UMA ANÁLISE SOBRE FILMES USANDO R E RSTUDIO

Felipe Mateus Evangelista[1], Filipe Pereira da Silva[2], William Salvador Antonelli[3]

1 INTRODUÇÃO E OBJETIVOS

O principal objetivo deste resumo foi fazer uma análise estatística sobre dados de filmes. A partir de uma base de dados, buscamos investigar informações sobre algumas variáveis dos filmes, classificar e descrever estas variáveis. Também buscamos criar tabelas sobre estes dados, criar gráficos para visualizar informações e entender como esses dados se comportam, criar métricas com os dados e discutir os resultados alcançados.

2 MATERIAIS E MÉTODOS

A base de dados escolhida foi uma base com dados relacionados a filmes. O conjunto de dados foi removido do site da Kaggle, que possui uma base do TMDb (The Movie Database) com 5 mil filmes, mas utilizamos somente uma amostra de 350 filmes. Nesta base existiam algumas variáveis que foram desconsideradas, sendo assim filtramos ela na sintaxe R para analisarmos somente os dados desejados. Sendo excluídas as variáveis, homepage, id, keywords, overview, spoken_languages, status, tagline, original_title, runtime. Outrossim, outra medida necessária foi ignorar alguns filmes cujo os dados não estavam satisfatórios, sendo por ter dados nulos, zerados ou incongruentes.

As variáveis definidas para análise foram:

- *Custo(budget)*: Informa o valor gasto na construção do filme. Sendo classificada como uma variável quantitativa contínua.

  Foi escolhida porque permite comparar o orçamento com a receita gerada e calcular a rentabilidade do filme. Essa informação é importante para avaliar se o filme foi financeiramente bem-sucedido ou não.
- *Gênero(genre)*: Informa qual é o gênero do filme. Sendo classificada como uma variável qualitativa nominal
- *Língua original(original language)*: Informa a língua original na qual o filme foi produzido. Sendo classificada uma variável qualitativa nominal.
- *Título original (original title)*: Informa o título do filme. Sendo classificada como uma variável qualitativa nominal
- *(Popularidade)popurlaty*: Informa a quantidade de pessoas que assistiram o filme no cinema. Sendo classificada como uma variável quantitativa discreta.
- *Produtora(production_company)*: Informa a produtora do filme. Sendo classificada como uma variável qualitativa nominal.
- *País(production_country*: Informa o país que produziu o filme. Sendo classificada como uma variável qualitativa nominal.

- *Data de lançamento(release_date)*: Informa a data de lançamento do filme. Sendo classificada como uma variável qualitativa ordinal
- *Receita(revenue)*: Informa a receita do filme. Sendo classificada como uma variável quantitativa contínua.

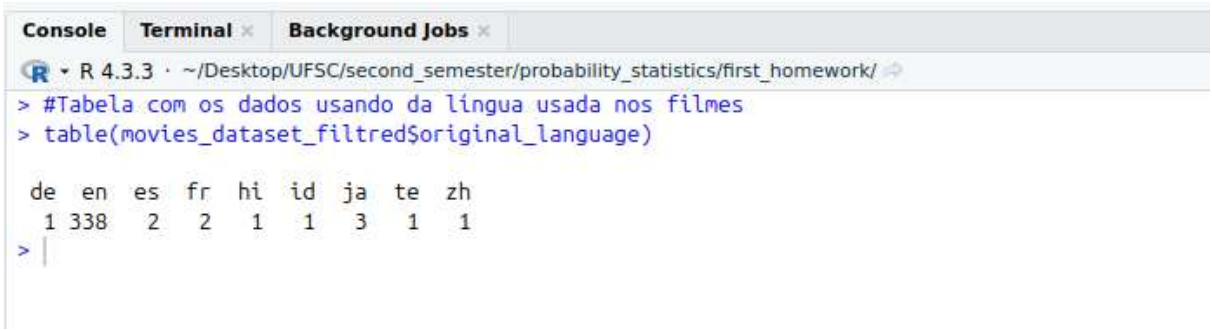
  Foi escolhida porque é importante para avaliar o retorno sobre o investimento e é importante para entender o impacto financeiro da produção.
- *Média dos votos(vote_average)*: Informa a média da nota do filme. Sendo classificada como uma variável quantidade contínua. É quantitativa contínua.
- *Quantidade de votos(vote_count)*: Informa a quantidade de votos que o filme recebeu. Sendo classificada com uma variável quantitativa discreta.

3 RESULTADOS E DISCUSSÕES

3.1 Tabelas

Primeiro vamos analisar o comportamento de algumas variáveis qualitativas nominais por meio de tabelas de frequência, são elas, Língua original (original language), Gênero (genre), Produtora (production_company), País (production_country).

Figura 1 - Tabela de frequência para Língua original.



Conforme podemos ver na imagem acima, é notável uma predominância da língua inglesa, sendo ela majoritariamente compondo a maioria das linguagens do filmes que são criados de acordo com a nossa amostra.

Figura 2 - Tabela de frequência para Gênero

Diferente da Figura 1, onde os dados de língua ficavam concentrados em uma única língua, na tabela sobre gênero podemos ver uma melhor distribuição dos dados, ainda assim é possível notar que os gêneros Drama, Comédia e Ação tem uma maior quantidade de filmes representados.

