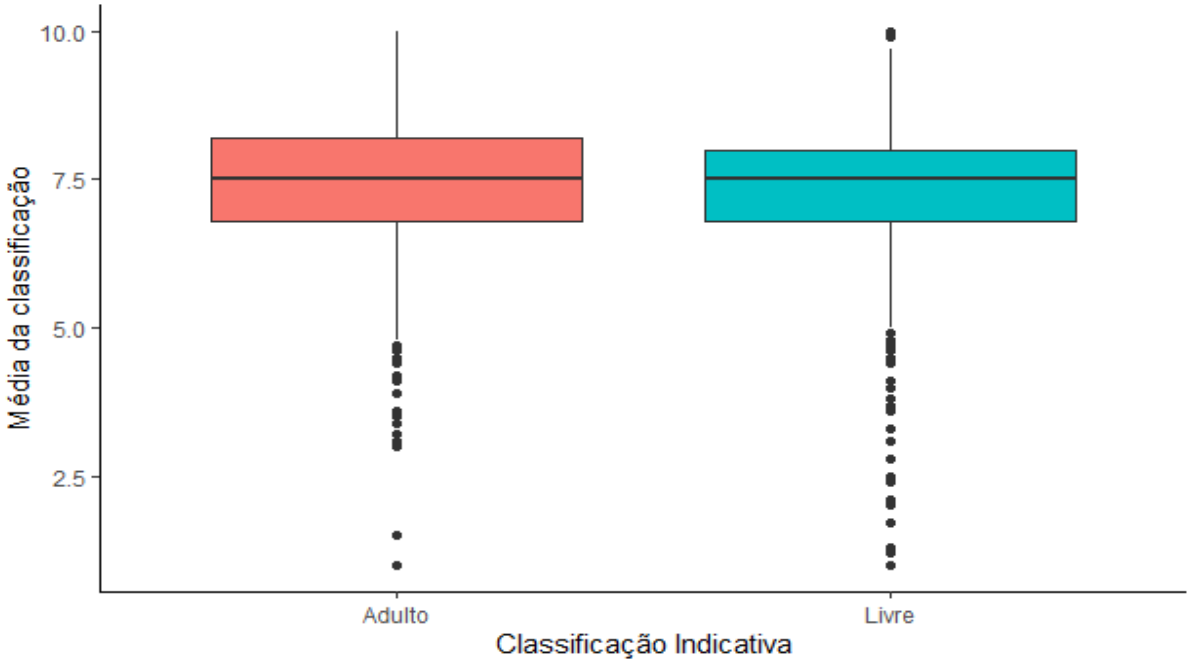


Figura 33: Boxplot relacionando a média da classificação com a classificação indicativa do título.

Através da figura 33, podemos observar que os filmes de classificação adulta (353) têm tanto a mediana quanto as amplitudes totais parecidas com os filmes de classificação livre (1000).

A figura 34, apresenta a dispersão das médias de classificação com os minutos de duração.

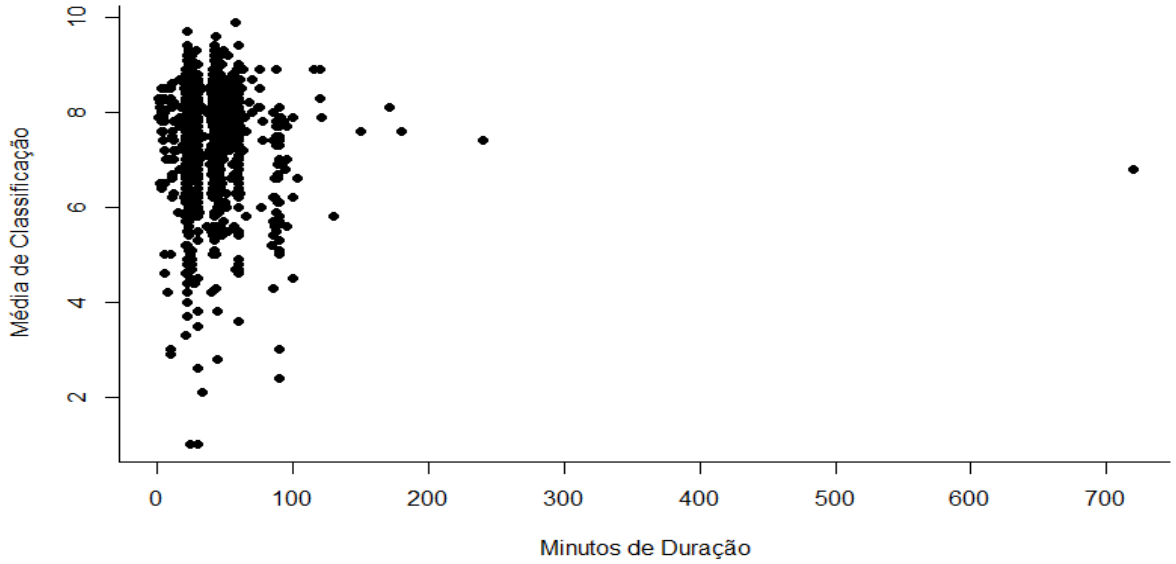


Figura 34: Gráfico de dispersão relacionando a média da classificação com os minutos de duração do título.

