# LISTA DE TABELAS

# LISTA DE FIGURAS

# 1. INTRODUÇÃO

A indústria cinematográfica norteamericana movimenta bilhões de dólares no mundo inteiro. Prevê-se que, somente em 2016, a arrecadação mundial de bilheteria de filmes norteamericanos será de 38 bilhões de dólares. Para 2020, a previsão é que a arrecadação mundial chegue a 50 bilhões de dólares. A visita ao cinema faz parte da cultura norteamericana e de grande parte do mundo, e mais de 1,2 bilhões de ingressos foram vendidos somente nos Estados Unidos ano passado. As estatísticas nos dizem que 14% dos americanos vão aos cinemas pelo menos uma vez ao mês, 7% duas ou mais vezes ao mês, e 37% dizem irem algumas vezes ao ano (STATISTA – The Statistics Portal, 2016). O que se passa nas telonas, o sucesso ou fracasso de filmes, podem nos dizer muito de uma sociedade. Embora o foco deste trabalho não seja o ponto de vista sociológico dos filmes, vamos aqui estudar o desempenho dos filmes no que tange à arrecadação em salas de cinema norteamericanas, baseado em várias variáveis selecionadas para o estudo. Algumas dessas variáveis incluem o número de salas nas quais o filme foi exibido, seu gênero, qual mês ele estreiou, e até avaliações de críticos e do público.

O principal objetivo deste trabalho é estimar qual será a arrecadação final de um filme após o seu tempo em cartaz nos cinemas norteamericanos. Para isso, faremos uma análise de regressão utilizando como variável resposta a porcentagem que a bilheteria do final de semana de estréia de um filme tem em sua bilheteria final. Isto quer dizer que o modelo que apresentaremos aqui neste trabalho somente poderá ser utilizado após o filme já ter a informação sobre sua arrecadação no final de semana de estréia. Essa informação é muito importante para prever o sucesso de um filme: tipicamente, o final de semana de estréia de um filme é responsável por 30% da bilheteria final. É a força da arrecadação da estréia de um filme que define todas as decisões importantes referentes ao seu destino financeiro, uma vez que a competição por telas de cinema é feroz, fazendo com que os proprietários das salas de cinema não queiram gastar mais do que as obrigatórias duas semanas de exibição de um filme que não tem longevidade. Quanto menor for a porcentagem do final de semana de estréia na bilheteria final, mais longevidade e melhor divulgação “boca a boca” um filme teve. Optou-se por utilizar a “Porcentagem da Bilheteria do Final de Semana de Estréia na Bilheteria Final” como variável resposta do modelo de regressão devido ao

fato de esta ser uma variável muito mais simples de estudar que a própria bilheteria final. Com o valor predito da variável resposta e com a própria bilheteria de estréia, chegar em um valor previsto para bilheteria final não é uma tarefa complicada.

O tema escolhido neste trabalho já foi estudado em 2015, por Jeffrey Simonoff, da New York University, embora de forma muito mais simples e abreviada (Simonoff, 2015). Em seu artigo entitulado "*Predicting Total Movie Grosses After One Week"*, Simonoff analisou o comportamento de todos 147 filmes lançados em mais de 1.000 salas de cinema no ano de 2013 nos Estados Unidos. Sua variável resposta foi a arrecadação final, e suas variáveis explicativas foram a arrecadação do final de semana de estréia e uma variável que indica qual a nota média que os críticos deram para o filme. Esta última variável foi extraída do site RottenTomatoes e também está incluída no presente trabalho. Simonoff conseguiu um modelo com bom ajuste ($R^2$ ajustado de 92.04%) e concluiu que ambas variáveis são estatisticamente significantes para prever a bilheteria final de um filme. Com o intuito de aprimorar este estudo, mais variáveis foram testadas pelo presente trabalho, que utilizou um banco de dados de 700 filmes. No final deste trabalho, tentaremos encontrar um modelo mais completo que possa ajudar a prever a bilheteria final de um filme com base na sua arrecadação de estréia.

Na próxima seção, apresentaremos os objetivos mais detalhadamente. Na seção de Materiais e Métodos, vamos descrever os dados, as variáveis e explicar como o banco de dados foi montado. Na seção Análise Exploratória dos Dados, vamos analisar o comportamento de cada variável e apresentar gráficos e estatísticas descritivas para cada uma. Na seção Resultados e Discussões, apresentaremos os resultados do ajuste dos modelos de regressão, faremos uma análise do modelo final, e destacaremos as limitações do estudo. Por fim, faremos as devidas conclusões e recomendações para futuros estudos.

# 2. OBJETIVOS

## 2.1 *Objetivo Geral:*

O objetivo geral deste trabalho é estimar qual a porcentagem que a arrecadação obtida por um filme em seu final de semana de estréia representa na sua arrecadação final, por meio de um modelo de regressão que possa ser utilizado para estimar a bilheteria final arrecadada em função de variáveis que estejam disponíveis no momento da estréia do filme. Outro objetivo geral deste trabalho, é explorar a relação entre as características de um filme.

## 2.2 *Objetivos Específicos:*

- Localizar quais características de um filme são relevantes na previsão da sua bilheteria final.
- Encontrar um modelo de regressão com bom ajuste, ausência de multicolineridade, erros com distribuição normal e variância constante, e com todas variáveis explicativas relevantes colocadas na forma correta.
- Entender os fatores que influenciam na divulgação to tipo “boca a boca” de um filme, entendido como a porcentagem da bilheteria de estréia na bilheteria final arrecadada.

# 3. MATERIAIS E MÉTODOS

## 3.1 *Descrição dos dados*:

Para este estudo foi montado uma base de dados com 700 filmes e, para cada filme, foram coletados dados em 15 variáveis. Foram selecionadas as 70 maiores bilheterias dos últimos 10 anos completos. Os dados vão de 2006 a 2015 e refletem somente a performance dos filmes no mercado norteamericano. O software estatístico utilizado neste trabalho foi o programa R.

Grande parte da base de dados para este estudo foi obtida através do site BoxOfficeMojo, site líder em acessos com respeito de bilheterias de filmes norteamericanos. Este site é operado pelo IMDb (*Internet Movie Database*), o maior site de banco de dados sobre filmes e cinema do mundo. BoxOfficeMojo é atualizado diariamente com várias novas informações, entre elas a bilheteria diária de todos filmes em cartaz. É referência para o acompanhamento detalhado de diversos tipos de rankings, lançamentos, artigos, etc. O site disponibiliza livre acesso aos dados, de maneira que qualquer pessoa tem acesso ao banco de dados aqui utilizado. As variáveis mais relevantes para o acompanhamento e conhecimento do mercado cinematográfico norteamericano usadas neste estudo, e que foram extraídas do site BoxOfficeMojo, foram:

- Porcentagem da Bilheteria do Final de Semana de Estréia na Bilheteria Final, Bilheteria final, Ano de Estréia, Salas de Estréia, Crescimento de Salas, Mês de Estréia, Gênero, Bilheteria do Final de Semana de Estréia, Studio, Elenco, Premiação, Adaptação e Classificação.

Outros dois sites também foram utilizados para montar o banco de dados. São eles: RottenTomatoes e CinemaScore. Enquanto no primeiro extraímos dados a respeito da opinião dos críticos sobre cada filme, no segundo, os dados dizem respeito à opinião do público. Ambos os sites também são referência na área, e são frequentemente citados em artigos sobre o tema (inclusive em artigos no site BoxOfficeMojo). A priori, há de se imaginar que ambas variáveis são muito importantes no estudo das bilheterias, especialmente quando estamos estudando

indiretamente o “boca a boca” de um filme. Mais adiante no trabalho, veremos se estas suposições são realmente confirmadas.

## 3.2 *Métodos*:

Nesta sub-seção, vamos definir todas as variáveis incluídas no banco de dados. Depois, para que possamos responder as perguntas levantadas e assim entender o comportamento das bilheterias dos filmes norteamericanos, precisamos fazer uma análise exploratória de todas as 15 variáveis do nosso banco de dados, e em seguida, efetuar uma análise de regressão com a variável resposta “Porcentagem da Bilheteria do Final de Semana de Estréia na Bilheteria Final” e 10 variáveis candidatas a explicativas. A análise de regressão é a ferramenta estatística ideal para responder os problemas delineados no presente trabalho.

