

**ALEXANDRE DE CASTRO SANTOS**

**Alexandre\_mc10@hotmail.com – 20192520038**

**LARYSSA GUIMARÃES**

**raryssa.guimaraes@gmail.com - 20192520024**

**TRABALHO DE ESTATÍSTICA**

Neste projeto iremos observar:

O comportamento numérico do

cinema durante os anos, e como

esse comportamento pode ser

identificado através de uma

Base de dados.

**Trabalho.**

**Base de dados - Filmes**

Rio de Janeiro

23 de setembro de 2021

* **INTRODUÇÃO**

O surgimento do cinema é datado em 1985, e em 1923 através do intelectual italiano Ricciotto Canudo, foi considerado a sétima arte. Com os passar dos anos a arte do cinema passou a virar um mercado. Fazendo com que a produção de obras cinematográficas crescesse vertiginosamente, e assim se tornando um mercado que movimenta até os dias de hoje bilhões em receita.

O tema abordado será, a construção de uma base de dados que mostre estatisticamente a evolução ou involução do cinema mundial através das estatísticas utilizando o programa “R data”. Utilizaremos como parâmetro uma base de dados que nos fornece notas de usuários cadastrados no site, e que realizam a avaliação de filmes no site. Sendo assim uma avaliação totalmente popular, onde os números estatísticos irão dizer a realidade sobre cada filme, ano, gênero, etc...

* **OBJETIVO**

Podemos dizer que o objetivo do trabalho a ser apresentado é, comparação aprofundada sobre qualidade e aspectos em específicos do cinema mundial durante os anos. Através da base de dados iremos desvendar os segredos do cinema durante os anos, qual gênero de maior sucesso? Qual ano de maior gloria do cinema? Comparado à períodos históricos o cinema teve alteração em suas representações artísticas? Essas e outras perguntas iremos responder ao longo do processo de análise das estatísticas.

Com a análise dos dados devemos chegar em algumas conclusões, durante todos os anos do cinema houve fórmula do sucesso? Um gênero específico uma metragem específica, poderemos também ver isso na base de dados.

* **METODOLOGIA**

O método científico utilizado foram os de uma pesquisa quantitativa, que utiliza diferentes técnicas estatísticas para qualificar informações para um determinado estudo. Nesta pesquisa em específico, foi utilizado o resultado de enquetes de opiniões presente no site de avaliações (Será citado oportunamente), e com o resultado dessas pesquisas montado uma pesquisa quantitativa sobre a relevância de determinado filme. Com o resultado desta pesquisa numérica, através dos métodos de pesquisas quantitativas para tratamento dos dados, pois ela prioriza os dados numéricos do estudo proposto, podemos avaliar completamente e tirar um resultado plausível dessas informações.

Os princípios básicos para uma pesquisa quantitativa são; Tem como objetivo medir (mensurar através de dados numéricos); não há espaço para ambiguidades; Análise de respostas objetivas; Uso de questionários de múltipla escolha, entrevistas individuais, etc.; Métodos de coleta inflexível e previamente estruturado; Pesquisador-observador (não pode interferir nos resultados); Resultados apresentados em gráficos, tabelas e índices (menos espaço para interpretações subjetivas). Podemos observar que todos os critérios foram respeitados e seguidos para realização do estudo sobre a base de dados.

* **DESCRIÇÃO DETALHADA DA BASE DE DADOS**

O primeiro passo para início da realização foi a escolha do tema que iria seguir para realização da pesquisa quantitativa, escolhi tema filmes. Logo passei para a fase de procura da base de dados, achei sem maiores dificuldades as do site IMDB o maior site de quando o assunto é avaliação cinematográfica. O site em questão deixa público alguma de suas *database* para uso livre, algumas com características muito ruins para se trabalhar e outras muito boas, de cara, percebi que teria problemas para realização da ideia que tive, quando a base de dados com as notas apenas continha o *contente ID* (nome do filme para o site ex: tt10272386), e as notas com número de avaliadores, já outra base de dados separada, continha os dados complementares (ano, nome, gênero, subgênero, etc...). Primeira ação que tive que tomar foi o cruzamento das duas bases de dados, no excel, através de uma fórmula (procv) ligando as informações através do *contente ID* que era o único dado igual entre todas as *database*.

Partindo desse momento, decidi filtrar os dados, afinal haviam um 1 milhão e 600 mil linhas de dados. Como decidi que o tema da pesquisa seria filme filtrei apenas para categoria “*Movies*”, me deixando com uma base de avaliação de 180 mil analisados. Porém, ainda estava faltando algo, e decidi por retirar todos os filmes que haviam menos de mil avaliações contabilizadas, para fins de pesquisa quantitativa, filmes cujo nota é de 9.9 e tem apenas 5 avaliações não mostra os dados em sua total realidade, o que me deixou com o número final de 21.182 linhas de informação e aproximadamente 9 colunas de complementares.