Na figura 34, não é possível enxergar nenhuma tendência da média ao longo dos minutos de duração, podemos ver também grandes quantidades de títulos com duração próximos a 0 minutos, o que mostra a inconsistência da variável, além de um título (War of Independence) que apresente 177 minutos de duração.

A figura 35, apresenta a dispersão das médias de classificação com o ano de produção.

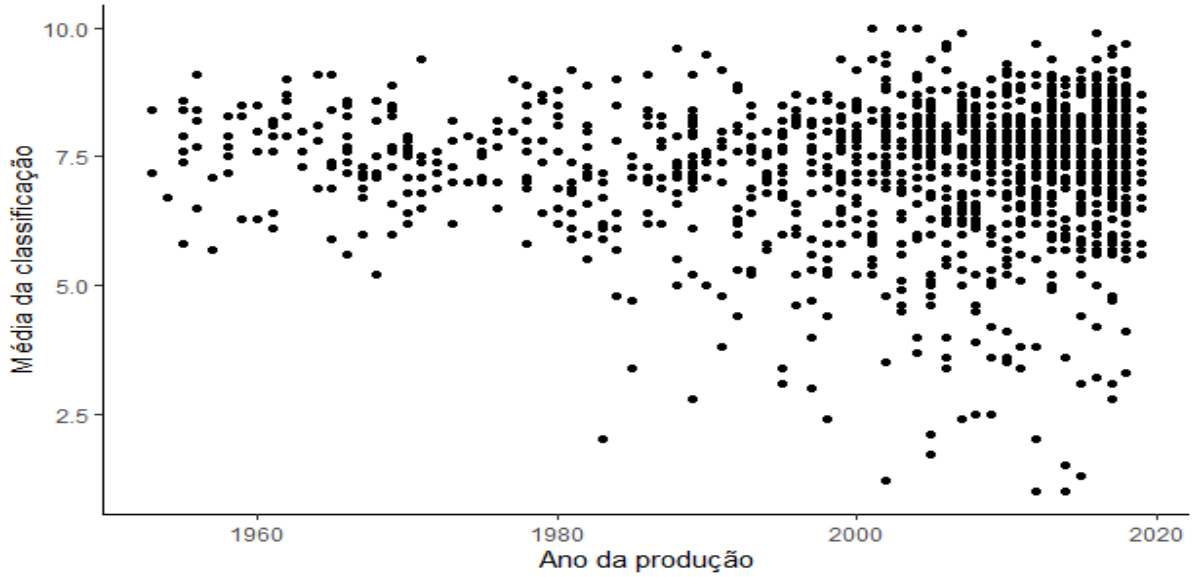


Figura 35: Gráfico de dispersão relacionando a média da classificação com o ano da produção do título.

Na figura 35, é possível ver que ao longo dos anos a dispersão das médias aumenta, assim como a quantidade de títulos avaliados. Além disso, não é possível destacar nenhuma tendência na média ao longo dos anos.

A figura 36, apresenta o Boxplot das médias de classificação com a originalidade.