Figura 3 - Tabela de frequência para Produtora

Source

Console | Terminal | Background Jobs

R 4.3.3 · ~/Desktop/UFSC/second_semester/probability_statistics/first_homework/

2929 Productions	40 Acres & A Mule Filmworks	A Loopy Production LLC	Alcon Entertainment	Alliance Atlantis Communications	Amblin Entertainment
1	1	1	2	1	2
Amen Ra Films	American Zoetrope	Anchor Bay Films	Appian Way	Arka Media Works	Artisan Entertainment
1	1	1	1	1	1
Bac Films	BBC	BBC Films	Beijing New Picture Film Co. Ltd.	Blumhouse Productions	Bold Films
1	1	3	1	1	1
Boram Entertainment Inc.	Brooksfilms	Canal+	Cannon Films	Castle Rock Entertainment	CBS Films
1	1	1	1	2	1
Celluloid Dreams	Central Cinema Company Film	Charles K. Feldman Group	Cineplex-Odeon Films	Columbia Pictures	Columbia Pictures Corporation
1	1	1	1	22	6
Conundrum Entertainment	Cookout Productions	Corrino Media Corporation	Crossbow Productions	Crystal Sky Worldwide	Cube Vision
1	1	1	1	1	1
Dark Horse Entertainment	Davis Entertainment	Davis-Films	DC Comics	Destination Films	Di Bonaventura Pictures
1	1	1	1	1	2
Dimension Films	DiNovi Pictures	DreamWorks	DreamWorks Animation	DreamWorks SKG	Dune Entertainment
3	1	1	2	4	2
Duplass Brothers Productions	Electric Entertainment	Emmett/Furla Films	Eon Productions	Estudios Picasso	Figment Films
1	1	1	1	1	1
Film4	Fine Line Features	Flicks Motion Pictures	Focus Features	Fox 2000 Pictures	Fox Entertainment Group
1	1	1	2	2	1
Fox Searchlight Pictures	France 2 Cinéma	France 3 Cinéma	Franchise Pictures	Gaumont	Gracie Films
6	2	1	1	1	1
Gravier Productions	Grindstone Entertainment Group	Imagine Entertainment	Incentive Filmed Entertainment	Ingenious Film Partners	Ingenious Media
1	1	2	1	2	2
Itaca Films	Jerry Bruckheimer Films	Jerry Weintraub Productions	Journeyman Pictures	Keith Barish Productions	Kennedy Miller Productions
1	2	1	1	1	1
Kennedy/Marshall Company, The	Killer Films	Lakeshore Entertainment	Larger Than Life Productions	Laurel Group	Lawrence Bender Productions
1	1	4	1	1	1
Legendary Pictures	Lions Gate Films	Lionsgate	Live Entertainment	Marvel Studios	Metro-Goldwyn-Mayer (MGM)
2	2	4	1	2	2
Mike Zoss Productions	Miramax Films	National Geographic Entertainment	New Amsterdam Entertainment	New Line Cinema	New World Pictures
1	8	1	1	13	1
Nick Wechsler Productions	NULL	Original Film	Paramount Pictures	Paramount Vantage	Participant Media
1	4	1	30	1	1
Plan B Entertainment	Playtone	Punch 21 Productions	Punch Productions	Red Envelope Entertainment	Regency Enterprises
1	1	1	1	1	2
Revolution Studios	Rhombus Media	River Road Entertainment	Robert Simonds Productions	Ruby in Paradise	Saturn Films
1	1	2	1	1	1
Scott Rudin Productions	Screen Gems	Screen Gems, Inc.	Shaw Brothers	Sidney Kimmel Entertainment	Silver Pictures
2	1	1	1	2	1
Sony Pictures Classics	Spyglass Entertainment	Studio Babelsberg	Studio Ghibli	StudioCanal	Summit Entertainment
1	4	2	2	3	4
Sunswept Entertainment	Tapestry Films	The Weinstein Company	Thinkfilm	This Is That Productions	Tig Productions
1	1	2	1	1	1
Touchstone Pictures	Tree Line Films	Trilogy Entertainment Group	TriStar Pictures	Twentieth Century Fox Film Corporation	United Artists
5	1	1	6	14	5
Universal Pictures	Universal Studios	USA Films	Vertigo Films	Village Roadshow Pictures	Walt Disney Pictures
24	2	1	1	6	12
Warner Bros.	Warner Bros. Pictures	Warner Brothers/Seven Arts	Worldview Entertainment	X-Filme Creative Pool	
9	1	1	1	1	

Na tabela para a variável produtora podemos notar uma heterogeneidade dos dados, ficando eles bem espalhados e tendo um número bem alto de produtoras, o que mostra que no mundo dos filmes temos uma grande variedade de empresas que os produzem.

Figura 4 - Tabela de frequência para línguas e países

Console | Terminal | Background Jobs

R 4.3.3 · ~/Desktop/UFSC/second_semester/probability_statistics/first_homework

```
> #Tabela com os dados relacionando paises que produziram os filme
> table(movies_dataset_filtred$production_country, movies_dataset_
```

	de	en	es	fr	hi	id	ja	te	zh
Argentina	0	1	0	0	0	0	0	0	0
Australia	0	3	0	0	0	0	0	0	0
Canada	0	11	0	0	0	0	0	0	0
China	0	1	0	0	0	0	0	0	1
Czech Republic	0	1	0	0	0	0	0	0	0
France	0	11	0	1	0	0	0	0	0
Germany	1	14	0	0	0	0	0	0	0
Hungary	0	1	0	0	0	0	0	0	0
India	0	2	0	0	1	0	0	1	0
Indonesia	0	0	0	0	0	1	0	0	0
Ireland	0	1	0	0	0	0	0	0	0
Italy	0	1	0	0	0	0	0	0	0
Japan	0	3	0	0	0	0	3	0	0
Mexico	0	2	1	0	0	0	0	0	0
Netherlands	0	2	0	0	0	0	0	0	0
New Zealand	0	2	0	0	0	0	0	0	0
Peru	0	1	0	0	0	0	0	0	0
Spain	0	1	1	0	0	0	0	0	0
Switzerland	0	0	0	1	0	0	0	0	0
United Kingdom	0	27	0	0	0	0	0	0	0
United States of America	0	253	0	0	0	0	0	0	0

Figura 5 - Tabela de frequência para Países

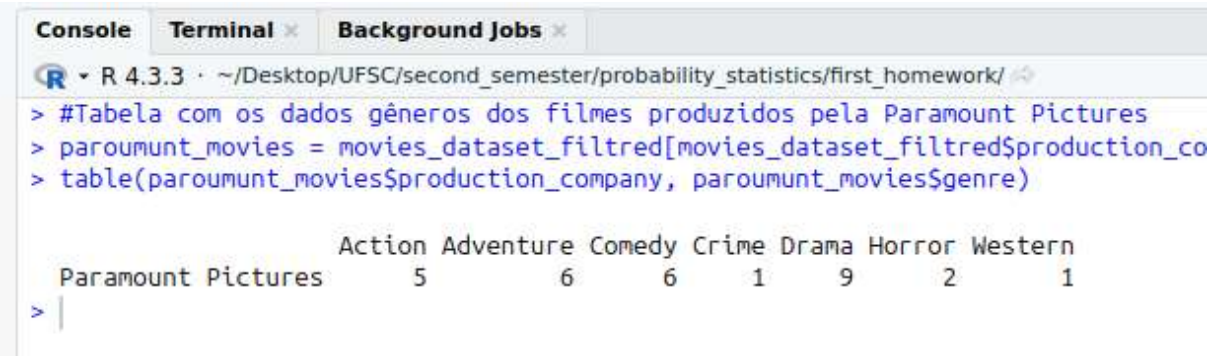
A tabela de países segue a mesma dinâmica da tabela de línguas dos filmes, podemos ver que a maioria dos dados ficam concentrados em um país, os Estados Unidos, tendo os outros países um número muito menor de frequência.

Podemos também analisar estas variáveis conjuntas para poder ver correlações entre os dados. Por exemplo, ao analisarmos a figura 4, notamos que a quantidade de filmes em língua inglesa se dá pelo fato da maioria dos filmes serem produzidos pelos Estados Unidos. Cabe também destacar um fato curioso, a França e a Alemanha, apesar de terem línguas próprias, francês e alemão, respectivamente, possuem mais filmes feitos em língua inglesa do que em suas línguas nativas.

Por último, podemos analisar a variável gênero do filme por país, como nas figuras abaixo. Na figura 6 estamos analisando os gêneros dos filmes produzidos pela França e na Paramount Pictures, em ambas podemos perceber que os gêneros são bem distribuídos, sem haver gêneros que se sobressaiam expressivamente perante aos outros.