A interpretação moderna de análise de regressão, de acordo com Gujarati, “diz respeito ao estudo da dependência de uma variável, a variável dependente, em relação a uma ou mais variáveis, as variáveis explanatórias, visando estimar e/ou prever o valor média (da população) da primeira em termos dos valores conhecidos ou fixados (em amostragens repetidas) das segundas” (GUJARATI, 2008, p.39). No nosso caso, a variável dependente, ou variável reposta, é a porcentagem da arrecadação do final de semana de estréia na arrecadação final. Assim, por meio de uma regressão, apresentaremos um modelo para estimar seu valor médio em termos de 10 possíveis variáveis explanatórias conhecidas, as quais testamos para determinar suas respectivas significâncias no modelo. Das 15 variáveis incluídas na análise exploratória dos dados (dentre elas, a variável resposta), três não foram consideradas na análise de regressão, pois estas não são conhecidas no momento após o final de semana de estréia de um filme, e por isso, não teriam utilidade prática no modelo. São elas: Bilheteria Final, Crescimento de Salas e Premiação. A variável Ano de Estréia também não foi considerada no modelo de regressão, uma vez que não podemos reproduzir um ano uma vez que este já se deu por encerrado.

A seguir, apresentaremos a definição de todas variáveis utilizadas no trabalho. Além de apresentar mais detalhadamente cada uma delas, também está detalhado como elas foram coletadas. Em alguns casos, as variáveis foram extraídas de forma

crua. Em outros, sempre que julgado enriquecedor, foi adicionado algum conhecimento a priori para ajustar os valores de variáveis muito subjetivas.

### 3.2.1) Definição das variáveis:

1) **Porcentagem da Bilheteria do Final de Semana de Estréia na Bilheteria Final ("FDS.com.final"):**

Variável resposta do nosso estudo. Necessária para se fazer a análise de regressão, esta variável é do tipo contínua. Visa demonstrar o quanto a bilheteria arrecadada no final de semana de estréia representa na bilheteria final arrecadada. Por meio desta variável e da bilheteria arrecadada no final de semana de estréia, podemos chegar a um valor estimado para a bilheteria final arrecadada.

A arrecadação no final de semana de lançamento de um filme, em média, é responsável por 30.8% da bilheteria final arrecadada, logo, espera-se, a priori, que este dado seja altamente preditivo da bilheteria final. Filmes que apresentam baixos valores percentuais deste dado, indicam que sua demanda se manteve relativamente estável após a estréia ou que até cresceu com o passar do tempo em exibição. Logo, pode-se enxergar esta variável como uma variável que mede a longevidade, ou o "boca a boca" de um filme.

2) **Bilheteria Final ("BilheteriaFinal"):**

Variável contínua, que representa o total, em milhões de dólares, arrecadado pelo filme nos cinemas norteamericanos durante todo o tempo em cartaz. Embora esta variável não esteja incluída diretamente no nosso modelo de regressão, ela foi usada no cálculo dos valores da variável resposta.

3) **Ano de Estréia ("AnoEstreia"):**

Ano em que o filme estreiou. Variável que assume valores entre 2006 a 2015. Caso um filme tenha estreiado nos últimos dois dias do ano, ele foi considerado como do ano seguinte. Esta variável não foi candidata a variável explicativa no modelo de regressão.

4) **Número de Salas de Estréia ("SalasEstreia"):**

Variável discreta que diz respeito ao número de salas em que o filme foi exibido no seu final de semana de estréia. Aqui, só foram considerados filmes que alcançaram status de "*Wide Release*", isto é, filmes cuja exibição ultrapassaram 1.000 salas de cinema em seu final de semana de estréia (esta mesma restrição foi utilizada no artigo de Simonoff). A decisão de somente considerar filmes de larga escala de divulgação acabou eliminando somente 3 filmes do nosso banco de dados. Entretanto, para manter a ideia de incluir 70 filmes por ano analisado, estas exclusões foram compensadas pelos próximos filmes dos rankings anuais de bilheteria final.

5) **Taxa de Crescimento de Salas ("CrescSalas"):**

Esta variável contínua foi calculada a partir da diferença entre o número máximo de salas de exibição atingido ("NumSalasMAX") e o número de salas em exibição durante o final de semana de estréia do filme, dividido pelo número de salas do final de semana de estréia, como mostra a expressão que segue:

[ (NumSalasMAX) – (SalasEstreia) ] / (SalasEstreia)

A variável está expressa em porcentagem. Teoricamente, esta variável seria outra forma de medir o efeito do "boca a boca" de um filme. Entretanto, esta variável pode não ser tão eficaz para medir este efeito, uma vez que um filme de grande expectativa, já estréia em um número alto de salas, tendo assim pouco espaço para ter um crescimento relativo alto, independente de seu "boca a boca". Esta variável também não está incluída no nosso modelo de regressão, mas sua relação com a variável resposta ("FDS.com.final") está detalhada na seção de Análise Exploratória dos Dados.

6) **Mês de Estréia ("MesEstreia"):**

O mês de lançamento de um filme foi analisado por duas perspectivas. A intenção aqui era ter no banco de dados ambas perspectivas, e, somente na hora de analisar o resultado do modelo, selecionar qual perspectiva é mais significante para a previsão da bilheteria de um filme.

Para a primeira perspectiva, foi atribuído um valor percentual para cada mês, calculado da seguinte forma: pegou-se a bilheteria total de cada um dos 12 meses para cada um dos últimos 10 anos. Somou-se todos os valores dos mês de janeiros, todos os valores do mês de fevereiro, e assim por diante. Desta soma, dividiu-se cada mês pela bilheteria total dos últimos 10 anos, para assim saber o peso que cada mês teve,

em média, nos últimos 10 anos. Os resultados (em milhões de dólares), e os pesos (em %) para cada mês foram os seguintes:

**Tabela 1 – Distribuição de Frequência da Bilheteria Segundo o mês de estréia do filme (2006-2015)**

| Mês | Soma das bilheterias no mês (em milhões de dólares) | Peso |
|---|---|---|
| Janeiro: | 3988.3 | 3.86% |
| Fevereiro: | 6333 | 6.14% |
| Março: | 8344.3 | 8.09% |
| Abril: | 5848.7 | 5.67% |
| Maio: | 11675.9 | 11.31% |
| Junho: | 11934 | 11.56% |
| Julho: | 11299.7 | 10.95% |
| Agosto: | 7423.8 | 7.19% |
| Setembro: | 5435.2 | 5.27% |
| Outubro: | 6276.5 | 6.08% |
| Novembro: | 11595.7 | 11.24% |
| Dezembro: | 13042.1 | 12.64% |
| Total: | | **100.00%** |

Pela segunda perspectiva foi testada a utilização de categorias para representar a *temporada* de estréia do filme (variáveis "Mes.DUMMY"). Observando a tabela acima, podemos ver que temos claramente quatro temporadas de estréia de filmes, e essa distinção foi usada para separar as categorias. Vale destacar que a divisão dessas temporadas, embora não uniformemente distribuídas em número de meses, é padrão nos sites aqui usados como referência.

Categoria de Referência → Janeiro a Abril ("*spring")*

Dummy 1 → Indicadora dos meses de Maio, Junho e Julho ("*summer")*

Dummy 2 → Indicadora dos meses de Agosto, Setembro e Outubro ("*fall*")

Dummy 3 → Indicadora dos meses de Novembro e Dezembro ("*winter / holiday*").

Em ambos os casos, caso um filme tenha estreiado nos últimos dois dias do mês, ele foi considerado com sendo do mês seguinte.

**7) Gênero ("Genero"):**

Variável categórica que indica ao gênero do filme. Para ser utilizada no modelo de regressão, foram definidas as seguintes variáveis *dummies*:

Categoria de Referência → Gênero Comédia

Dummy 1 → Indicadora do gênero Drama

Dummy 2 → Indicadora do gênero Animação / Família

Dummy 3 → Indicadora do gênero Ficção Científica / Fantasia

Dummy 4 → Indicadora do gênero Ação / Aventura

Dummy 5 → Indicadora do gênero Terror / Suspense.