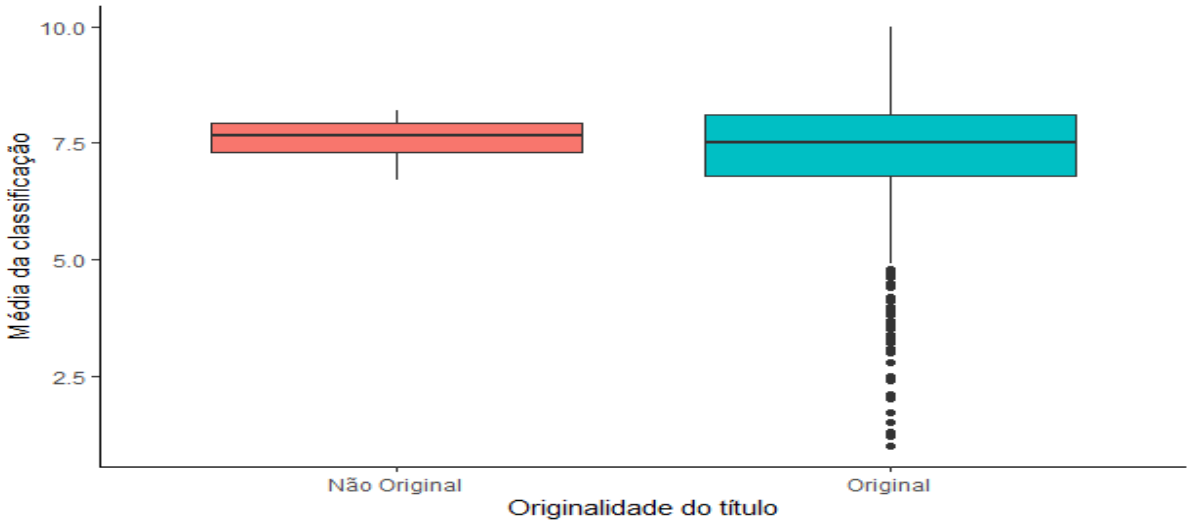


Figura 36: Boxplot relacionando a média da classificação com a originalidade de título.

Na figura 36, podemos notar que visualmente a mediana das duas categorias são parecidas e a dispersão dos títulos originais é maior. No entanto, é importante destacar que os títulos originais aparecem em maior quantidade nos dados, uma vez que 6 títulos são classificados como não originais e 1347 como originais.

A figura 37, apresenta o Boxplot das médias de classificação com a variável gênero.

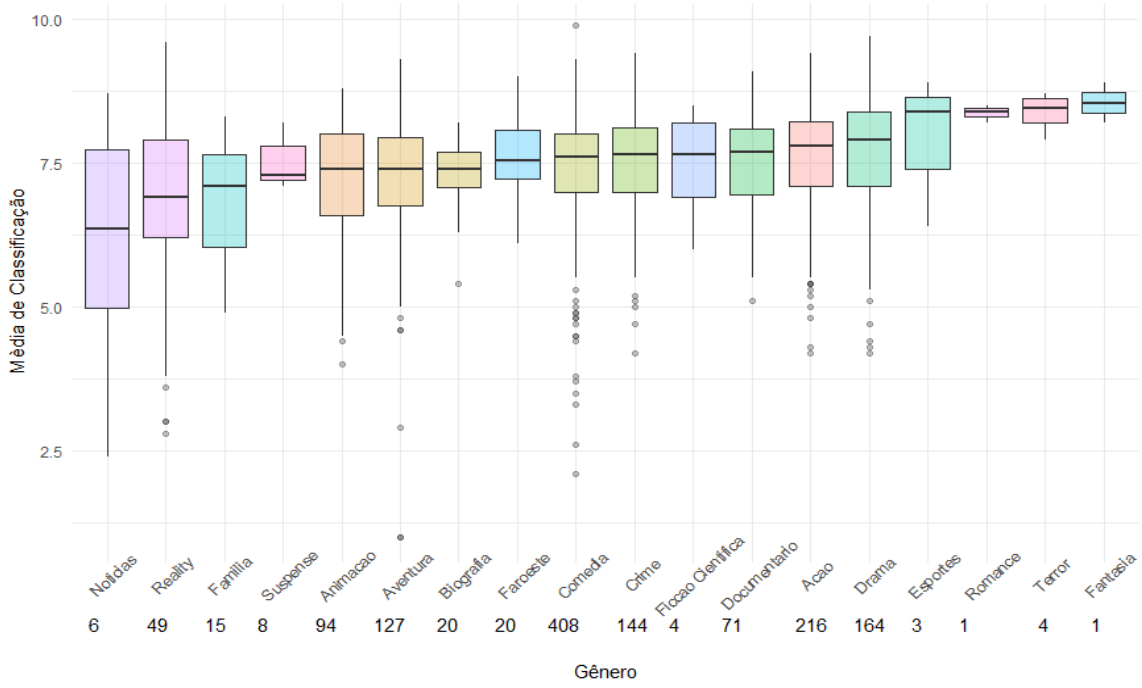


Figura 37: Boxplot relacionando a média de classificação e o gênero do título, com o valor da quantidade em cada grupo.

Através da análise da figura 28, podemos destacar que: visualmente os gêneros fantasia e notícias apresentam respectivamente a maior e menor mediana na amostra. A análise da dispersão em cada nível da variável gênero fica comprometida, devido ao desbalanceamento da quantidade de ocorrência de cada nível (por exemplo: romance e fantasia aparecem somente uma vez, enquanto comédia possui 408 ocorrências).

4.3 – Resultados dos Modelos Ajustados

Os ajustes foram feitos segundo as especificações acima, alguns pontos foram decisivos para o comprometimento da qualidade do resultado.