Figura 6 - Tabela de frequência dos gênero dos filmes produzidos pela França

Figura 7 - Tabela de frequência dos gênero dos filmes produzidos pela Paramount Pictures



3.2 Gráficos

Com a base de dados também buscamos criar gráficos para conseguir visualizar o comportamento dos dados e como eles se relacionam individualmente ou agrupados, os primeiro gráficos a serem analisados serão os de densidade.

Figura 8 - Gráfico de densidade para Orçamento (budget)

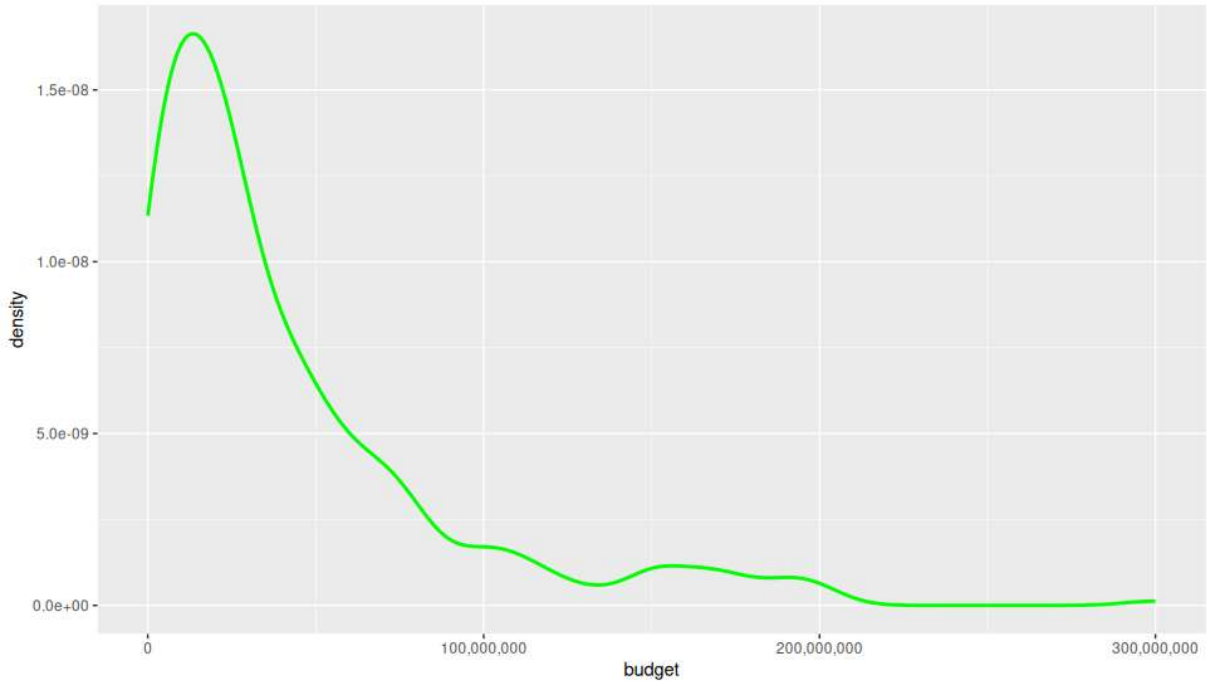


Figura 9 - Gráfico de densidade para Receita (revenue)

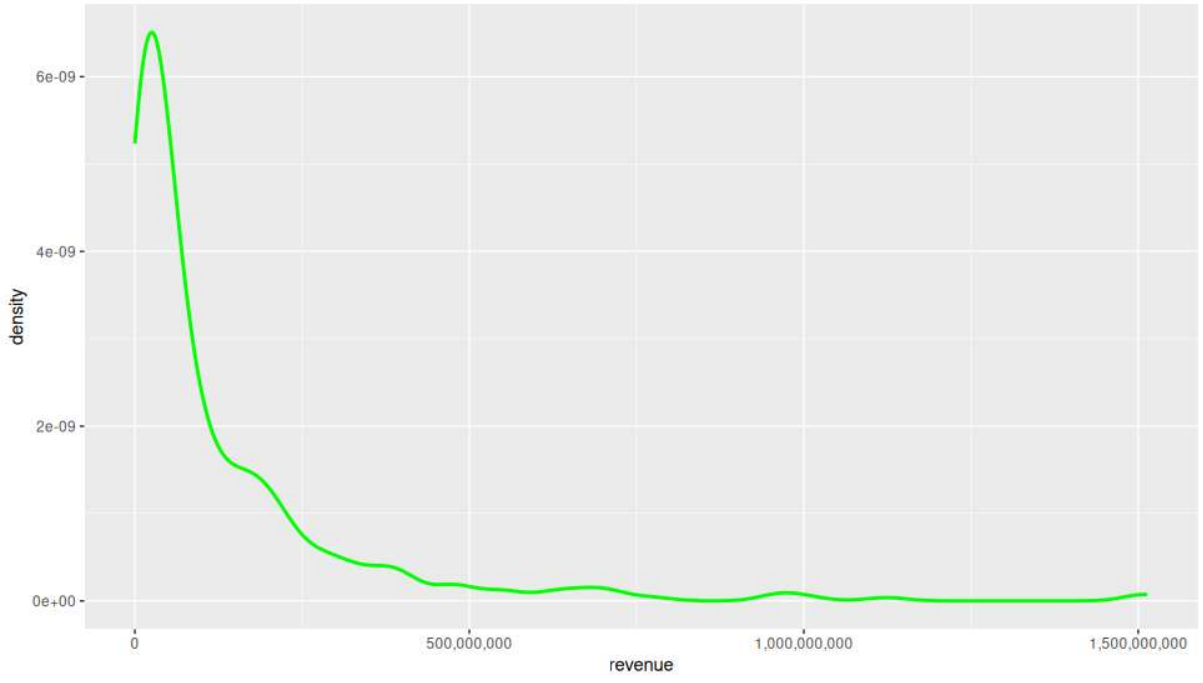
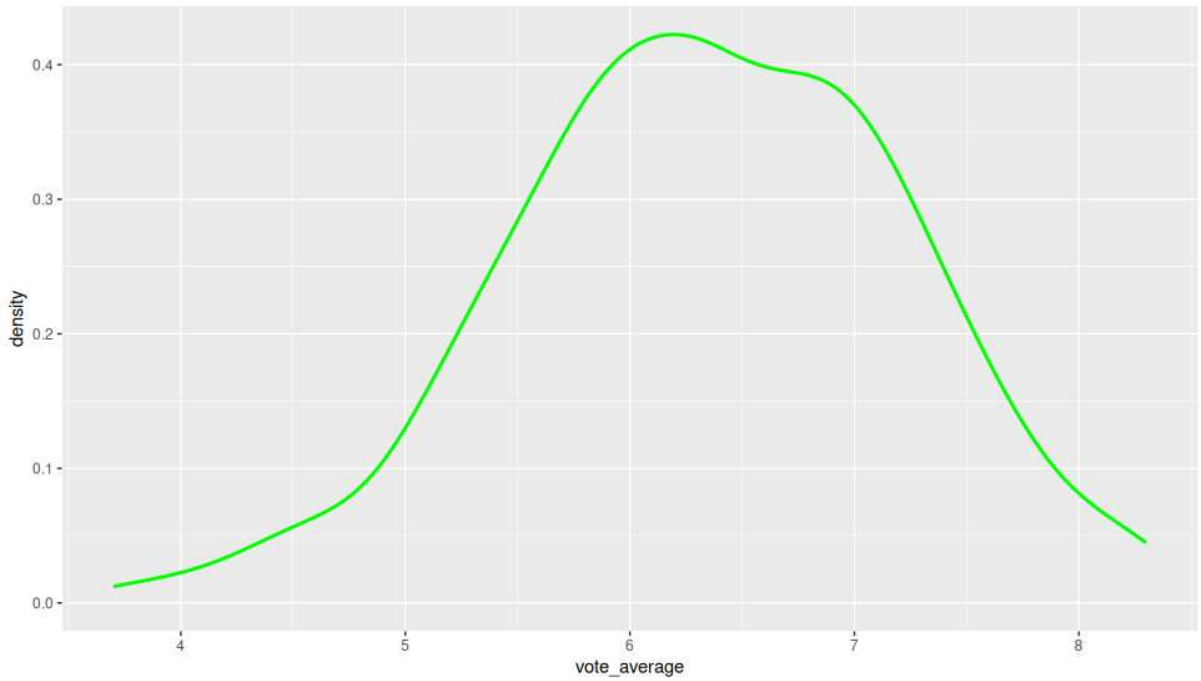
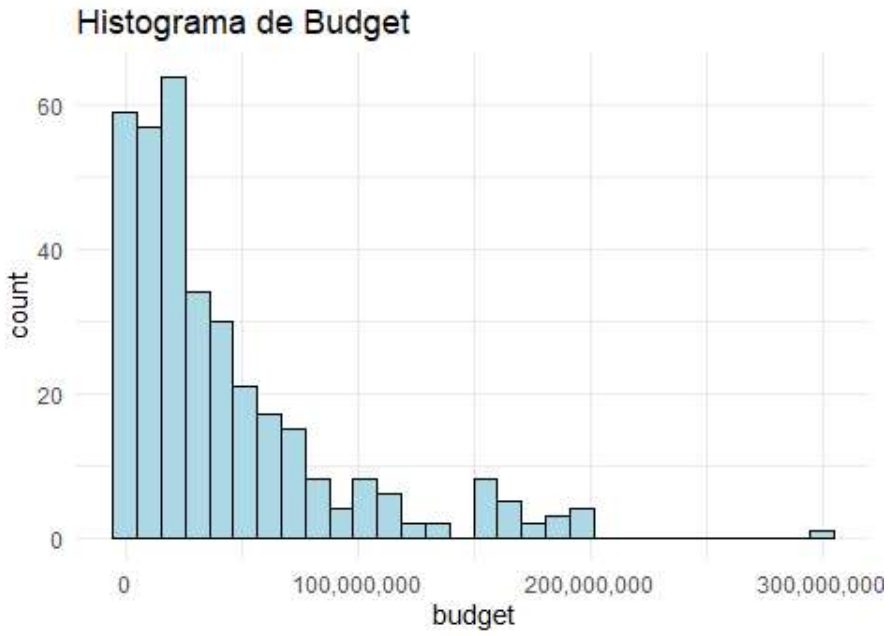