Agora nossa planilha conta com mais de 20 mil linhas de informação prontas para serem analisadas. Temos todos os campos necessários para uma avaliação que vai nos dizer os comportamentos dos filmes durantes os anos, como por exemplo campos de gênero e subgênero, ano de lançamento ou nota, sendo esses alguns dos exemplos, podemos traçar de cabeça diversas combinações possíveis para nos mostrar a realidade com base em uma ampla pesquisa opinativa e não obrigatória sobre o resultado/desempenho dos filmes.

Na base de dados são apresentados os seguintes campos de dados:

|  |
| --- |
| **ID** |
| **Nota** |
| **Votos** |
| **Estilo** |
| **Ano** |
| **Minutos** |
| **Gênero** |
| **Subgênero** |
| **Subgênero2** |

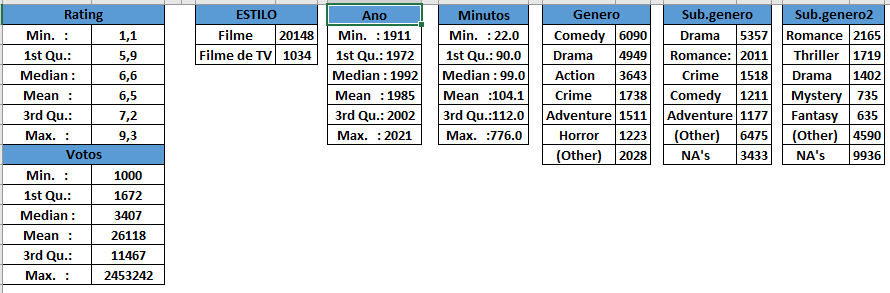
Onde ID é o nome do filme para localização direta no site, exemplo: (https://www.imdb.com/title/tt0110912/) onde “tt011091” seria o ID do filme “*Pulp Fiction*”. Nota é o campo mais interessante para análise, pois é uma média avaliativa das notas dadas por todos os pesquisados, ou seja, um consenso numérico sobre qual a média de execução cinematográfica o filme tem (cada pesquisado possui o seu próprio critério avaliativo) sendo assim, quando mais próximo a 0 pior é esta execução e quando mais próximo a 10 melhor é essa execução, por cara pesquisado possuir um critério diferente o resultado obtido é necessário levar em consideração pois se agrada 90% dos avaliadores que não tem correlação preliminar nenhuma, provavelmente este filme foi bem executado.

Votos é quantidade numérica de votos obtidos para aquele filme e específico, ou seja, quantidade de pessoas que avaliaram aquele filme (independente da nota dada por cada um), estilo se é um filme de produtoras independente ou filmes feitos por televisões, campo ano é a data do lançamento, que se subdivide em duas partes quando forem analisadas as estatísticas no R graficamente, para os filmes anteriores de 1966, são considerados da primeira fase do cinema (cinema mudo, preto e branco ou simples), já os considerados da segunda fase, os mais novos, partindo de 1967 (primeiras grandes produções).

Minutos, apenas a quantidade de tempo de tela que filmes apresentam, e finalizando com gêneros e subgêneros, sendo os determinantes para a classificação dos filmes, podemos exemplificar com filmes de guerra, em sua maioria são considerados dramas e o tema “guerra” é mais presente na parte de subgênero, tornando mais interessante ainda a análise dos motivadores para tal fato.

* **ANÁLISE DE RESULTADOS**

Partindo desse ponto, iremos analisar todos os resultados obtidos com a base de dados analisada.

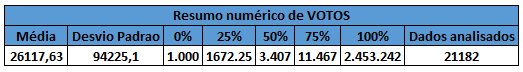


Fonte: Base de dados – Filmes

Começando pelo resumo numérico de toda a base de dados, pode-se analisar diversas estatísticas que nos ajudam a interpretar a evolução do cinema durante os anos. No resumo numérico podemos destacar já algumas estatísticas interessantes, que não estariam presentes graficamente, como por exemplo mais de 90% dos filmes se encontra sobre os gêneros, comédia, drama, ação, crime, aventura ou horror. Em um breve cruzamento de dados com a tabela de sub gêneros, estatisticamente falando a combinação que mais se repete seria os filmes de comédia dramática.