No caso dos modelos para os grupos curtas, seriados e programas de tv, o algoritmo de estimação (máxima verossimilhança) falhou na estimação de um dos coeficientes. Sendo assim, a matriz de Informação de Fisher apresenta problemas de singularidade, comprometendo a estimação dos erros padrão dos coeficientes estimados. Dessa forma, a avaliação da qualidade do ajuste ficou comprometida.

Uma das hipóteses levantadas para essa falha na estimação seria a multicolinearidade, posto isto, novas tentativas foram feitas considerando diferentes combinações das covariáveis, mas em todas estas a qualidade do ajuste não foi satisfatória.

Para o ajuste do grupo filmes, o problema encontrado foi a adequação da distribuição do componente aleatório, já que o gráfico envelope apresenta grande quantidade de pontos fora das bandas de confiança. O gráfico envelope para este ajuste se encontra no Apêndice.

Considerando os problemas apontados acima, os resultados obtidos não devem ser interpretados, uma vez que a confiabilidade dos resultados está seriamente comprometida pela violação das suposições do MLG.

4.3.2 – Teste Paramétricos e Não Paramétricos

É importante lembrar que os fatores podem ter algumas relações entre si, considerando a hipótese de haver interação entre eles. Para tentar enxergar se há o efeito de interação, foi feito um gráfico (figura 38), relacionando o gênero e o tipo ao longo da média de classificação. Se não houvesse interação, deveria observar o comportamento do tipo de título em uma linha paralelo. Visualmente acredita-se que há interação, portanto, tal efeito foi inserido no modelo.

Desta forma, podemos observar na figura 38, que há interação entre as variáveis gênero e tipo do título.

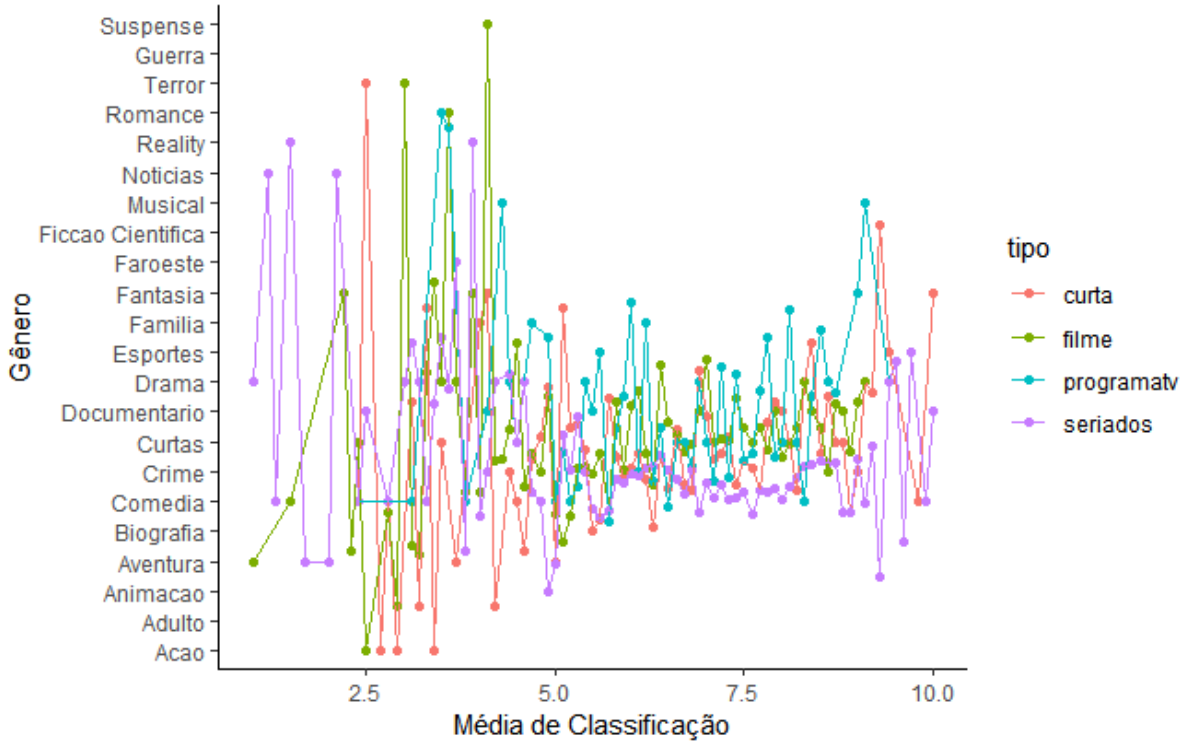


Figura 38: Gráfico de interação entre gênero e tipo do título.