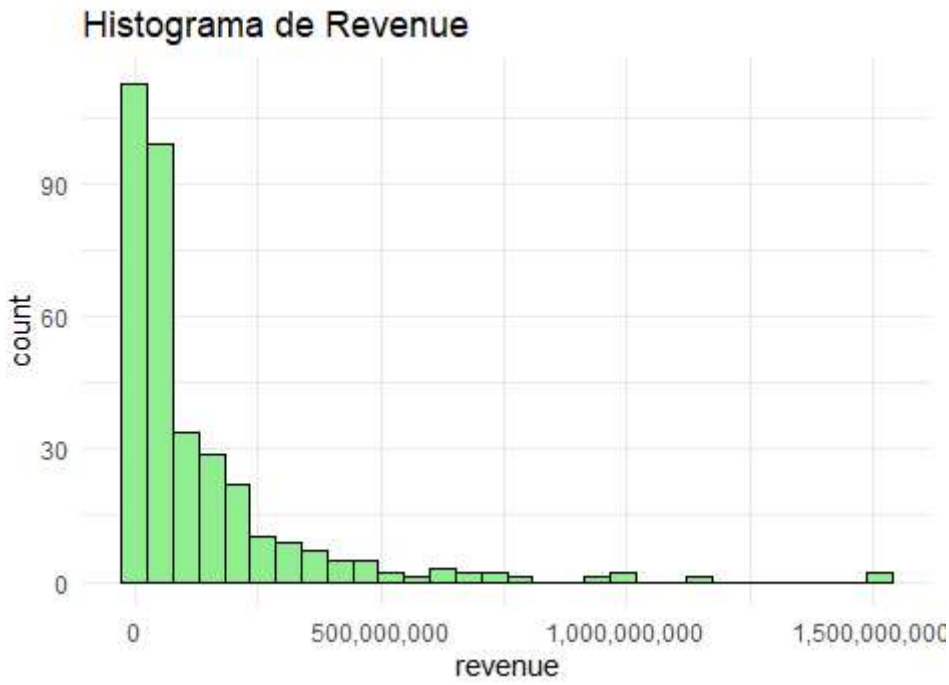
Figura 10 - Gráfico de densidade para Notas dos filmes (vote_average)



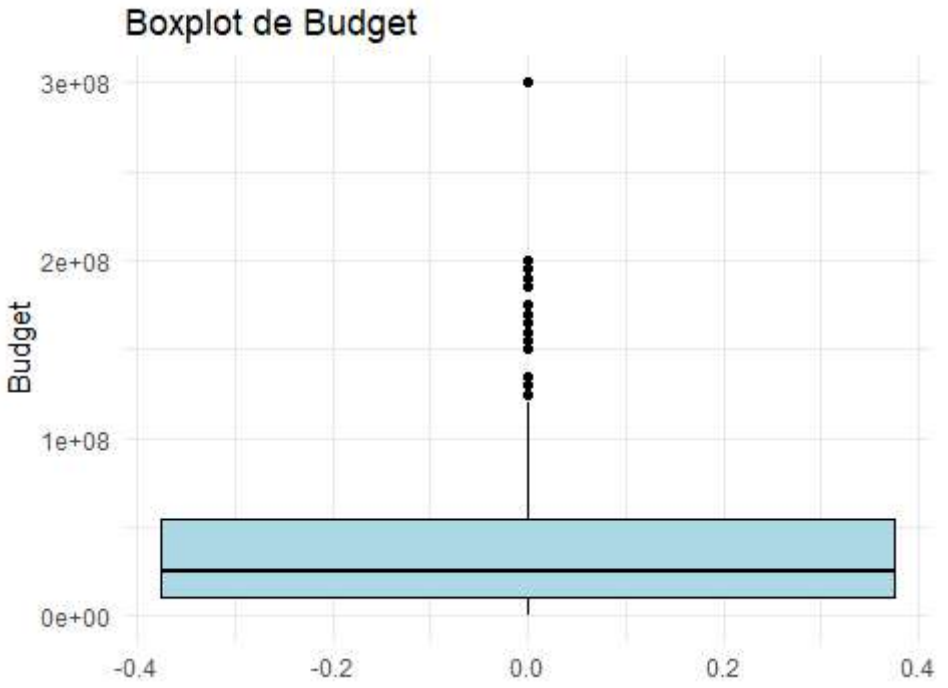
Com os gráficos das figuras 8, 9 e 10 podemos perceber alguns comportamentos, como, nos gráficos de orçamento e receita temos uma maior pico de densidade para filmes que tem um baixo orçamento e filmes com uma baixa receita, também percebemos a presença de outliers, o que faz com que estes gráficos fiquem com as “caudas” mais esticadas e que podem influenciar métricas como cálculo da média. Já no gráfico da figura 10, temos um comportamento diferente, nesse gráfico podemos ver que as notas são bem distribuídas entre a média central, fazendo com com que o pico fique próximo no centro e também mostrando que os dados estão bem distribuídos pelo gráfico, sem um local central de concentração.