Para esta variável, caso um filme tivesse mais de uma classificação de gênero, sendo estas classificações de diferentes grupos, foi considerado o primeiro gênero listado na descrição do filme.

**8) Bilheteria do Final de Semana de Estréia ("FDS.Estreia"):**

Variável contínua que se refere ao valor total da bilheteria arrecadada nos cinemas norteamericanos no final de semana de estréia do filme. Nos casos dos filmes que estreiaram em poucas salas de exibição, e depois passaram a barreira das 1.000 salas ("*Wide Release*"), foi considerado a bilheteria do final de semana em "*Wide Release*" como o final de semana de estréia. Com a informação sobre esta variável e seu respectivo valor estimado pela variável resposta "FDS.com.final", podemos prever a bilheteria final de um filme.

**9) Estúdio ("Studio"):**

Variável categórica que indica os principais estúdios responsáveis pelo patrocínio e distribuição dos filmes. Para ser utilizada no modelo de regressão, foram definidas as seguintes variáveis *dummies*:

Categoria de referência → Estúdio Buena Vista

Dummy 1 → Indicadora do estúdio Universal

Dummy 2 → Indicadora do estúdio Fox

Dummy 3 → Indicadora do estúdio Sony

Dummy 4 → Indicadora do estúdio Paramount

Dummy 5 → Indicadora do estúdio Warner Bros

Dummy 6 → Indicadora de outros estúdios.

**10) Impacto do Elenco ("Elenco"):**

Variável categórica ordinal, para medir o poder de arrecadação que certo elenco traz ao filme. Na maioria dos casos, para calcular esta variável, foi feito o

seguinte cálculo: pegou-se os dois principais protagonistas do filme (ou os dois primeiros atores listados na descrição do filme) mais o diretor, e calculou-se a média de arrecadação dos filmes estrelados por cada um deles. Depois, foi calculado a média da média desses três dados, e, a ao valor dessa média, foi atribuída uma nota da seguinte forma:

Categoria de referência 0 → se a média de arrecadação dos filmes estrelados pelas três pessoas ficou entre 0 e 24 milhões de dólares

Dummy 1 → Indicadora para a categoria na qual a média de arrecadação dos filmes estrelados pelas três pessoas ficou entre 25 e 49 milhões de dólares

Dummy 2 → Indicadora para a categoria na qual a média de arrecadação dos filmes estrelados pelas três pessoas ficou entre 50 e 74 milhões de dólares

Dummy 3 → Indicadora para a categoria na qual a média de arrecadação dos filmes estrelados pelas três pessoas ficou entre 75 e 100 milhões de dólares

Dummy 4 → Indicadora para a categoria na qual a média de arrecadação dos filmes estrelados pelas trêspessoas foi maior do que 100 milhões de dólares.

Entretanto, as notas descritas acima não foram seguidas à risca. Em alguns casos, um filme foi dado uma nota maior do que o cálculo acima indicaria. Este é o caso de filmes com elenco extensivo, como por exemplo os filmes "The Expendables" ("Os Mercenários"), "Midnight in Paris" ("Meia Noite em Paris"), "He's Just Not That Into You" ("Ele não estão tão afim de você") e "Valentine's Day" ("Idas e Vindas do Amor"). Outro caso em que isto ocorreu foram os filmes estrelados por celebridades da televisão ou da música, como por exemplo "Hannah Montana: The Movie", "The Muppets" e "Justin Bieber: Never Say Never". Para filmes de animação, foi utilizado o elenco responsável pelas dublagens dos personagens.

**11) Indicação à Premiação do Oscar ("Premiacao"):**

Variável categórica binária que indica se um filme foi indicado a um Oscar, independentemente de ter ganho, nas seguintes categorias, selecionadas como as mais importantes:

- Melhor filme;
- Melhor diretor;
- Melhor ator / ator coadjuvante;
- Melhor atriz / atriz coadjuvante;
- Melhor roteiro original;

- Melhor roteiro adaptado;

As categorias da variável foram definidas da seguinte forma:

Categoria de referência → se o filme não ganhou, nem foi indicado, à *nenhuma* das categorias citadas acima.

Dummy 1 → Indicadora para o caso de o filme ganhada, ou ter sido indicado, a *qualquer* uma das categorias citadas acima.

Esta variável não foi incluída no modelo de regressão, pois não teria utilidade prática. Afinal de contas, não se sabe se um filme vai ser indicado, muito menos se vencerá qualquer uma das categorias citadas acima no Oscar, até que sua exibição já tenha se encerrado ou esteja próxima de encerrar (a cerimônia de premiação do Oscar normalmente acontece dois meses após o fim do ano de referência).

**12) Adaptação / Continuação ("Adaptacao"):**

Variável categórica binária que indica se um filme é uma adaptação / continuação ou se ele é um filme original. Para que filmes fossem considerados uma continuação, não importou se eles contam histórias que procedem o filme original ("*sequel*"), ou se contam histórias que precedem o filme original ("*prequel*"). Para certo filme ter sido considerado uma adaptação, foi analisado se o filme era baseado em algum livro, conto infantil ou personagem famoso. Um exemplo de um filme que foi considerado adaptação por conta de se tratar de um personagem famoso, mesmo sendo um filme de roteiro original, é o filme "Sherlock Homes". Mas também houve limites na categorização de um filme com respeito a essa variável. Filmes que simplesmente contam a história de uma pessoa, como por exemplo "A Rede Social", ou "O Lobo de Wall Street", não foram considerados adaptação.

As categorias da variável foram definidas da seguinte forma:

Categoria de referência → se o filme for uma adaptação ou uma continuação

Dummy 1 → Indicadora para o caso de o filme não ser nem uma adaptação nem uma continuação.

**13) Classificação Etária ("Rating"):**

Variável categórica que indica qual a classificação etária do filme. O sistema de classificação utilizado foi o usado atualmente pelo *Motion Picture Association of America*, e é a classificação que fica exposta em todos os cartazes norteamericanos. A variável foi separada da seguinte forma:

Categoria de Referência → Classificação "G" (*General Audiences*): todas idades permitidas

Dummy 1 → Indicadora para a Classificação "PG" (*Parental Guidance Suggested*): algum material não apropriado a crianças

Dummy 2 → Indicadora para a Classificação "PG-13" (*Parents Strongly Cautioned*): algum material inapropriado para crianças menores de 13 anos

Dummy 3 → Indicadora para a Classificação "R" (*Restricted*): menores de 17 anos precisam estar acompanhados dos pais ou dos responsáveis.

Existe mais uma classificação oficial: a "NC-17". Nesta classificação, nenhuma pessoa menor de 17 anos é permitida entrada. Entretanto, não houve nenhum filme com essa classificação no nosso banco de dados.

**14) Avaliação dos críticos ("RottenTomatoes"):**

Neste trabalho, iremos analisar a relação entre a opinião dos críticos e a bilheteria de um filme. Para isso, foi incluída esta variável discreta, que varia de 0 a 100 pontos, e representa a nota média que os principais críticos de cinema norteamericano deram ao filme. O site *Rotten Tomatoes*, muito conhecido e citado em vários artigos online de previsão de bilheteria de filmes, é a referência que compilha todas as críticas de um certo filme e atribui esta nota. Todos os dados desta variável foram retirados deste site.

**15) Avaliação do público ("CinemaScore"):**

Queremos também analisar a relação entre a opinião do público e a bilheteria de um filme. Todos os dados desta variável foram retirados de um site que também é considerado referência para este tipo de informação. Este site chama-se *Cinema Score*. A nota atribuída a um filme é calculada da seguinte maneira: a empresa responsável aplica um questionário a pessoas que acabaram de assistir ao filme na sua estréia, e pedem para elas dêem um conceito variando de F até A+. A média dessas notas é publicado no site. Como a nota média do público atribuída a um filme é disponível somente em conceitos, para este trabalho, foi dada a seguinte nota correspondente (assim, temos uma variável que varia entre 56 e 99):

**Tabela 2 – Conceitos atribuídos aos filmes pelo CinemaScore e suas respectivas Notas**

| Conceito | Valor Atribuído |
|---|---|
| A+ | 99 |
| A | 96 |
| A- | 93 |
| B+ | 89 |
| B | 86 |
| B- | 83 |
| C+ | 79 |
| C | 76 |
| C- | 73 |
| D+ | 69 |
| D | 66 |
| D- | 63 |
| F | 56 |

# 4. RESULTADOS E DISCUSSÕES:

## 4.1 ANÁLISE EXPLORATÓRIA DOS DADOS:

Nesta seção, faremos uma análise das estatísticas descritivas de cada variável. Como cada variável tem uma natureza diferente, seja ela contínua ou discreta, categórica ou não, as medidas de posição e dispersão apresentadas serão as mais adequadas para cada tipo de variável. Além de estatísticas descritivas individuais, também analisaremos a relação entre cada uma das variáveis com nossa variável dependente, a porcentagem do final de semana de estréia na bilheteria final.