A fim de validar as suposições deste modelo, foi utilizado gráficos de resíduos (figura 39) e testes como de Shapiro Wilk (para verificar normalidade) e Levene (para verificar homocedasticidade) (WICKHAN, 2019, p. 213).

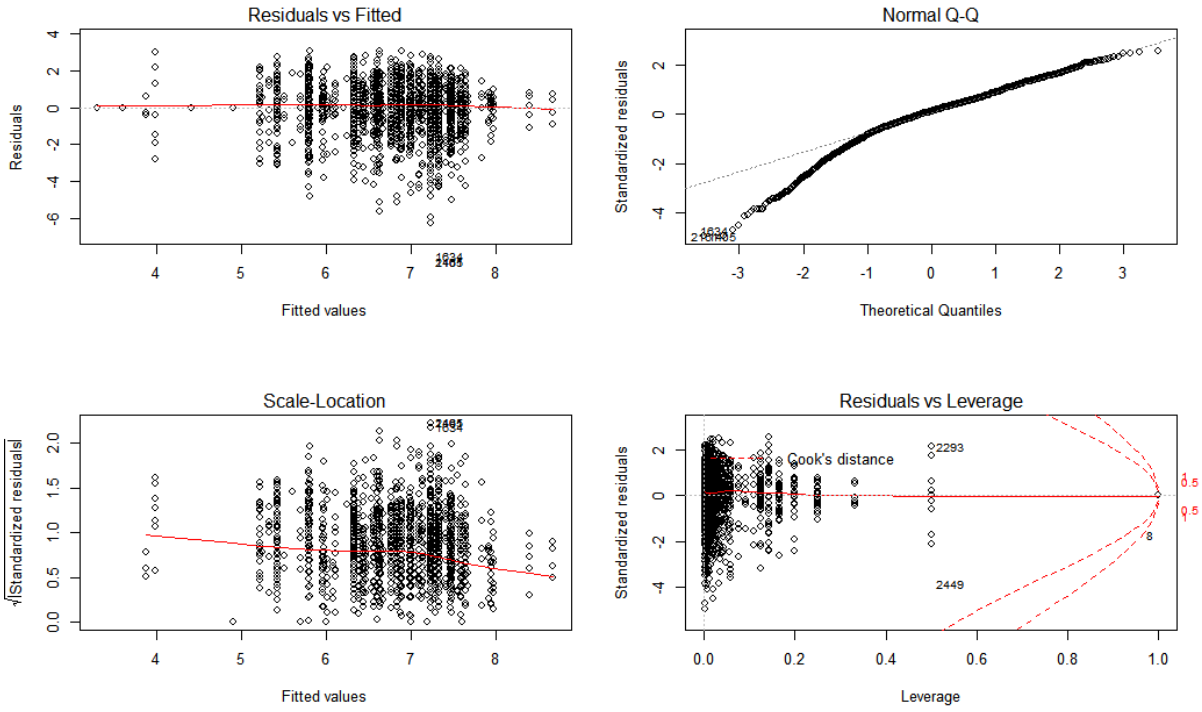


Figura 39: Gráficos de resíduos do modelo.

Tabela 7: Resultado dos testes da análise de resíduos.

Teste	P-Valor
Shapiro Wilk	<0,05
Levene	<0,05

Tendo em vista a análise dos gráficos expostos na figura 39, podemos acreditar que visualmente as suposições do modelo parecem violadas. Os resultados dos testes apresentados na tabela 7, mostram que, ao nível de significância de 5%, as suposições de normalidade e homocedasticidade dos resíduos foram descumpridas.

Diante disto, uma alternativa é considerar o ajuste utilizando alguma técnica não paramétrica. Porém, o desbalanceamento dos dados dificulta que uma única análise seja feita considerando os dois fatores (e assim perde-se a interpretação que levaria em conta o efeito de interação).

A técnica escolhida para comparar os níveis de cada um dos fatores é o teste de Kruskal-Wallis (Teste de Kruskal-Wallis, 2020).

Para comparar a mediana de avaliação entre os níveis da variável gênero as hipóteses são as seguintes:

$$\{H_0: \beta_1 = \beta_2 = \cdots = \beta_{22}\ H_1: \beta_i \neq \beta_j\ para\ pelo\ menos\ um\ par\ i,j\ ;\quad i = 1, \dots, 22\ ; j = 1, \dots, 22$$

O teste de Kruskall Wallis obteve o seguinte resultado:

Tabela 8: Resultado do teste de Kruskall Wallis para a variável gênero.