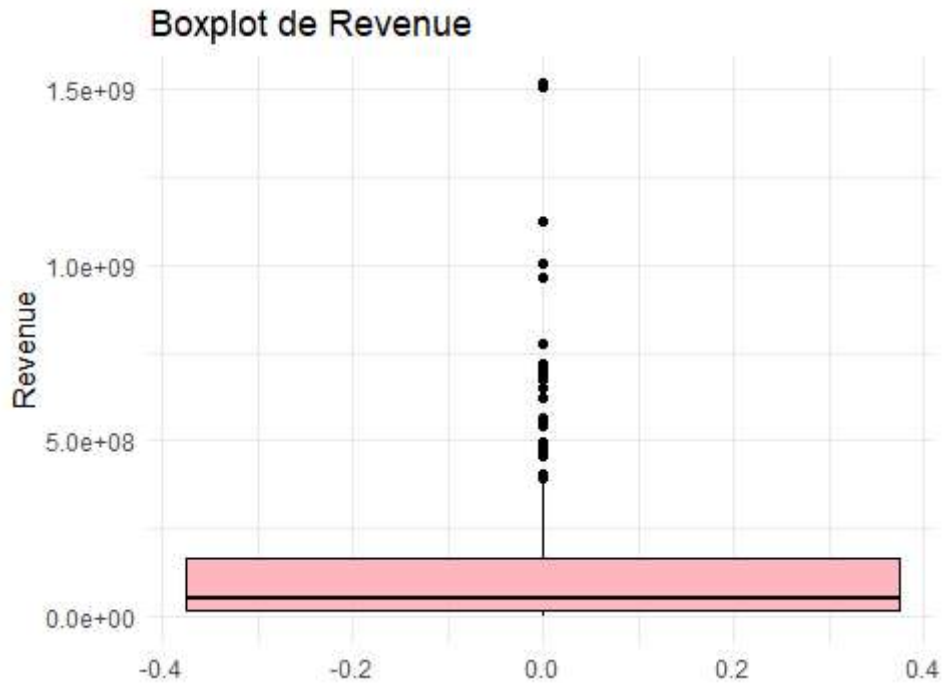
O gráfico acima mostra que a maior parte dos filmes têm orçamentos baixos, poucos filmes têm orçamentos muito altos, o que gera a cauda longa à direita.



O gráfico acima mostra que a grande maioria dos filmes gera receitas menores. Existem alguns poucos filmes com faturamentos altíssimos, a distribuição é assimétrica à direita.



O gráfico acima mostra que a maioria dos valores de budget está concentrada entre valores relativamente baixos. Existem muitos outliers, o que indica que há filmes com orçamentos muito maiores que a média. A mediana (linha preta dentro da caixa) é bem próxima de 0 comparada ao máximo, o que mostra que a maioria dos filmes tem orçamento baixo em relação a poucos filmes com orçamento muito alto. O budget é altamente assimétrico para a direita. Há uma grande desigualdade no orçamento dos filmes: poucos filmes gastam muito, a maioria gasta pouco.



O gráfico acima mostra que assim como o orçamento, o revenue também apresenta muitos outliers. A maioria dos filmes arrecada valores baixos, mas alguns poucos arrecadam muito. A mediana também é baixa comparada ao valor máximo. O faturamento dos filmes também é muito desigual, com a maior parte dos filmes faturando pouco e poucos filmes faturando muito. A distribuição é assimétrica à direita, indicando que existem filmes que se destacam muito em receita.

3.3 Medidas de Resumo

```
> summary(movies_dataset_filtred$budget)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
4 10.875.000 25.000.000 41.818.896 54.750.000 300.000.000

[1] "Variância:  2.160.957.918.892.042"
[1] "Desvio Padrão:  46.486.104,5785947"
[1] "Coeficiente de Variação:  111,16052600433 %"
```

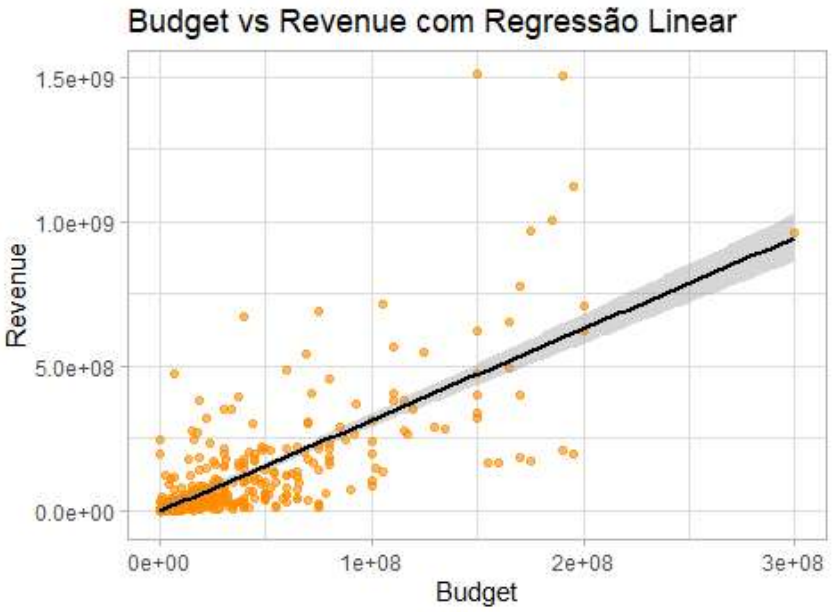
O desvio padrão é muito alto em relação à média, o que indica que os valores de budget são muito dispersos. O coeficiente de variação acima de 100% mostra que a variabilidade é maior que a média. Como a média é maior que a mediana, a distribuição é assimétrica à direita.

```
> summary(movies_dataset_filtred$revenue)
 Min.   1st Qu.    Median      Mean      3rd Qu.      Max.
    7 17.520.000 50.770.000  128.200.000 164.300.000 1.514.000.000
```