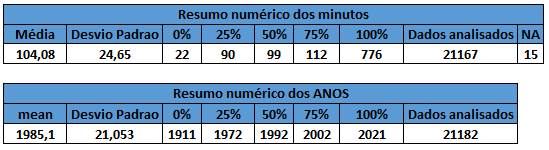
Outra estatística a ser analisada seria a nota, a mediana é 6,6 e a média das notas 6,5 com um desvio padrão de 1.11 que nos ajuda a verificar a média de dispersão dos dados, por esse desvio ser baixo é próximo da média. Porém quando interligada ao número de votos, é observado algo totalmente inverso, com a mediana dos votos sendo 3407 votos, e a média sendo 26118, o desvio padrão nesse caso é muito grande que mostra que os dados estão espalhados por uma ampla gama de valores. Isso acontece pois na base de dados existem menos filmes que fizeram sucesso nas suas avaliações como por exemplo o primeiro que obteve 2.4milhões de votos, em compensação de outros milhares que tem o mínimo de notas para entrar na base de dados avaliativa (mil votos), fazendo com que a média e mediana se tornem pouco confiáveis.

Abaixo o gráfico mostra a variável VOTOS de uma forma melhor



Fonte: Base de dados – Filmes

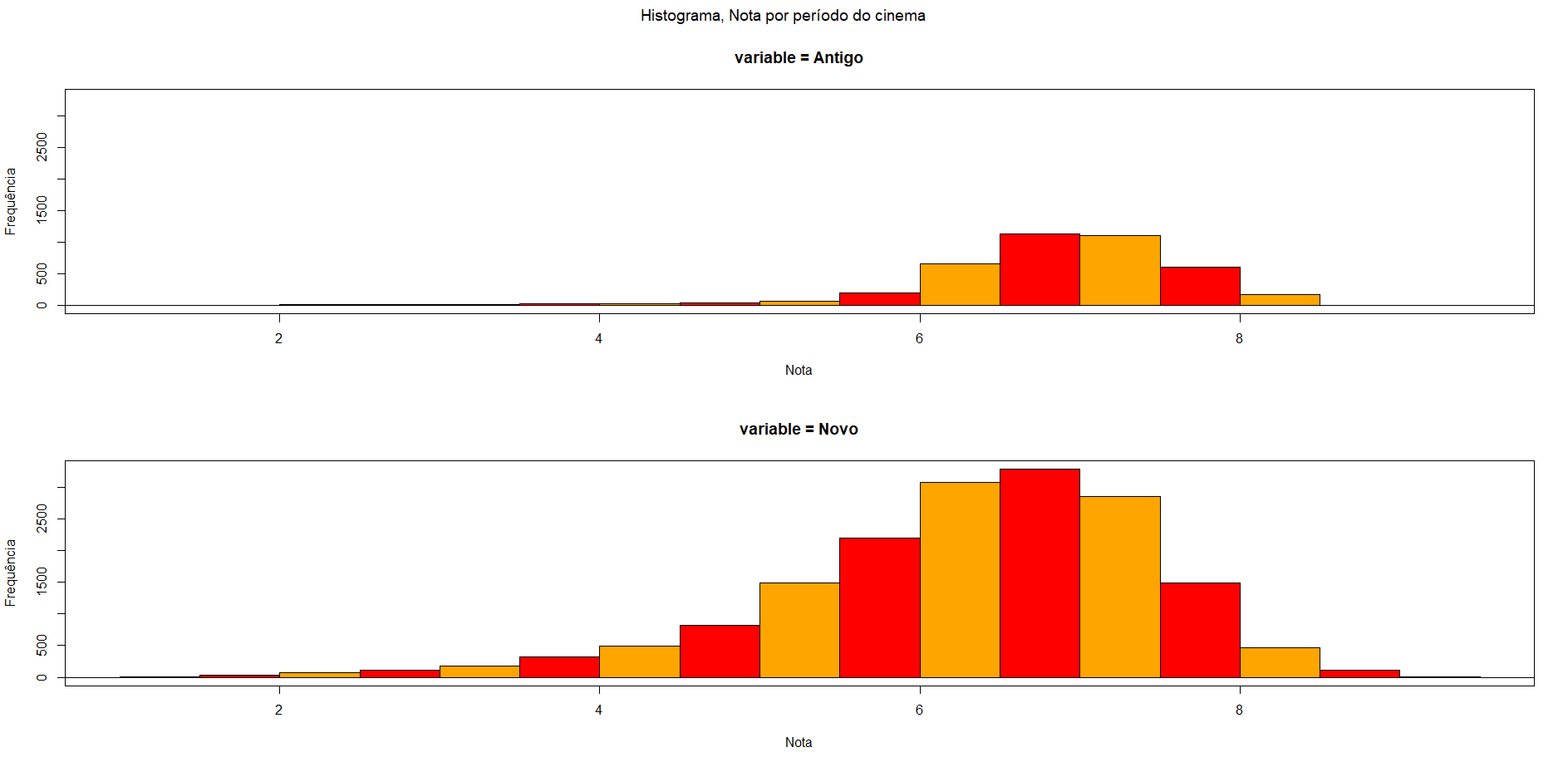
* **Resumo numérico minutos e Ano**



Fonte: Base de dados – Filmes

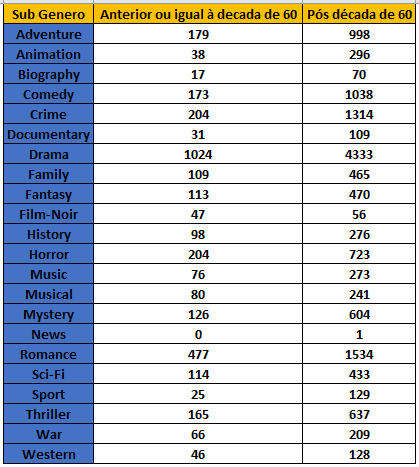
No resumo apresentado pelas tabelas acima, podemos verificar algumas estatísticas muito importantes como a média de minutos de duração dos filmes, que é de 104 minutos, o desvio padrão é relativamente alto para os dados analisados. Partindo do terceiro quartil os filmes começam a ficar com uma duração muito elevada, entre 120 minutos até 770 minutos de duração.