Graus de Liberdade	P-valor
21	<0,05

Com o resultado obtido (tabela 8), rejeitamos a hipótese nula de que a mediana é igual dentro de todos os níveis da variável gênero. A rejeição de tal hipótese implica que há pelo menos um par de gêneros que apresentam mediana diferentes, tornando necessária a realização de comparações múltiplas, que foram realizadas por meio do teste de Dunn, utilizando a correção de Bonferroni (a fim de controlar a probabilidade de cometer pelo menos um erro do tipo I).

A partir do teste de Dunn, com um nível de significância de 5% foi possível observar que o gênero notícias, tem média de avaliação diferente dos gêneros: ação, animação, comédia, crime, esporte, faroeste, ficção científica e musicais. Enquanto

o gênero terror difere dos gêneros: ação, animação, comédia, crime, documentários, drama, esporte, ficção científica e musical. O gênero adulto tem média de avaliação diferente dos gêneros: musical e crime. Por fim, o gênero documentário difere na média dos gêneros: aventura, curtas e drama. A tabela com os resultados e p-valores para todas as combinações foi inserida no Apêndice.

Para comparar a mediana de avaliação entre os níveis da variável tipo as hipóteses são as seguintes:

$$H_0: \tau_1 = \tau_2 = \cdots = \tau_4 \; H_1: \tau_i \neq \tau_j \; para \; pelo \; menos \; um \; par \; i, j \; ; \quad i = 1, \ldots, 4 \; ; j = 1, \ldots, 4$$

O teste de Kruskall wallis obteve o seguinte resultado:

Tabela 9: Resultado do teste de Kruskall Wallis para a variável tipo

Graus de Liberdade	P-valor
3	<0,05

Com o resultado obtido na tabela 9, rejeitamos a hipótese nula de que a mediana é igual dentro de todos os níveis da variável tipo. A rejeição de tal hipótese implica que há pelo menos um par de tipos que apresentam mediana diferentes, para investigar quais são os tipos de títulos que diferem entre si, foi realizado novamente o teste de Dunn, utilizando a correção de Bonferroni.

A partir do teste de Dunn, com 5% de significância, é possível concluir que, não há diferença nas medianas das avaliações dos programas de tv e dos curtas. É possível concluir também que todas as outras combinações entre os tipos que foram especificados são diferentes entre si.

5 – CONCLUSÃO

O presente trabalho teve como eixo central, o estudo sobre os dados de entretenimentos (curtas, longas, filmes para cinema, filmes para tv, programas de televisão, comerciais para televisão e especiais), contidos em vários arquivos TSV extraídos do site IMDb. Através deste estudo, conseguimos integrar todos dados, gerando um único arquivo que serviu como fonte de dados. Com ele, utilizamos técnicas de amostragem para definir, calcular e colher uma amostra.

Na conclusão do trabalho, seguimos as etapas: extração dos arquivos, concatenação dos mesmos, utilizando a ferramenta Python, estudo da melhor metodologia para extração de uma amostra, geração da amostra com a ferramenta Python e modelagem com a ferramenta R e por fim, elaboração das conclusões.

Para melhor análise, percebemos a necessidade de separar os dados por tipos de títulos de entretenimento e agrupá-los em 4 categorias distintas, com os atributos semelhantes. Por exemplo: para títulos do tipo filme as características em geral são de entretenimentos para serem lançados em cinemas com uma duração mais longa, diferentemente dos títulos do tipo curtas, que são mais para programas de tv com o tempo de exibição menor. Já os programas de tv são mais voltados para família e não são exibidos em cinemas, desta forma, muda o público alvo. Em resumo, os grupos com qualidades similares foram analisados em conjunto para não diferenciar as características do público alvo de quem votou, com isso obtivemos melhores resultados.

Como apresentado, os resultados obtidos no modelo MLG com distribuição gama não foram conclusivos, uma vez que o ajuste do modelo não pôde ser validado, já que apresentou problemas de singularidade, multicolinearidade e subdispersão dos dados.

Desta forma, seguimos com os métodos paramétricos ANOVA, com o objetivo de entender se existem diferença na média de avaliação dos grupos, dentre os tipos de títulos (curta, filme, seriado e programa de tv) e gêneros (ação, adulto, animação, aventura, biografia, comedia, crime, curtas, documentário, drama, esportes, família, fantasia, faroeste, ficção cientifica, musical, notícias, reality, romance e terror), já que o método ANOVA é usado para 3 ou mais grupos dentre de cada variável. Mesmo não tendo a distribuição normal, que é uma das características das análises paramétricas, neste caso o tamanho da amostra é suficientemente grande para que a distribuição seja usada por aproximação. O teste mostrou que existem diferenças entre as variâncias dos grupos e não satisfizeram as suposições necessárias, fazendo com que se tornasse necessário o uso de técnicas não paramétricas, pois as técnicas não paramétricas têm menos suposições para serem validadas do que as paramétricas.