Estudaremos a correlação entre elas, acompanhado de histogramas, boxplots ou gráficos de dispersão. Por fim, sempre se que julgou enriquecedor entender como as variáveis explicativas se relacionavam uma com as outras, foi feita uma análise conjunta de duas variáveis, independentemente se tais informações forem relevantes para o modelo de regressão apresentado na próxima seção. No final desta seção, ficará evidenciado o comportamento e o relacionamento entre variáveis importantes para melhor entendermos o mercado cinematográfico norteamericano.

## 1) Porcentagem da Bilheteria do Final de Semana de Estréia na Bilheteria Final ("Fds.com.final"):

A Tabela 3 apresenta as estatísticas descritivas para a variável porcentagem da bilheteria do final de semana de estréia na bilheteria final. As figuras 1 e 2 apresentam a representação gráfica da distribuição dos valores dessa variável utilizando os gráficos histograma e da caixa (boxplot), respectivamente.

**Tabela 3 - Estatísticas Descritivas de "FDS.com.final"**

| Mínimo | 1º Quartil | Mediana | Média | 3º Quartil | Máximo | Variância | Desvio Padrão | Coef. de Variação |
|---|---|---|---|---|---|---|---|---|
| 0.0554 | 0.2500 | 0.3048 | 0.3079 | 0.3679 | 0.6333 | 0.0087 | 0.0933 | 30.30% |

**Figura 1 – Histograma de FDS.com.final**

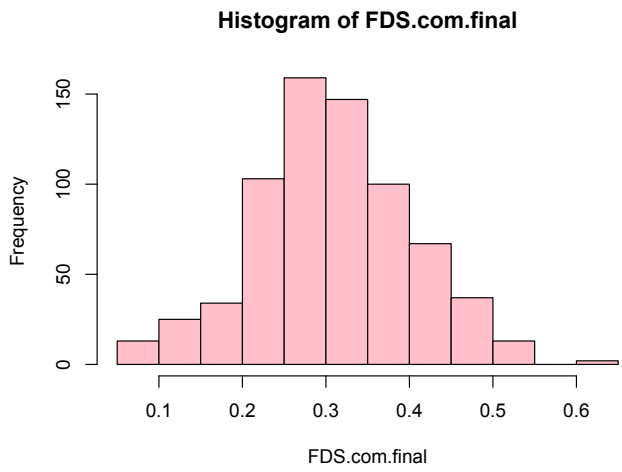
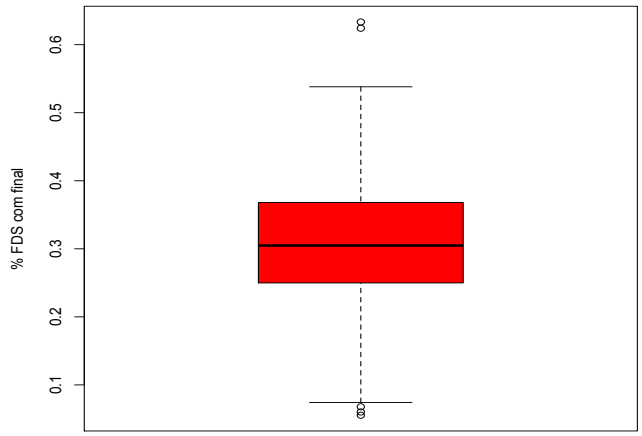



**Figura 2 – BoxPlot de FDS.com.final**




Muito pode-se dizer sobre o "boca a boca" de um filme por esta variável. Ela nos diz o peso arrecadado no final de semana de estréia relativo à bilheteria final de um filme. As estatísticas descritivas nos mostram que o menor peso que um final de semana de estréia teve na bilheteria final de um filme foi de 5.54%, enquanto o maior peso foi de 63.33%. O desvio padrão de 9.33% nos leva a um Coeficiente de Variação de 30.3%. O histograma e o boxplot nos mostram que esta variável tem uma distribuição quase simétrica, com uma leve assimetria à direita (a mediana é levemente menor que a média). O teste de normalidade de Shapiro-Wilk obteve um p-valor de 0.035, quase acima do nível de significância de 5%, o que nos leva à rejeição da hipótese de normalidade para a distribuição dos valores desta variável.

## 2) Bilheteria Final (“BilheteriaFinal”):

A Tabela 4 apresenta as estatísticas descritivas para a variável Bilheteria Final. As figuras 3 e 4 apresentam a representação gráfica da distribuição dos valores dessa variável utilizando os gráficos histograma e da caixa (boxplot), respectivamente.

**Tabela 4 - Estatística Descritiva "BilheteriaFinal" (em milhões de dólares)**

| Mínimo | 1º Quartil | Mediana | Média | 3º Quartil | Máximo | Variância | Desvio Padrão | Coef. de Variação |
|---|---|---|---|---|---|---|---|---|
| 38.32 | 59.94 | 85.07 | 118.30 | 142.70 | 936.70 | 8561.37 | 92.53 | 78.22% |

**Figura 3 – Histograma de BilheteriaFinal**

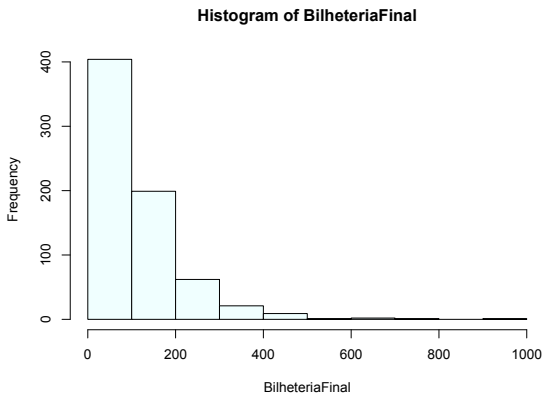
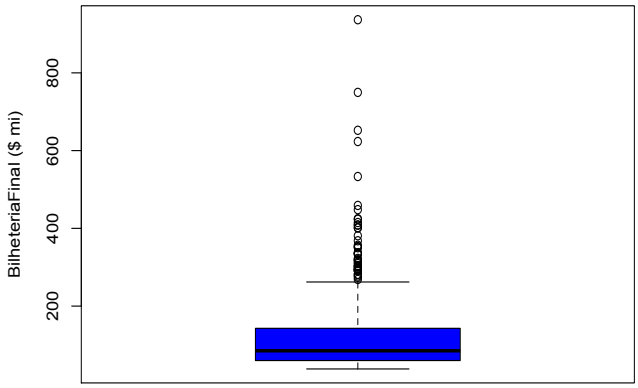


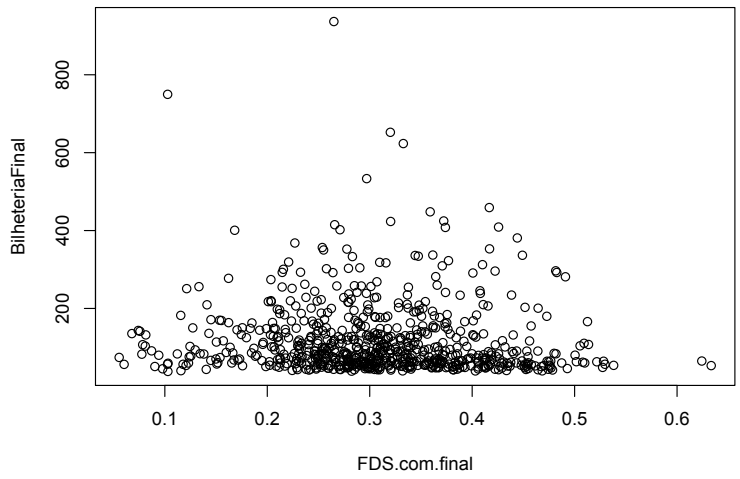

**Figura 4 – Boxplot de BilheteriaFinal**




As estatísticas descritivas acima nos mostram que o menor filme do nosso banco de dados rendeu U$ 38.32 milhões de dólares, isto é, o pior septuagéssimo colocado dos últimos 10 anos. A mediana é bastante menor que a média, e isto é evindenciado pelo histograma, que nos mostra uma distribuição assimétrica positiva a direita. O desvio padrão de 92.53 nos leva a um Coeficiente de Variação de 78.22%. Uma vez que pegamos somente as 70 maiores bilheterias de cada ano, era de se esperar que o boxplot somente apresentasse dados discrepantes em sua parte superior. Note que o intervalo interquartílico é pequeno em comparação com toda a amplitude dos nossos dados.