Sobre o resumo numérico dos anos de lançamento, o primeiro filme da nossa base de dados é datado de 1911, e o mais atual sendo do ano de construção dessa análise (2021). Metade dos filmes da nossa base de dados ou seja 10.591 foram lançados até 1992, em um intervalo de 81 anos foram lançados esses 50%, porém do ano de 1992 até o ano de 2021 foi lançada a outra metade dos filmes em um intervalo de 29 anos, quase três vezes menos tempo para lançar a mesma quantidade de filmes. Com essa estatística podemos notar o crescimento dos filmes partindo da década de 90 (Mais precisamente 80), para o que veio a se tornar a indústria que mais movimenta dinheiro durantes muitos anos.



Fonte: Base de dados – Filmes

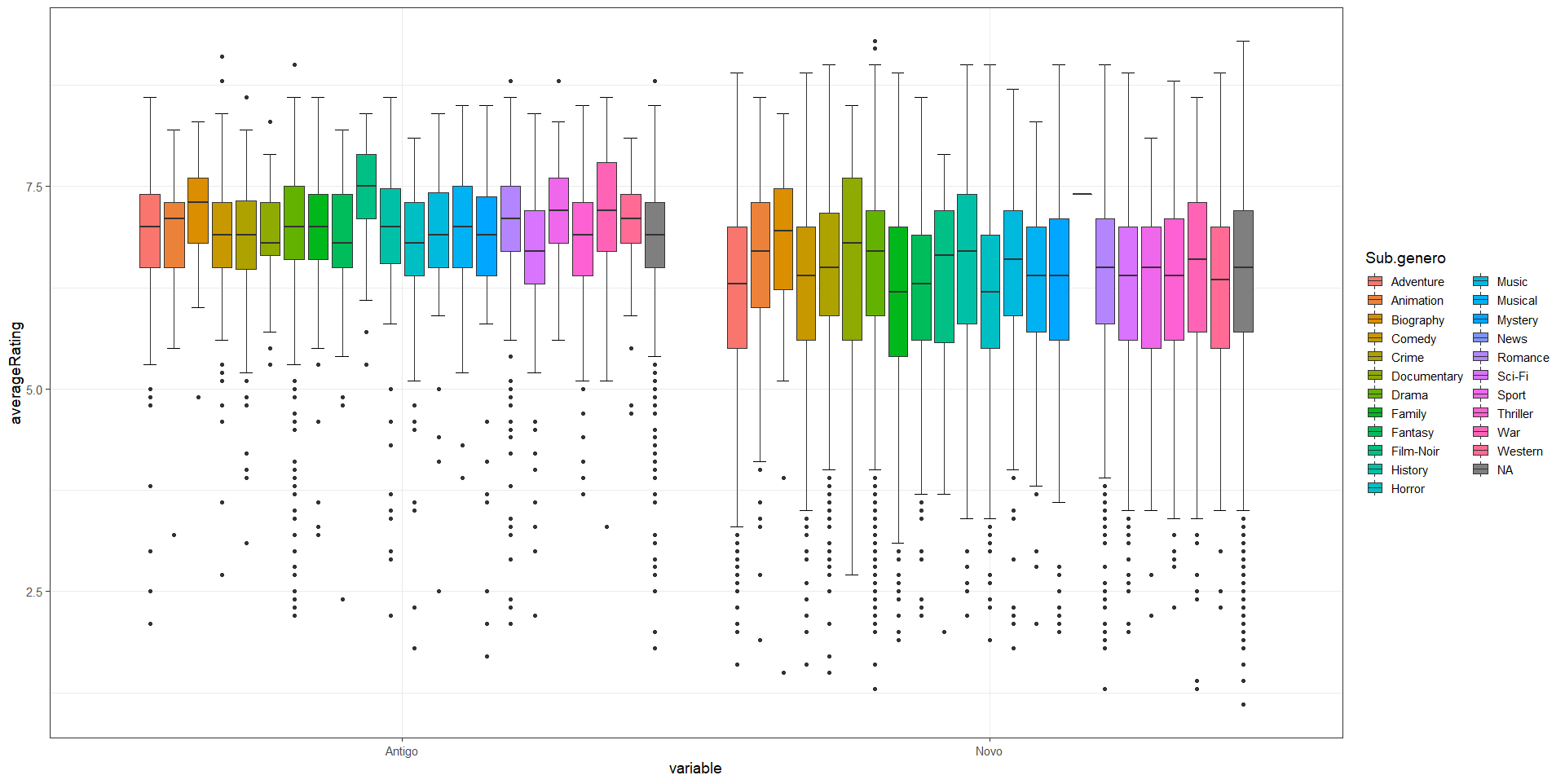
Ainda falando sobre períodos do cinema no gráfico acima, podemos ver a avaliação (nota) dos filmes por períodos do cinema, anteriores à 1966 são considerados antigos, e posteriores à 1967 são os considerados os “novos”. A avaliação que deve ser feita é que, os filmes do denominados período antigo tem menor avaliação no geral quando comparados aos novos. Sendo assim, entende-se que os melhores filmes (com maior ranking de avaliações positivas) foram feitos depois do ano de 1966. Outro tipo de análise pode ser feito, a concentração de notas dos filmes denominados antigos, é concentrado entre 5,75 e 8,75 de nota, com poucos filmes que contém notas baixas, já nos filmes novos vemos uma dispersão maior para notas menores (e até maiores que 8,75). Sendo assim é possível dizer que por mais que haja uma frequência maior de filmes com notas boas no período atual, a uma presença menor de filmes “ruins” no período antigo do cinema.