A partir das técnicas não paramétricas utilizadas, foi possível fazer um primeiro teste e identificar que havia diferença tanto entre os gêneros quanto entre os tipos de títulos. Como o primeiro teste, apenas detectava se havia alguma diferença ou não, surgiu o interesse em saber onde estavam as diferenças apontadas por ele. Ao realizar um segundo teste (teste de Dunn) foi possível identificar quais grupos da variável gêneros ou tipos eram diferentes entre si, como foi mostrado nas tabelas de Dunn.

O grande problema apresentado neste estudo, foi justamente a base de dados, pois ela apresenta dados incoerentes, sendo necessários tratamentos como por exemplo, a retirada de valores nulos, minutos de filmes com valores zerados, votação em alguns títulos menores que 5 votos, números de episódios ou temporadas com valores zeros, enfim, mesmo trabalhando na base ainda não conseguimos ter resultados satisfatórios.

Mesmo assim, acreditamos que o ganho neste trabalho foi justamente poder aplicar a grande parte das técnicas da estatística como a amostragem, análise descritiva, regressão não linear simples e múltiplas, técnicas paramétricas e não paramétricas.

Fica a lição que, nem todos os dados são adequados para uma correta interpretação estatística, podendo assim, como neste trabalho, apresentar inferências sem resultados coerentes. Mas, no caso deste trabalho, podemos destacar os passos seguidos para chegar a um resultado desejado. Desta forma, ele ajudaria como um ponto de partida para trabalhos futuros e ainda, aqueles que quiserem dar sequência ao estudo sobre entretenimentos seguiriam as etapas descritas nas seções 3 e 4, acrescentando somente novos dados, dados externos, a fim de complementar as informações para conseguir gerar resultados mais satisfatórios.

6 - REFERÊNCIAS

Teste de Kruskal-Wallis. Disponível em: https://www.inf.ufsc.br/~vera.carmo/Testes_de_Hipoteses/Teste_Nao_parametrico_Kruskal-Wallis.pdf. Acesso em: 5 mar. 2020.

Teste de Kruskal-Wallis (1952). Disponível em: https://edisciplinas.usp.br/pluginfile.php/1064940/mod_resource/content/2/Teste%20de%20KW.pdf. Acesso em: 5 mar. 2020.

TSV, Extensão de arquivos. Disponível em: https://www.reviversoft.com/pt/file-extensions/tsv. Acesso em: 22 jul. 2019.

AIC E BIC, Portal Action. Disponível em: http://www.portalaction.com.br/analise-de-regressao/2715-aic-e-bic. Acesso em: 22 set. 2009.

COSTA, Marcelo Azevedo. **Introdução aos Modelos Lineares Generalizados - MLG**. 1. ed. 2010

HUANG, Jae. **IMDb Data - Machine Learning (predicting movie gross**). Acesso em 15 abril 2020. Disponível em: "https://medium.com/@jae.huang111/imdb-data-machine-learning-predicting-movie-gross-2113513513bb".

IMDb, Internet Movie Database. Disponível em: https://m.imdb.com. Acesso em: 22 jul. 2019.

JUNIOR, Fabio Alves. **Introdução a Linguagem de Programação Python**: 1. ed. São Paulo: Editora Ciência Moderna, 2018.

OLIVEIRA, Uanderson Rebula de. **Estatística I (para leigos): aprenda fácil e rápido**. 1. ed. São Paulo: Editora Saraiva Publique-se, 2007**.**

KISH, L. Survey Sampling. Nova Iorque: John Wiley; Sons.

PESSOA, Djalma; SILVA, Pedro Nascimento. Análise de Dados Amostrais Complexos. Disponível em: https://djalmapessoa.github.io/adac/index.html. Acesso em: 10 ago. 2020.