Entretanto, embora esta variável não esteja incluída diretamente no nosso modelo de regressão, queremos estudar sua relação com nossa variável dependente. Sabemos que, com a bilheteria arrecadada no final de semana de estréia e com o valor estimado da nossa variável dependente, podemos prever a bilheteria final de um filme. Mas qual a relação entre a bilheteria final arrecadada e o percentual do final de semana de estréia na bilheteria final? Esta pergunta pode ser respondida por meio da análise do diagrama de dispersão para os dados das duas variáveis, apresentado na Figura 5. Veja que, embora as maiores bilheterias tiveram valores relativamente baixos de "FDS.com.final", não há uma clara tendência de relacionamento entre as duas variáveis. O coeficiente de correlação linear de Pearson entre as duas variáveis é **-0.073**, o que corrobora com esta falta de correlação entre as duas variáveis.

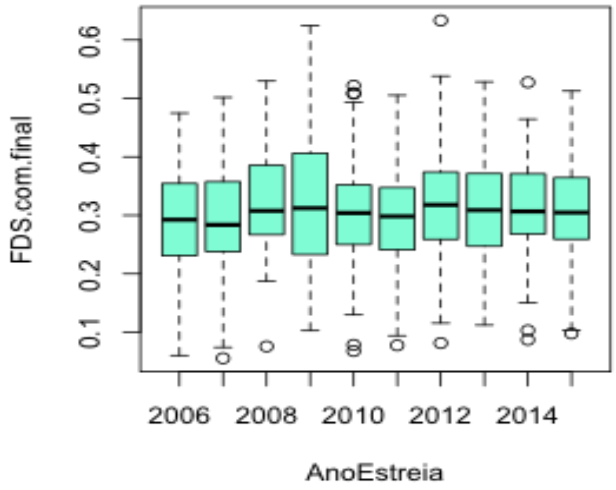
**Figura 5 – Relação entre BilheteriaFinal e FDS.com.final**




**3) Ano de Estréia ("AnoEstreia"):**

Como nossos dados não foram extraídos aleatoriamente, uma vez que pegamos somente as 70 maiores bilheterias dos últimos 10 anos, não faria sentido falar em estatísticas descritivas aqui (obviamente, temos 70 observações para cada ano, entre 2006 e 2015). Na Figura 6, apresentamos a distribuição da variável "FDS.com.final" segundo AnoEstreia:

**Figura 6 – Boxplot de AnoEstréia**




O boxplot acima não nos evidencia nenhuma tendência linear no comportamento dos 70 maiores filmes nos últimos 10 anos. Isto é, não podemos dizer que ao passar dos anos, nossa variável resposta tem uma queda ou crescimento no seu valor esperado. Entretanto, podemos notar que o ano de 2009 foi o ano que apresentou maior intervalo interquartílico entre os 10 anos analisados.

**4) SalasEstreia ("SalasEstreia"):**

A Tabela 5 apresenta as estatísticas descritivas para a variável Salas de Estréia. As Figuras 7 e 8 apresentam a representação gráfica da distribuição dos valores dessa variável utilizando os gráficos histograma e da caixa (boxplot), respectivamente.

**Tabela 5 - Estatísticas Descritivas de "SalasEstreia"**

| Mínimo | 1º Quartil | Mediana | Média | 3º Quartil | Máximo | Variância | Desvio Padrão | Coef. de Variação | Assimetria | Curtose |
|---|---|---|---|---|---|---|---|---|---|---|
| 1003 | 2894 | 3222 | 3203 | 3641 | 4468 | 402491 | 634 | 19.79% | -0.83 | 1.16 |

**Figura 7- Histograma de SalasEstreia**

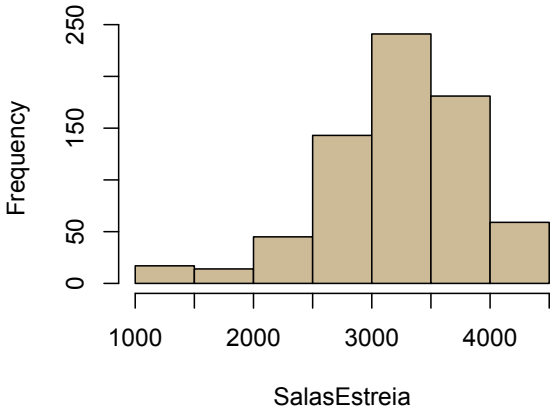
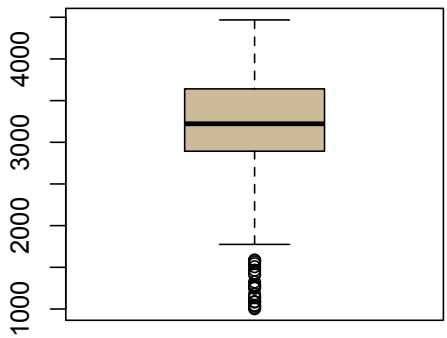


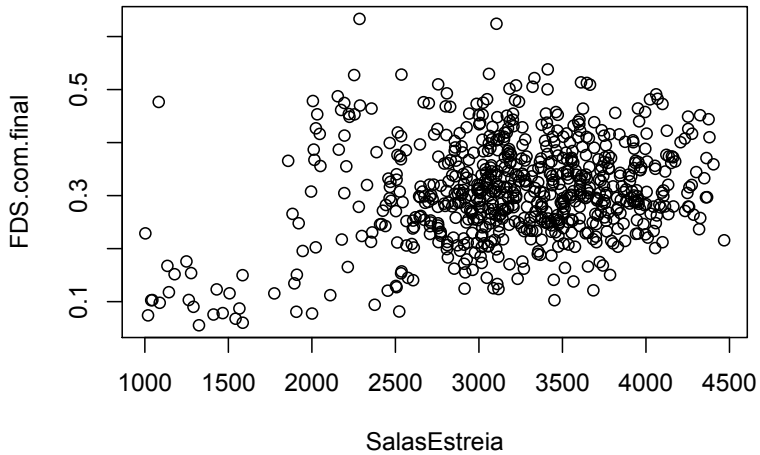

**Figura 8- Boxplot de SalasEstreia**




A variável "SalasEstreia", que representa o número de salas que certo filme foi exibido no seu final de semana de estréia, é uma variável discreta, com uma distribuição de assimetria negativa e à esquerda, como vemos na Figura 7. A mediana é maior que a média, e o coeficiente de variação é de 19.79%. Há dados atípicos somente na parte inferior do boxplot. Isto nos indica que, para um filme chegar nas 70 maiores bilheterias do ano, ele provavelmente precisará de ser exibido em, aproximadamente, mais de 1800 salas de cinema (representado pela linha que separa os dados típicos do atípicos no boxplot) no seu final de semana de estréia.

O coeficiente de correlação linear de Pearson com a variável dependente FDS.com.final é de 0.252. Filmes que estréiam em muitas salas tendem a ser filmes mais aguardados, com maiores legiões de fãs. Estes filmes, na maioria das vezes, acabam tendo um peso forte na bilheteria do final de semana de estréia (alguns exemplo são os filmes das séries Harry Potter, Crepúsculo, Jogos Vorazes). Devido a isso, a correlação é positiva. A Figura 9 evidencia que, embora seja fraca a relação, filmes que estréiam em muitas salas tendem a ter um peso maior de arrecadação no seu final de semana de estréia.