Fonte: Base de dados – Filmes

No resumo numérico acima, temos ainda usando o mesmo critério da análise anterior, nessa vemos qual utilização de subgênero durante os períodos. Podemos destacar alguns como o gênero guerra, que teve três vezes mais filmes pós anos 70 do que antes nos anos de 60, pois devido ao breve curto período de tempo depois que a guerra acabou ambas as partes ainda estavam tentando entender os acontecidos, isso foi acontecer melhor posterior aos anos 60 onde diversos filmes com subgênero de guerra foram lançados.

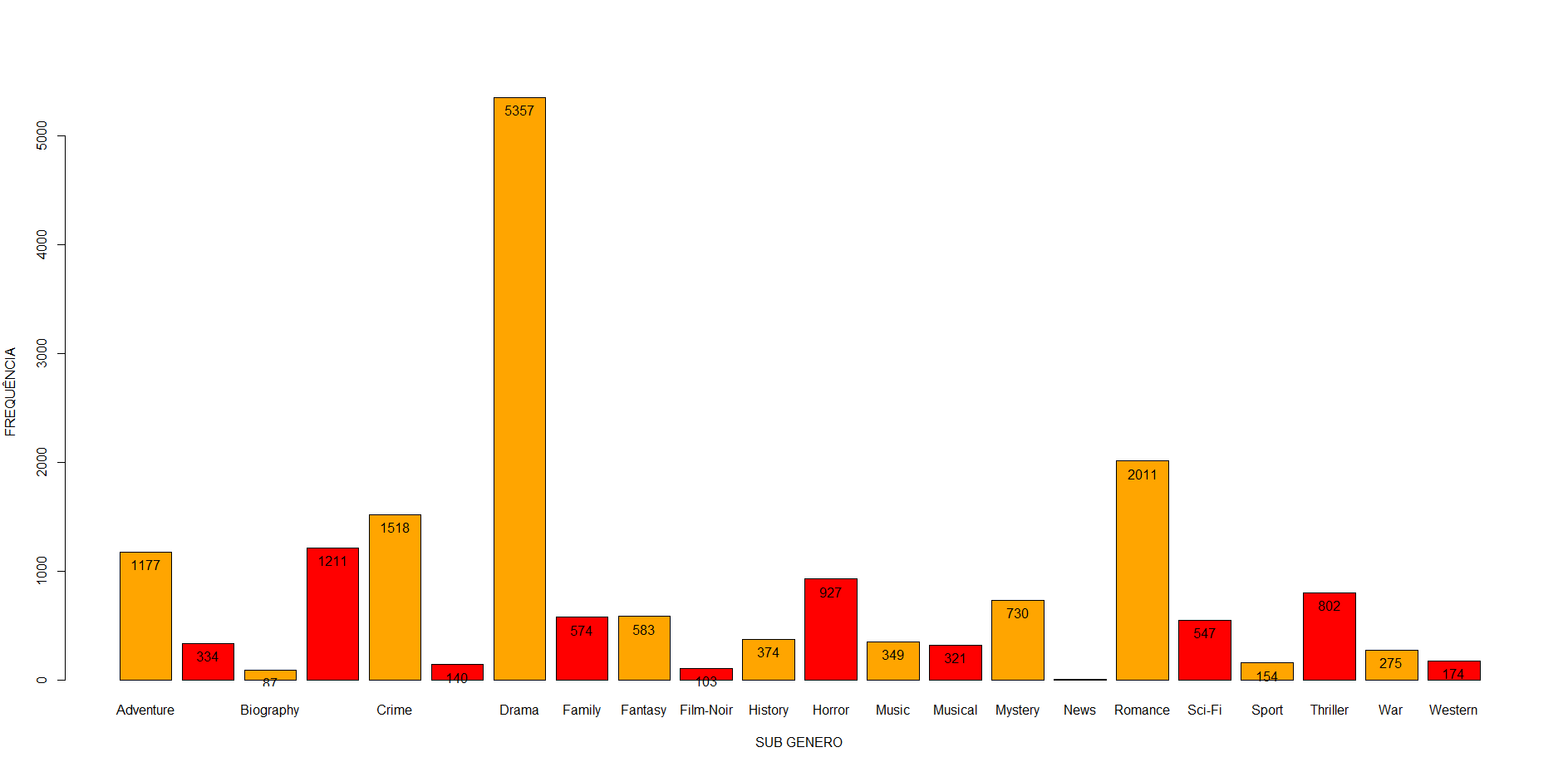
Outro destaque são as animações, que eram muito caras e difíceis de se fazer na época impossibilitando uma produção em massa, hoje com a Disney a distribuição ampliada de suas animações e a evolução tecnológica possibilitou a produção de quase dez vezes mais filmes que os produzidos até o final dos anos 60. Drama é um outro expoente dos filmes, cresceu mais que quatro vezes a produção de filmes com subgênero drama, por diversos motivos um deles é que o gênero drama é um coringa e abrange diversos temas e pode ser usado para descrever qualquer conflito gerado dentro do filme, por esse motivo ele é muito utilizado na descrição dos filmes.