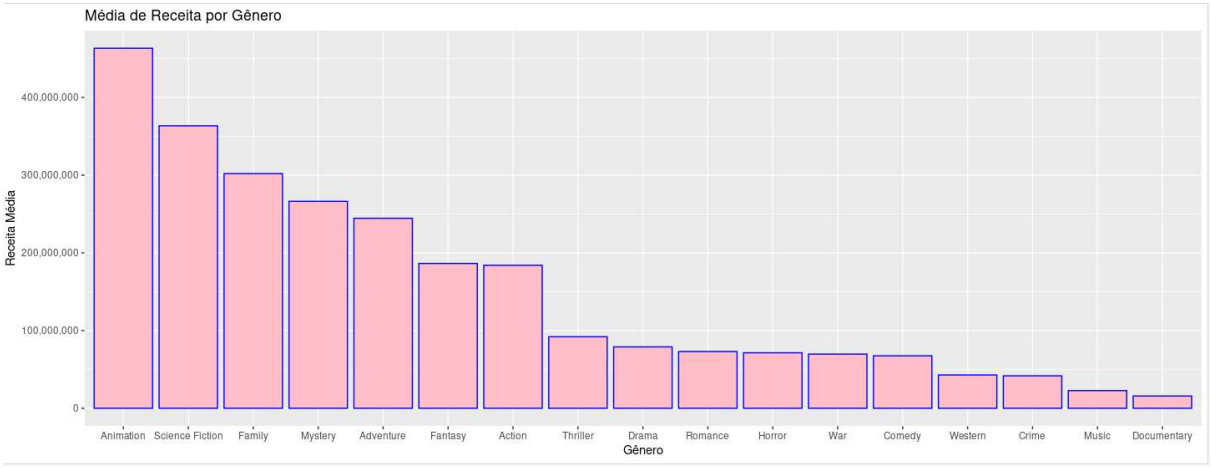
[1] "Variância: 41.004.528.060.731.096"
[1] "Desvio Padrão: 202.495.748,253466"
[1] "Coeficiente de Variação: 157,950738665499 %"

O desvio padrão é ainda mais alto que no caso do budget, reforçando que o faturamento dos filmes é altamente variável. O coeficiente de variação acima de 150% mostra uma variabilidade extrema, muito maior que a média. A diferença grande entre média e mediana também indica uma assimetria forte à direita, com poucos filmes faturando bilhões e muitos faturando valores baixos.

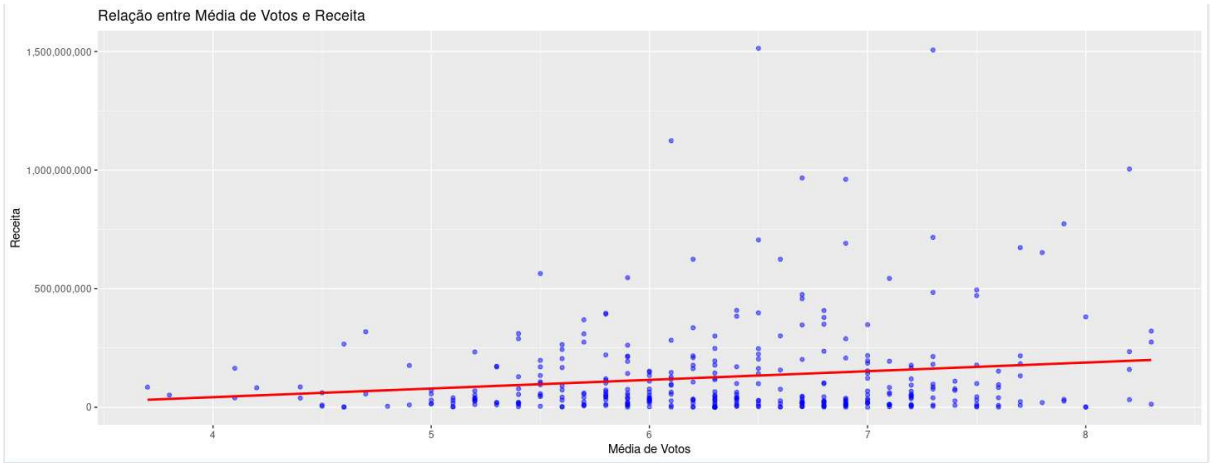
3.4 Gráficos Relacionando Variáveis



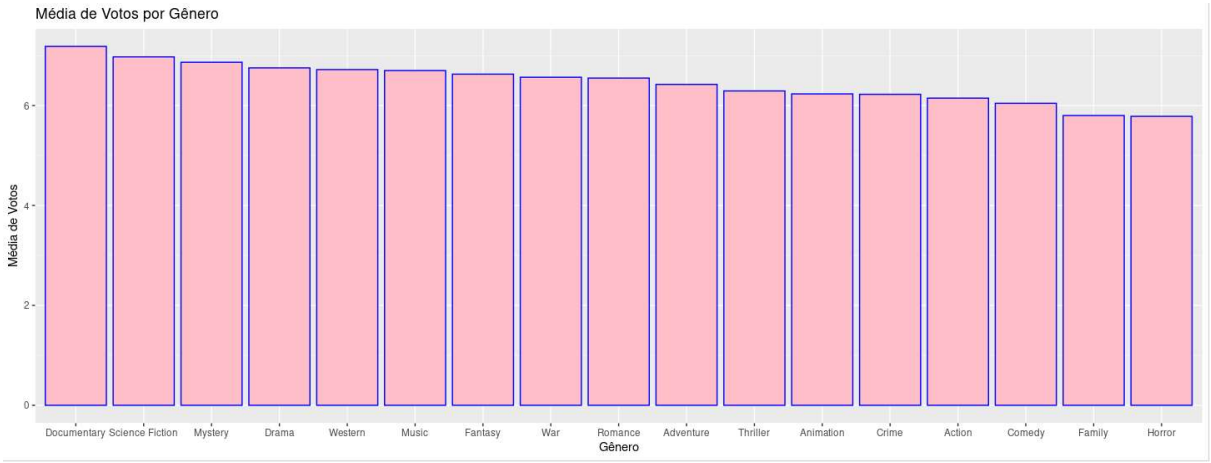
O gráfico mostra que existe uma correlação positiva entre o budget e o revenue: em geral, quanto maior o orçamento, maior a receita. Apesar disso, a dispersão é alta e nem todo filme caro fatura muito, e nem todo filme barato fatura pouco. A linha de regressão é inclinada positivamente, confirmando a tendência. Investir mais tende a aumentar a receita, mas não é garantia de sucesso.