**Figura 9- Relacao entre SalasEstreia e FDS.com.final**



### 5) Taxa de Crescimento de Salas (“CrescSalas”):

A Tabela 6 apresenta as estatísticas descritivas para a variável Crescimento de Salas. As figuras 10 e 11 apresentam a representação gráfica da distribuição dos valores dessa variável utilizando os gráficos histograma e da caixa (boxplot), respectivamente.

**Tabela 6 - Estatística Descritiva "CrescSalas"**

| Mínimo | 1º Quartil | Mediana | Média | 3º Quartil | Máximo | Variância | Desvio Padrão | Coef. de Variação |
|---|---|---|---|---|---|---|---|---|
| 0.00% | 0.00% | 0.30% | 3.05% | 1.81% | 148.70% | 1.18% | 10.88% | 356.72% |

**Figura 10 - Histograma de CrescSalas**

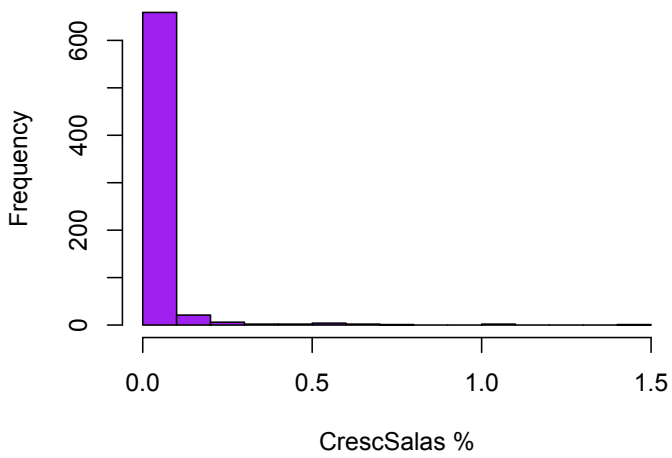
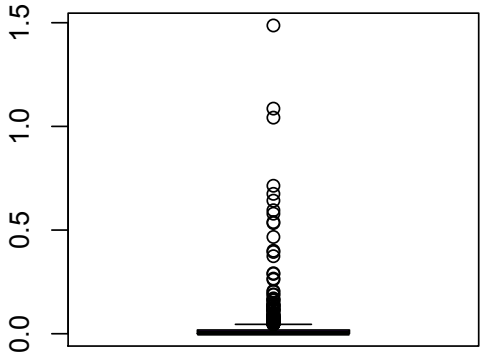



**Figura 11 - Boxplot de CrescSalas**




Relembrando o que foi dito na última seção, a variável contínua “CrescSalas” foi calculada da seguinte forma:

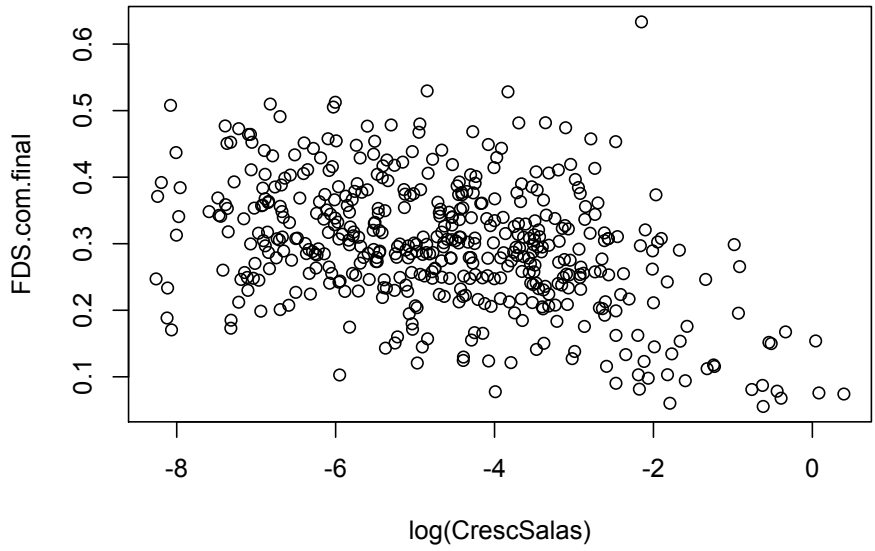
[ (NumSalasMAX) – (SalasEstreia) ] / (SalasEstreia)

A variável está em termos percentuais. As estatísticas descritivas, o histograma e boxplot nos mostram uma distribuição bastante assimétrica à direita (o coeficiente de assimetria equivale a 7.81). A variável apresenta muitos dados discrepantes positivos, o que é evidenciado por uma mediana próxima do primeiro quartil, e uma média bem acima do terceiro quartil. Também devido a esses dados discrepantes, temos um alto desvio padrão, e consequentemente, um alto coeficiente de variação. Dos 700 filmes

do banco de dados, 223 não tiveram nenhum aumento no número de salas em exibição após sua estréia e estão representados pela barra mais alta do histograma. A distribuição é bastante leptocúrtica, com uma curtose de 76.11.

O coeficiente de correlação linear de Pearson entre CrescSalas e FDS.com.final foi negativa, quase moderada e com coeficiente de correlação linear igual a -0.347. Afinal de contas, se um filme tem um crescimento grande de salas de exibição durante seu tempo em cartaz, é de se esperar que a bilheteria arrecadada no final de semana de estréia tende a representar uma parcela menor da bilheteria final (o que também explica o negativo no coeficiente de correlação). A Figura 12 mostra o gráfico de dispersão entre o logarítimo de CrescSalas e FDS.com.final, que corrobora com este coeficiente de correlação linear negativo e quase moderado.

**Figura 12 - Relacao entre logCrescSalas e FDS.com.final**




**6) Mês de Estréia ("MesEstreia")*:**

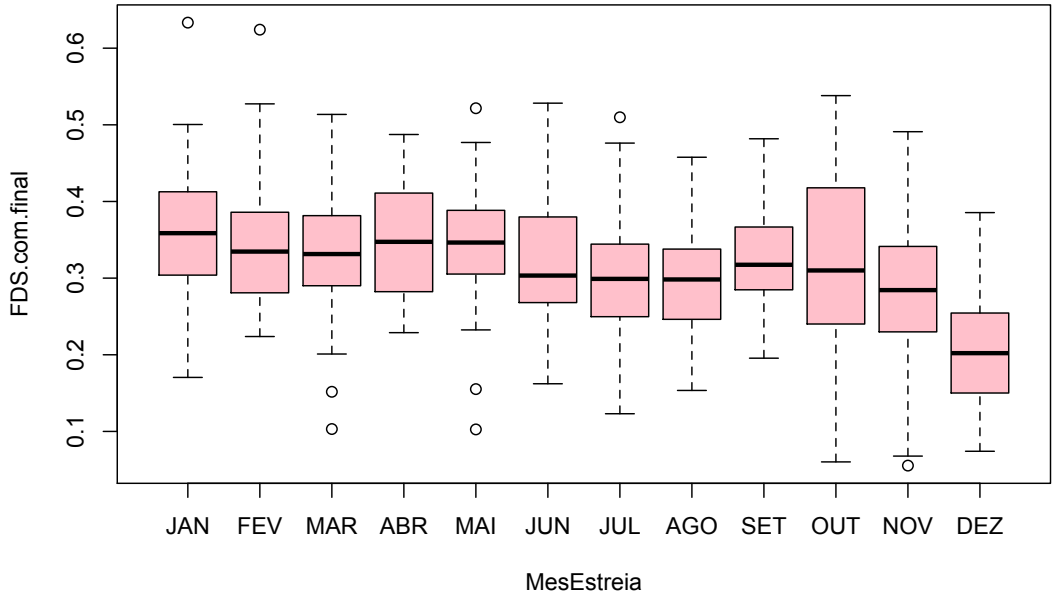
Como vimos na última sessão, esta variável foi analisada por dua perspectivas. Primeiramente, vamos analisar o mês de estréia de um filme pelos pesos percentuais de referência já mencionado. Depois, vamos analisar o mês de estréia pela ótica das temporadas.