Fonte: Base de dados – Filmes

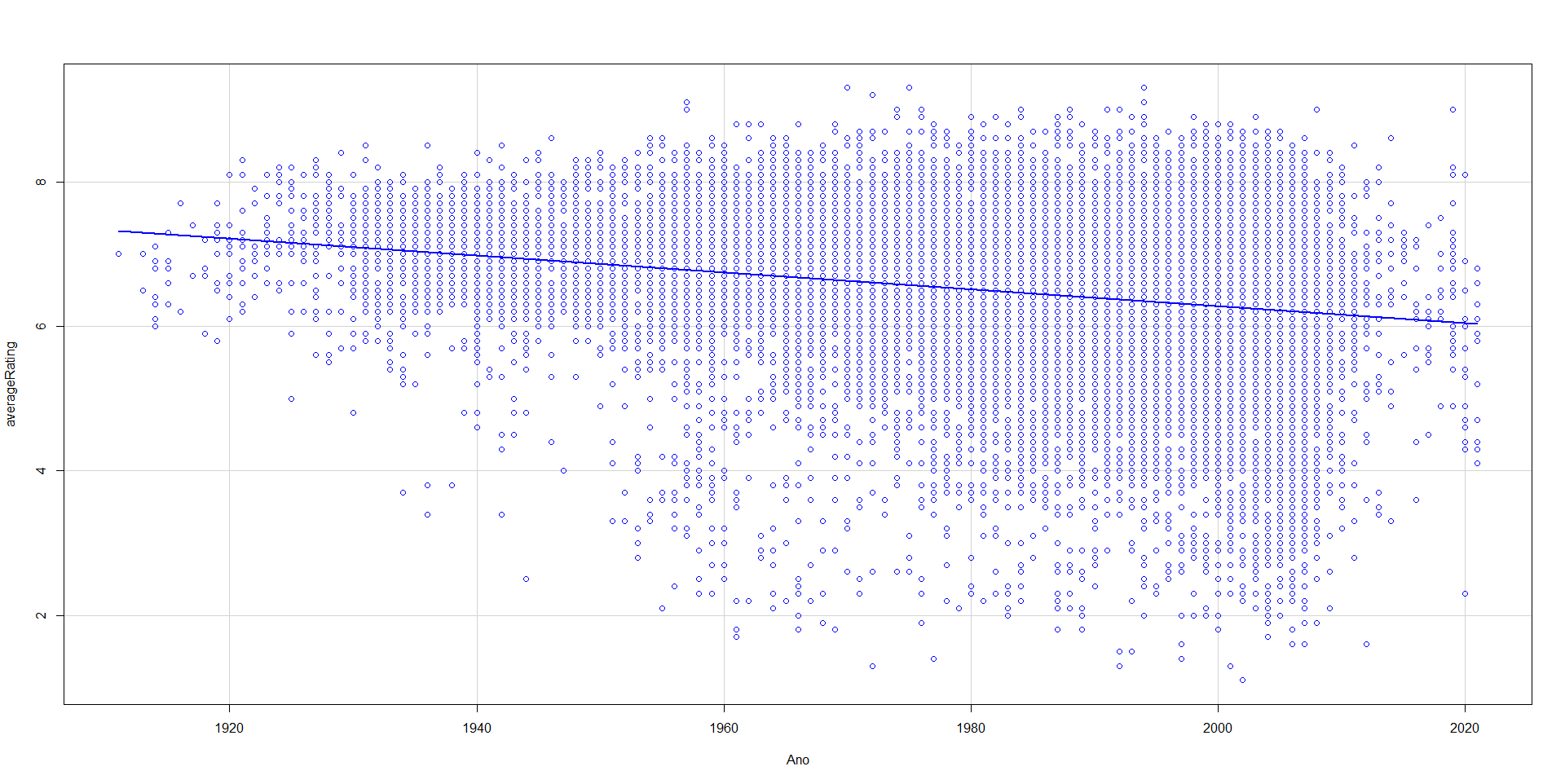
O boxplot (ou diagrama de caixa) é uma ferramenta gráfica que permite visualizar a distribuição e valores discrepantes dos dados, fornecendo assim um meio complementar para desenvolver uma perspectiva sobre o caráter dos dados. Além disso, o boxplot também é uma disposição gráfica comparativa. O boxplot tem como objetivo estudar as medidas estatística do conjunto de dados, como propriedades de locação, variabilidade, média e outliers.