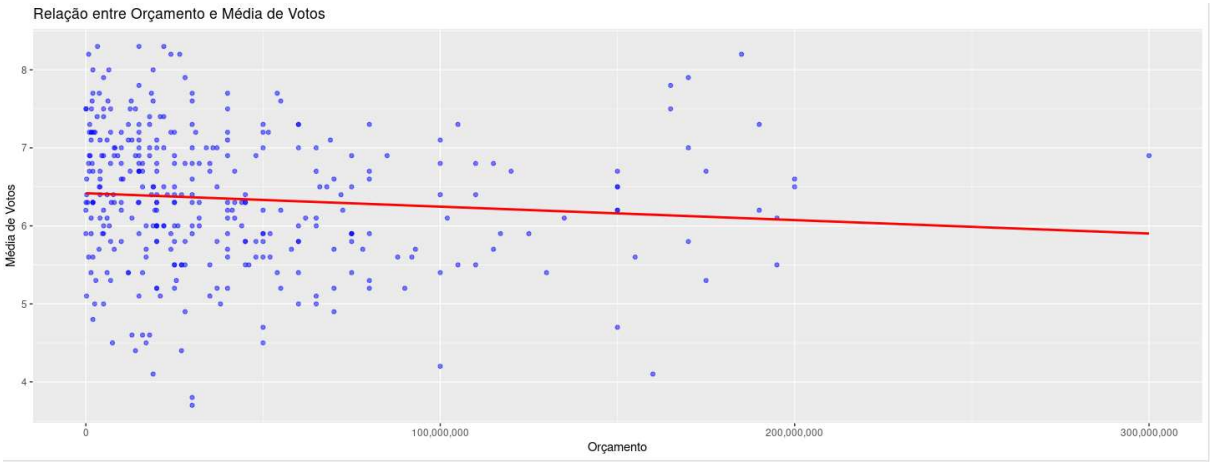
Ao analisar o gráfico que compara a receita média dos filmes por gênero fica evidente que filmes de nichos mais específicos possuem o faturamento médio muito menor quando comparados a filmes cujo público alvo é mais abrangente. Os filmes de animação são os que possuem o maior faturamento enquanto documentários seguem por último com uma diferença exorbitante. Essa disparidade pode ser atribuída a diversos fatores. Filmes de animação costumam atrair públicos de diferentes faixas etárias e contam com investimentos massivos em marketing com uma ampla distribuição internacional. Já os documentários geralmente têm públicos muito mais restritos e são lançados com campanhas de divulgação mais modestas.



O gráfico demonstra uma concentração grande de filmes com receita baixa independentemente da média de votos, entretanto, há uma leve tendência positiva na linha de regressão onde filmes com média de votos mais alta tendem a gerar receitas maiores. Apesar da presença de muitos outliers com receitas muito altas, isso demonstra que melhores avaliações podem estar associadas a maiores receitas, ainda que a correlação também pareça fraca.



O gráfico que compara a média de votos por gênero apresentou resultados bem equilibrados e com pouca variação. O gênero Documentário foi o que conquistou a maior média de votos, enquanto o gênero Horror ficou em último lugar. Isso se deve ao fato dos documentários possuírem um teor mais educacional e não necessariamente precisarem agradar o público quando comparado aos outros gêneros.



O gráfico acima mostra que a maioria dos pontos se concentra na parte esquerda do gráfico onde há orçamentos mais baixos. Apesar de haver filmes com orçamentos altos eles não apresentam uma média de votos consideravelmente maior. A linha de tendência tem uma leve inclinação negativa, indicando uma fraca correlação entre orçamento e a média de votos, entretanto, essa relação é muito fraca é provavelmente não é estatisticamente significativa.

4 CONSIDERAÇÕES FINAIS

A análise estatística realizada permitiu obter diversas informações relevantes sobre os filmes presentes na base de dados. Foi possível identificar padrões de produção e distribuição, como a predominância da língua inglesa e dos Estados Unidos como país produtor, além da diversidade de gêneros e produtoras envolvidas.

Através das medidas de tendência central e dispersão, observamos que tanto o orçamento quanto a receita dos filmes apresentam distribuições altamente assimétricas, com grande quantidade de outliers. Isso evidencia uma forte desigualdade no investimento e no retorno financeiro entre os filmes analisados.

Os gráficos relacionando as variáveis quantitativas demonstraram que há uma correlação positiva entre orçamento e receita, ou seja, filmes com maior investimento tendem a gerar maiores receitas, embora existam exceções. Já a relação entre média de votos e receita também mostrou uma tendência positiva, mas fraca, indicando que filmes bem avaliados podem, em média, ter melhores desempenhos financeiros. Por outro lado, não foi possível observar uma correlação significativa entre orçamento e média de votos, sugerindo que um orçamento elevado não garante necessariamente uma boa recepção por parte do público.

Por fim, o uso da linguagem R e da plataforma RStudio mostrou-se eficaz para o tratamento, análise e visualização dos dados, permitindo uma exploração detalhada e acessível de informações complexas. Esta análise evidencia como a estatística pode ser uma ferramenta poderosa para interpretar e compreender melhor os dados no contexto do entretenimento e da indústria cinematográfica.

REFERÊNCIAS BIBLIOGRÁFICAS

CHUAN, August Sun. **TMDB 5000 Movie Dataset.** 2017. Disponível em: <https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata?select=tmdb_5000_movies.csv>. Acesso em: 27 abr. 2025

R Foundation. ***Documentation***. 2025. Disponível em: <https://www.r-project.org/other-docs.html>. Acesso em: 27 abr. 2025.