A Tabela 7 apresenta a distribuição de frequência para a variável Mês de Estréia. A Figura 13 apresenta a representação gráfica da distribuição dos valores dessa variável utilizando o gráfico da caixa (boxplot).

**Tabela 7 - Frequência de MesEstreia**

| Mês de Estréia | Valor Referência | Frequência | Frequência Relativa |
|---|---|---|---|
| 1- Janeiro | 0.0386 | 35 | 5.00% |
| 2- Fevereiro | 0.0614 | 60 | 8.57% |
| 3- Março | 0.0809 | 60 | 8.57% |
| 4- Abril | 0.0567 | 45 | 6.43% |
| 5- Maio | 0.1131 | 59 | 8.43% |
| 6- Junho | 0.1156 | 72 | 10.29% |
| 7- Julho | 0.1095 | 66 | 9.43% |
| 8- Agosto | 0.0719 | 54 | 7.71% |
| 9- Setembro | 0.0527 | 39 | 5.57% |
| 10- Outubro | 0.0608 | 49 | 7.00% |
| 11- Novembro | 0.1124 | 74 | 10.57% |
| 12- Dezembro | 0.1264 | 87 | 12.43% |

**Figura 13- Boxplot de MesEstreia**




A tabela de freqüência nos mostra que janeiro é o mês menos frequente e com menor valor mensal de referência e dezembro é o mês que mais apresenta filmes na lista de 70 maiores bilheterias e tem o maior valor mensal de referência entre todos outros meses.

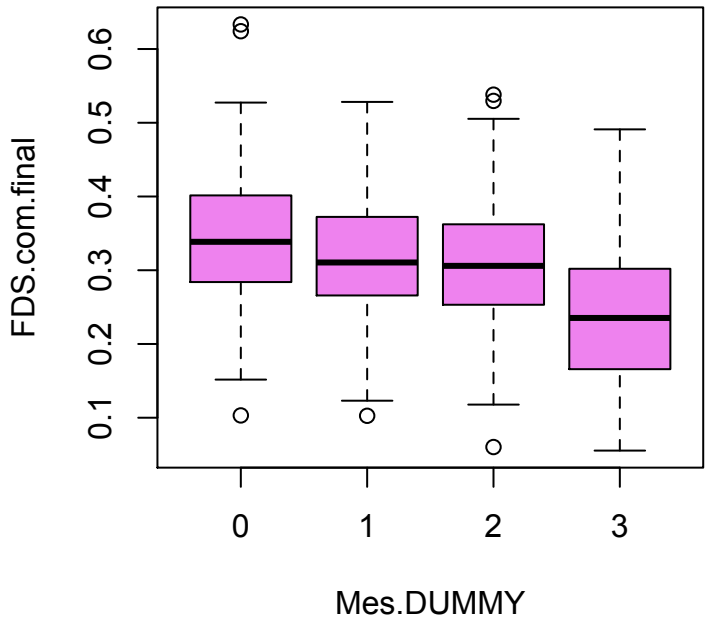
O boxplot de FDS.com.final e MesEstreia nos mostra uma tendência decrescente das medianas com o passar dos meses. Dezembro se destaca, sendo o mês onde filmes apresentam arrecadação mais bem distribuída, e possui os menores valores de primeiro quartil, mediana e terceiro quartil. A correlação entre FDS.com.final e MesEstreia é negativa e equivale a -0.310.

Agora, analisando o mês de estréia pela ótica de temporadas de estréia, podemos observar a Tabela 8, que traz a distribuição de frequência dos filmes por temporada, e a Figura 14, que traz o gráfico da caixa:

**Tabela 8 - Frequência de Mes.DUMMY**

| Categoria (Temporada) | Frequência | Frequência Relativa |
|---|---|---|
| 0 (Janeiro a Abril) | 201 | 28.71% |
| 1 (Maio a Julho) | 198 | 28.29% |
| 2 (Agosto a Outubro) | 141 | 20.14% |
| 3 (Novembro e Dezembro) | 160 | 22.86% |

**Figura 14- Boxplot de Mes.DUMMY**




A Tabela 8 nos mostra que a quantidade de filmes está bem distribuída entre as quatro temporadas. Entretanto, cada temporada tem quantidade de meses diferente, e, por isso, a análise mensal de freqüência ou representatividade no nosso banco de

dados parece ser melhor pela primeira perspectiva. Tanto pela Tabela 7 quanto pela Tabela 8, podemos concluir que novembro e dezembro são os meses com maior representatividade no nosso banco de dados.

O boxplot de FDS.com.final e Mes.DUMMY também nos mostra uma tendência temporal negativa. À medida que o ano passa, os filmes tendem a terem arrecadação mais bem distribuídas e menos concentradas no final de semana de estréia, especialmente na temporada representada pelos meses de novembro e dezembro, que possui os menores valores de primeiro quartil, mediana e terceiro quartil.

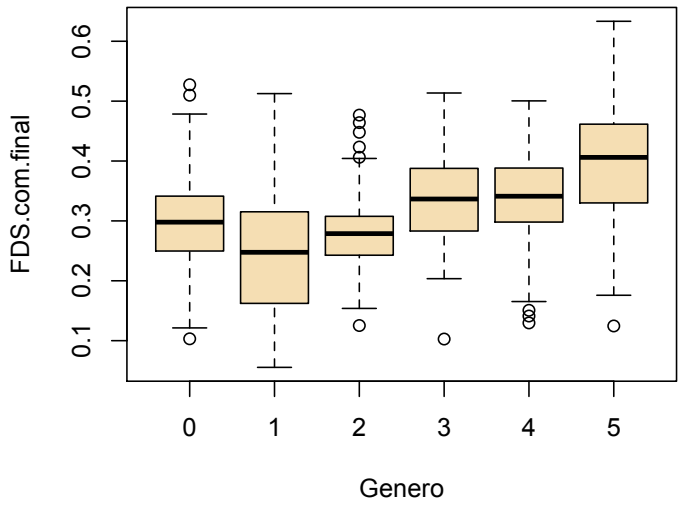
### 7) Gênero ("Genero"):

A Tabela 9 apresenta a distribuição de frequência para a variável Gênero. A Figura 15 apresenta a representação gráfica da distribuição dos valores dessa variável utilizando o gráfico da caixa (boxplot).

**Tabela 9 - Frequência de Genero**

| Gênero | Frequência | Frequência Relativa |
|---|---|---|
| 0 (Comédia) | 162 | 23.14% |
| 1 (Drama) | 134 | 19.14% |
| 2 (Animação / Família) | 113 | 16.14% |
| 3 (SciFi / Fantasia) | 77 | 11.00% |
| 4 (Ação / Aventura) | 137 | 19.57% |
| 5 (Terror / Suspense) | 77 | 11.00% |

**Figura 15- Boxplot de Genero**



Pela tabela de freqüência podemos ver que Comédia é o gênero mais frequente entre as 70 maiores bilheterias dos últimos 10 anos, enquanto empatados, em último, temos os gêneros Sci fi/Aventura e Terror/Suspense.

Pelo boxplot entre Gênero e a variável dependente FDS.com.final, podemos concluir que, em geral, Drama é o gênero com maior longevidade de arrecadação, cujo final de semana de estréia representa menor parcela da bilheteria final (sua mediana equivale a 25%). Terror/Suspense é o gênero com bilheteria arrecadada mais concentrada no final de semana de estréia (sua mediana equivale a 41%). Devido ao fato das caixas da categoria 3, SciFi/Aventura, e categoria 4, Ação/Aventura, serem muito semelhantes, e para facilitar a interpretação do nosso modelo de regressão, vamos agrupar ambos gêneros em somente uma categoria dummy. Desta forma, para nosso modelo de regressão, que será apresentado no próxima sub-seção, as novas categorias de gênero serão:

Categoria de Referência → Gênero Comédia

Dummy 1 → Indicadora do gênero Drama

Dummy 2 → Indicadora do gênero Animação / Família

Dummy 3 → Indicadora do gênero Ficção Científica / Fantasia e Ação / Aventura

Dummy 4 → Indicadora do gênero Terror / Suspense.

### 8) Bilheteria do Final de Semana de Estréia ("FDS.Estreia"):

A Tabela 10 apresenta as estatísticas descritivas para a variável Bilheteria do Final de Semana de Estréia. As Figuras 16 e 17 apresentam a representação gráfica da distribuição dos valores dessa variável utilizando os gráficos histograma e da caixa (boxplot), respectivamente.