No boxplot acima, são consideradas as variáveis faixa de ano, nota e subgênero (coloração). Fazendo uma correlação com o resumo numérico mostrado anteriormente. Assim, podemos ver qual a evolução da avaliação pública dos filmes categorizado pelos gêneros, e também algumas observações. Podemos destacar alguns dos boxplots, primeiro o do subgênero News que só teve uma aparição, já o subgênero drama é um caso bem curioso, no boxplot da parte antiga ele é assimétrico, essa assimetria mostra que existe mais filmes com notas maiores, e também, diversos outliars, porém, no boxplot que é representado pela parte dos filmes “novos” é o contrário, ele ainda é assimétrico só que a mediana aponta que existem mais filmes com notas menores, novamente diversos outliars (incluindo os 2 filmes mais bem avaliados de toda a base de dados).



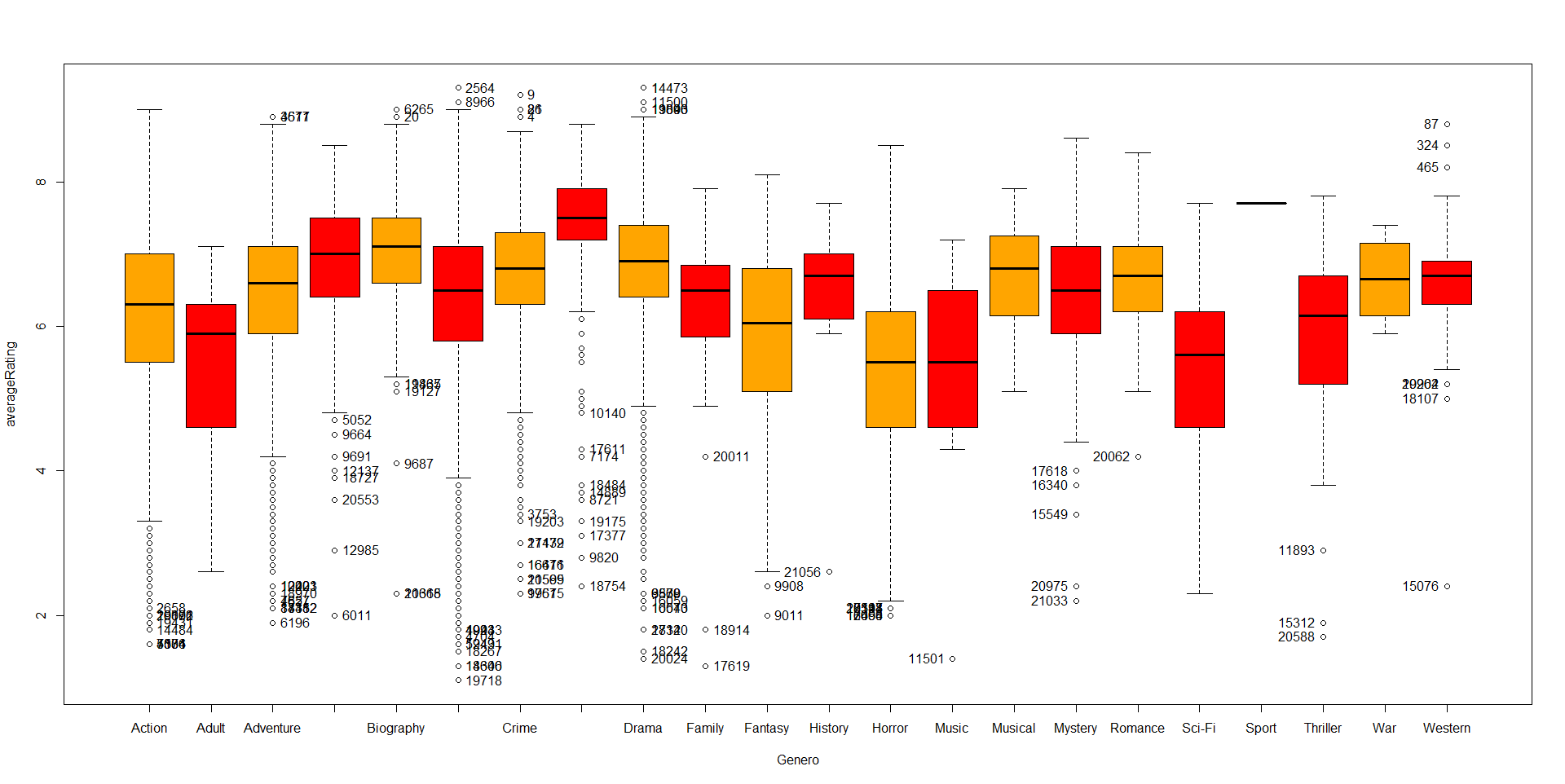
Fonte: Base de dados – Filmes

O gráfico de barras é utilizado em geral para representar dados de uma tabela de frequências associadas a uma variável qualitativa. Nesse tipo de gráfico, cada barra retangular representa a frequência ou a frequência relativa da respectiva opção da variável. Pode-se analisar neste gráfico de barras as frequências de quantas vezes aparecem os subgêneros, ou seja, na base de dados há 21 mil linhas e 3 colunas divididas em: gênero, subgênero e subgênero 2, esta é a de subgênero. O gráfico acima apresenta uma frequência de 5357, que tem como subgênero “drama”, cujo na tabela é o maior número e logo após tem-se o sub gênero de romance.



Fonte: Base de dados – Filmes

Um gráfico de dispersão é uma representação gráfica da associação entre pares de dados. Esse emparelhamento de dados é o resultado da associação de diferentes medições de uma certa causa com a medição correspondente da característica de qualidade, por exemplo. No gráfico acima pode-se analisar um gráfico de dispersão de notas por ano. Seguindo as possíveis correlações a serem identificadas em um gráfico de dispersão, chega-se à conclusão de que é um gráfico sem correlação, ou seja, pode-se dizer que não há correlação entre a variável Y e a variável X. Os seus valores não estão de maneira alguma atrelados. Apesar disso, quando traçado uma linha, o gráfico apresenta uma baixa tendência de queda das notas durante o passar dos anos.