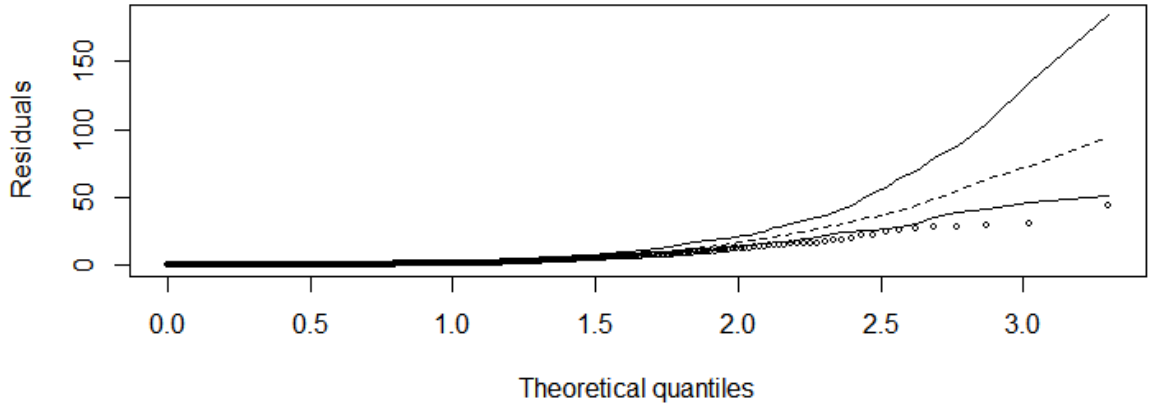
PINHEIRO, Conrad Elber. **Estatística Descritiva**. 1. ed. São Paulo: Editora Saraiva, 2017**.**

RUMSEY, Deborah J. **Estatística Para Leigos**. 2. ed. São Paulo: Editora Alta Books, 2011**.**

SARKAR, Tirthajyoti. Step-by-step guide to build your own ‘mini IMDb’ database**.** Acesso em 15 abril 2020.Disponível em: "https://towardsdatascience.com/step-by-step-guide-to-build-your-own-mini-imdb-database-fc39af27d21b".

WICKHAN, Hadley; GROLEMUND, Garrett. **R Para Data Science**. 1. ed. São Paulo: Editora Alta Books, 2019**.**

7 - APÊNDICE A: GRÁFICO RESÍDUOS DO MODELO DE REGRESSÃO GAMMA DA CATEGORIA FILMES.



8 - APÊNDICE B: TABELAS COM OS RESULTADOS DO TESTE DE DUNN PARA O GÊNERO DO TÍTULO.

	Acao	Adulto	Animacao	Aventura	Biografia	Comedia	Crime
Adulto	0.094						
Animacao	1.000	0.174					
Aventura	0.395	1.000	1.000				
Biografia	1.000	1.000	1.000	1.000			
Comedia	1.000	0.357	1.000	1.000	1.000		
Crime	1.000	0.014	1.000	0.048	1.000	1.000	
Curtas	1.000	1.000	1.000	1.000	1.000	1.000	0.254
Documentario	1.000	0.001	1.000	0.001	1.000	0.177	1.000
Drama	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Esportes	1.000	0.080	1.000	0.469	1.000	1.000	1.000
Familia	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Fantasia	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Faroeste	1.000	0.257	1.000	1.000	1.000	1.000	1.000
Ficcao Cientifica	1.000	0.519	1.000	1.000	1.000	1.000	1.000
Musical	1.000	0.011	1.000	0.078	1.000	1.000	1.000
Noticias	0.011	1.000	0.013	0.357	0.086	0.026	0.003
Reality	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Romance	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Terror	0.000	1.000	0.000	0.580	0.115	0.000	0.000
Guerra	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Suspense	1.000	1.000	1.000	1.000	1.000	1.000	1.000

	Crime	Curtas	Documentario	Drama	Esportes	Familia	Fantasia	Faroeste
Crime								
Curtas	0.254							
Documentario	1.000	0.031						
Drama	1.000	1.000	0.029					
Esportes	1.000	0.423	1.000	1.000				
Familia	1.000	1.000	1.000	1.000	1.000			
Fantasia	1.000	1.000	1.000	1.000	1.000	1.000		
Faroeste	1.000	1.000	1.000	1.000	1.000	1.000	1.000	
Ficcao Cientifica	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Musical	1.000	0.137	1.000	1.000	1.000	1.000	1.000	1.000
Noticias	0.003	0.777	0.001	0.051	0.002	0.482	1.000	0.007
Reality	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Romance	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Terror	0.000	1.000	0.000	0.004	0.004	1.000	1.000	0.005
Guerra	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Suspense	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

	Faroeste	Ficcao Cientifica	Musical	Noticias	Reality	Romance	Terror	Guerra
Ficcao Cientifica	1.000							
Musical	1.000	1.000						
Noticias	0.007	0.010	0.001					
Reality	1.000	1.000	1.000	0.290				
Romance	1.000	1.000	1.000	1.000	1.000			
Terror	0.005	0.032	0.000	1.000	0.876	1.000		
Guerra	1.000	1.000	1.000	1.000	1.000	1.000	1.000	
Suspense	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

9 - APÊNDICE C: TABELA COM OS RESULTADOS DO TESTE DE DUNN PARA O TIPO DE TÍTULO.

	curta	filme	programatv
filme	0.000		
programatv	0.501	0.000	
seriados	0.000	0.000	0.000