ANEXOS

Sintaxe do R utilizada na análise:

```
#Carrega bibliotecas necessárias
library(dplyr)
library(ggplot2)
library(jsonlite)


#Carrega dataset
movies_dataset = read.csv("tmdb_5000_movies.csv")

#Filtrando informações irrelevantes no dataset e tratando dados
columns_to_ignore = c("homepage", "id", "keywords", "overview",
"spoken_languages", "status", "tagline", "original_title")
movies_dataset_filtred <- movies_dataset[, !(names(movies_dataset) %in%
columns_to_ignore)]
movies_dataset_filtred <-
movies_dataset_filtred[movies_dataset_filtred$budget != 0, ]
movies_dataset_filtred <-
movies_dataset_filtred[movies_dataset_filtred$revenue != 0, ]

set.seed(15112000) # data de nascimento como parametro para criar amostra
movies_dataset_filtred =
movies_dataset_filtred[sample(nrow(movies_dataset_filtred), 350), ]

movies_dataset_filtred <- movies_dataset_filtred |>
  mutate(
    genre = as.character(sapply(lapply(genres, fromJSON), function(x)
x$name[1])),
    production_company = as.character(sapply(lapply(production_companies,
fromJSON), function(x) x$name[1])),
    production_country = as.character(sapply(lapply(production_countries,
fromJSON), function(x) x$name[1])),
  )

movies_dataset_filtred <-
  movies_dataset_filtred[, !(names(movies_dataset_filtred) %in% c("genres",
"production_companies", "production_countries"))]

#Tabelas individuais

#Tabela com os dados usando da língua usada nos filmes
table(movies_dataset_filtred$original_language)

#Tabela com os dados usando do Genêro usada nos filmes
```

```r
table(movies_dataset_filtred$genre)

#Tabela com os dados usando do Empresas que produziram os filmes
table(movies_dataset_filtred$production_company)

#Tabela com os dados usando dos paises que produziram os filmes
table(movies_dataset_filtred$production_country)

#Tabelas agrupando duas variáveis

#Tabela com os dados relacionando países que produziram os filmes e a língua
do do filme
table(movies_dataset_filtred$production_country,
movies_dataset_filtred$original_language)

#Tabela com os dados gêneros dos filmes produzidos pela França
france_movies =
movies_dataset_filtred[movies_dataset_filtred$production_country == "France",
]
table(france_movies$production_country, france_movies$genre)

#Tabela com os dados gêneros dos filmes produzidos pela Paramount Pictures
paroumunt_movies =
movies_dataset_filtred[movies_dataset_filtred$production_company ==
"Paramount Pictures", ]
table(paroumunt_movies$production_company, paroumunt_movies$genre)




# medidas de resumo
summary(movies_dataset_filtred$budget)
summary(movies_dataset_filtred$revenue)

# Variância budget
variancia_budget <- var(movies_dataset_filtred$budget)
print(paste("Variância: ", variancia_budget))

# Desvio Padrão budget
desvio_padrao_budget <- sd(movies_dataset_filtred$budget)
print(paste("Desvio Padrão: ", desvio_padrao_budget))

# Coeficiente de Variação budget
coeficiente_variacao_budget <- desvio_padrao_budget /
mean(movies_dataset_filtred$budget) * 100
```

```r
print(paste("Coeficiente de Variação: ", coeficiente_variacao_budget, "%"))


# Variância revenue
variancia_revenue <- var(movies_dataset_filtred$revenue, na.rm = TRUE)
print(paste("Variância: ", variancia_revenue))

# Desvio Padrão revenue
desvio_padrao_revenue <- sd(movies_dataset_filtred$revenue, na.rm = TRUE)
print(paste("Desvio Padrão: ", desvio_padrao_revenue))

# Coeficiente de Variação
coeficiente_variacao_revenue <- desvio_padrao_revenue /
mean(movies_dataset_filtred$revenue, na.rm = TRUE) * 100
print(paste("Coeficiente de Variação: ", coeficiente_variacao_revenue, "%"))

#Gráfico de desnisdade para budget
ggplot(movies_dataset_filtred, aes(x = budget, y =
                                     after_stat(density))) +
  geom_density(color = "green", linewidth = 1) +
  scale_x_continuous(labels = scales::label_number(big.mark = ",",
                                                   decimal.mark = "."))

#Gráfico de desensidade de revenue
ggplot(movies_dataset_filtred, aes(x = revenue, y =
                                     after_stat(density))) +
  geom_density(color = "green", linewidth = 1) +
  scale_x_continuous(labels = scales::label_number(big.mark = ",",
                                                   decimal.mark = "."))

#Gráfico de desnsidade de média das notas do filmes
ggplot(movies_dataset_filtred, aes(x = vote_average, y =
                                     after_stat(density))) +
  geom_density(color = "green", linewidth = 1) +
  scale_x_continuous(labels = scales::label_number(big.mark = ",",
                                                   decimal.mark = "."))

# Gráfico Gênero vs Receita
media_receita_genero <- movies_dataset_filtred |>
  group_by(genre) |>
  summarise(media_receita = mean(revenue)) |>
  arrange(desc(media_receita))

ggplot(media_receita_genero, aes(x = reorder(genre, -media_receita), y =
media_receita)) +
  geom_col(color = "blue", fill = "pink") +
  labs(title = "Média de Receita por Gênero", x = "Gênero", y = "Receita
Média") +
```

```
  scale_y_continuous(labels = scales::label_number(big.mark = ",",
decimal.mark = "."))

# Gráfico Média de Votos vs Receita
ggplot(movies_dataset_filtred, aes(x = vote_average, y = revenue)) +
  geom_point(color = "blue", alpha = 0.5) +
  geom_smooth(method = "lm", color = "red", se = FALSE) +
  labs(title = "Relação entre Média de Votos e Receita", x = "Média de
Votos", y = "Receita") +
  scale_y_continuous(labels = scales::label_number(big.mark = ",",
decimal.mark = "."))

# Gráfico Gênero vs Média de Votos
media_votos_genero <- movies_dataset_filtred |>
  group_by(genre) |>
  summarise(media_votos = mean(vote_average)) |>
  arrange(desc(media_votos))

ggplot(media_votos_genero, aes(x = reorder(genre, -media_votos), y =
media_votos)) +
  geom_col(color = "blue", fill = "pink") +
  labs(title = "Média de Votos por Gênero", x = "Gênero", y = "Média de
Votos") +
  scale_y_continuous(labels = scales::label_number(big.mark = ",",
decimal.mark = "."))

# Gráfico Orçamento vs Média de Votos
ggplot(movies_dataset_filtred, aes(x = budget, y = vote_average)) +
  geom_point(color = "blue", alpha = 0.5) +
  geom_smooth(method = "lm", color = "red", se = FALSE) +
  labs(title = "Relação entre Orçamento e Média de Votos", x = "Orçamento", y
= "Média de Votos") +
  scale_x_continuous(labels = scales::label_number(big.mark = ",",
decimal.mark = "."))

# Gráfico com linha de tendência entre budget e revenue
ggplot(movies_dataset_filtred, aes(x = budget, y = revenue)) +
  geom_point(alpha = 0.6, color = "darkorange") +
  geom_smooth(method = "lm", se = TRUE, color = "black") +
  labs(
    title = "Budget vs Revenue com Regressão Linear",
    x = "Budget",
    y = "Revenue"
  ) +
  theme_light()

# Histograma para budget
ggplot(movies_dataset_filtred, aes(x = budget)) +
```

```
  geom_histogram(fill = "lightblue", color = "black", bins = 30) +  # Cor das
barras e borda
  scale_x_continuous(labels = label_comma()) +  # Formata os números no eixo
X
  scale_y_continuous(labels = label_comma()) +  # Formata os números no eixo
Y
  theme_minimal() +
  labs(title = "Histograma de Budget")



# Histograma para revenue
ggplot(movies_dataset_filtred, aes(x = revenue)) +
  geom_histogram(fill = "lightgreen", color = "black", bins = 30) +  # Cor
das barras e borda
  scale_x_continuous(labels = label_comma()) +  # Formata os números no eixo
X
  scale_y_continuous(labels = label_comma()) +  # Formata os números no eixo
Y
  theme_minimal() +
  labs(title = "Histograma de Revenue")

# Boxplot de budget
ggplot(movies_dataset_filtred, aes(y = budget)) +
  geom_boxplot(fill = "lightblue", color = "black") +
  labs(
    title = "Boxplot de Budget",
    y = "Budget"
  ) +
  theme_minimal()

# Boxplot de revenue
ggplot(movies_dataset_filtred, aes(y = revenue)) +
  geom_boxplot(fill = "lightpink", color = "black") +
  labs(
    title = "Boxplot de Revenue",
    y = "Revenue"
  ) +
  theme_minimal()
```