Fonte: Base de dados – Filmes

O boxplot acima, representa a nota de cada gênero, diferente de outra análise prévia dos subgêneros o gênero é o conteúdo principal do filme, e assim, podemos ver que como gênero principal, guerra é desenhado como graficamente simétrico e sem outliars, sendo todos os seus filmes com notas maiores que 6. Na base de dados inteira apenas um filme categorizado como sport, o gênero ação é representado como simétrico também, porém só possui outliars de notas menores, nenhum filme do gênero ação atinge a terceira camada da divisão das notas (Ruim, boa, ótima).

* **Conclusão**

Com todas as análises feitas, temos todas as conclusões para poder analisar o desenvolvimento do cinema durante todos esses anos, desde 1911 data do primeiro filme analisado até o ano atual. Conseguimos responder algumas perguntas, por exemplo, a combinação de gêneros cinematográficos mais provável, seria comédia dramática logo em seguida Ação e drama.

O crescimento do mercado cinematográfico, foi comprovado estatisticamente a evolução na produção de filmes, com o resumo numérico dos anos onde foi mostrado que em 29 anos foi produzida a mesma quantidade de filmes (relatados na base de dados) que em 82 anos anteriores quando comparados por períodos.

Por fim, a análise que deve ser tirada da pesquisa quantitativa é igual a do gráfico de dispersão, não há qualquer fórmula mágica para conseguir fazer um filme de sucesso, pois assim como o um gráfico de dispersão não há qualquer sinal de que tenha correlação com gêneros, tempo de tela, ou ano de lançamento. Sendo assim, o único motivos para a pesquisa de notas analisada ser um sucesso e 100% confiável é que todos os avaliadores são livres para avaliar como desejar sem vícios e pré-conceitos, assim, fazendo com que os relatórios retirados sejam as informações condizentes com a realidade.