**Tabela 10 - Estatísticas Descritivas de FDS.Estreia (em milhões de dólares)**

| Mínimo | 1º Quartil | Mediana | Média | 3º Quartil | Máximo | Variância | Desvio Padrão | Coef. de Variação | Assimetria | Curtose |
|---|---|---|---|---|---|---|---|---|---|---|
| 3.40 | 18.51 | 25.97 | 35.80 | 40.88 | 247.97 | 902.73 | 30.05 | 83.94% | 2.77 | 10.28 |

**Figura 16- Histograma de FDS.Estreia**

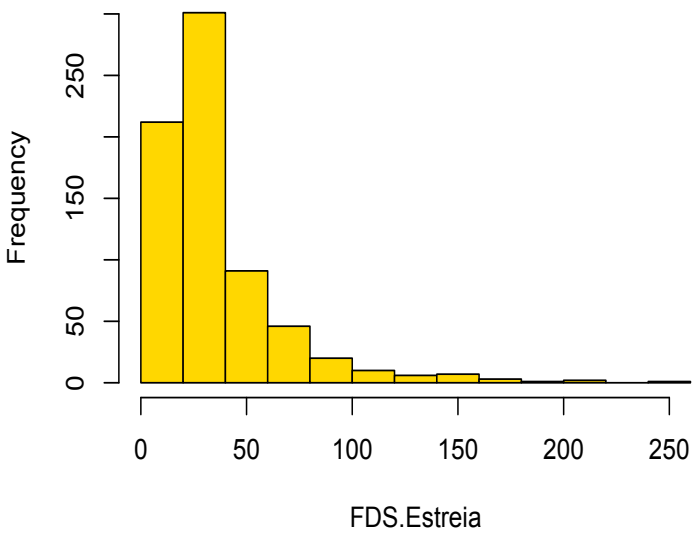
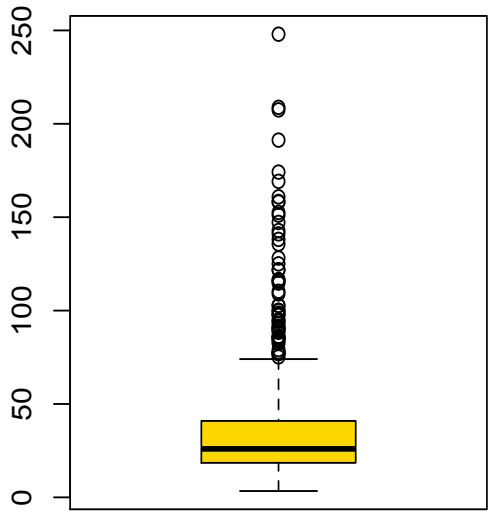


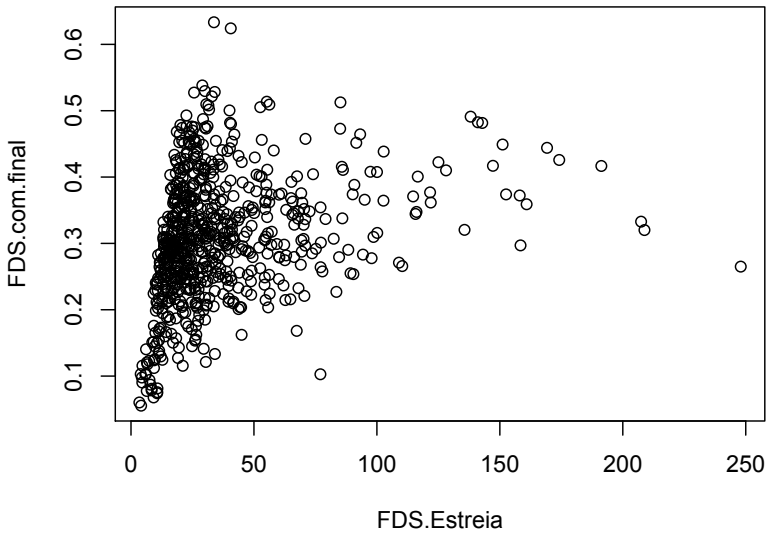

**Figura 17- Boxplot de FDS.Estreia**




A variável FDS.Estreia, que representa a bilheteria arrecadada no final de semana de estréia, tem uma distribuição assimétrica à direita e positiva, com uma mediana menor do que a média, influenciada por vários dados atípicos positivos. A assimetria é confirmada por um coeficiente de assimetria de 2.77, e a distribuição também é leptocúrtica, com um coeficiente de curtose de 10.28. O coeficiente de variação é alto, sendo que o desvio padrão representa quase 84% da média.

A correlação linear de Pearson entre FDS.Estreia com a variável dependente FDS.com.final é positiva, fraca e equivale a 0.273. Um fato que pode explicar a correlação ser positiva é que filmes que estréiam com grandes arrecadações podem acabar exaurindo sua demanda muito rápido. Para melhor visualizar a relação entre as duas variáveis, a Figura 18 mostra o gráfico de dispersão entre elas.

**Figura 18- Relacao entre FDS.Estreia e FDS.com.final**



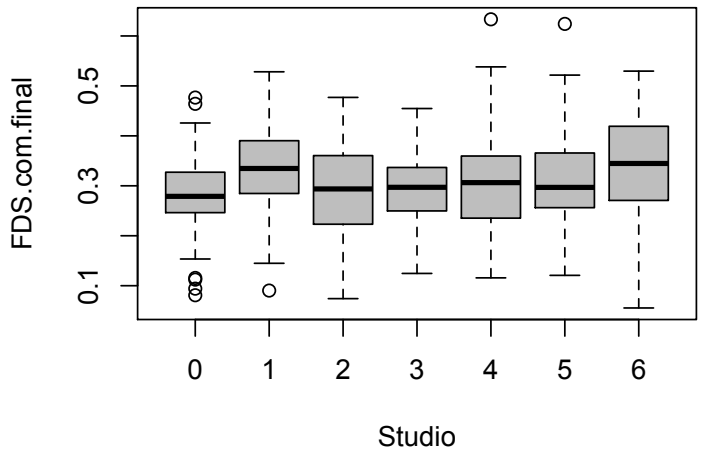
## 9) Estúdio ("Studio"):

A Tabela 11 apresenta a distribuição de frequência para a variável Studio. A Figura 19 apresenta a representação gráfica da distribuição dos valores dessa variável utilizando o gráfico da caixa (boxplot).

**Tabela 11 - Frequência de Studio**

| Studio | Frequência | Frequência Relativa |
|---|---|---|
| 0 (Buena Vista) | 116 | 16.57% |
| 1 (Universal) | 89 | 12.71% |
| 2 (Fox) | 103 | 14.71% |
| 3 (Sony) | 87 | 12.43% |
| 4 (Paramount) | 64 | 9.15% |
| 5 (Warner Bros) | 123 | 17.57% |
| 6 (Outros) | 118 | 16.86% |

**Figura 19- Boxplot de Studio**




A variável Studio foi dividida em sete categorias, sendo seis categorias para os principais estúdios norteamericanos e uma categoria que engloba todos os outros estúdios. Pela Tabela 11, podemos ver que Warner Bros e Buena Vista são os estúdios mais freqüentes nas 70 maiores bilheterias dos últimos 10 anos, enquanto Paramount foi o studio que menos apareceu no banco de dados.

O boxplot de FDS.com.final segundo o Studio não evidencia nenhuma diferença muito significativa entre os estúdios, isto é, não parece existir uma clara influência entre o estúdio de um filme e sua longevidade de arrecadação. Entretanto, podemos dizer que o estúdio Universal apresenta maior mediana entre os seis

principais estúdios, e portanto, seus filmes tendem a ter menor longevidade de arrecadação. Assim como em Gênero, devido ao fato das caixas das categorias 0 (Buena Vista), 2 (Fox), 3 (Sony), 4 (Paramount) e 5 (Warner) serem muito semelhantes, e para facilitar a interpretação do nosso modelo de regressão, vamos agrupar estes estúdios em somente uma categoria dummy. Desta forma, para nosso modelo de regressão, que será apresentado no próxima sub-seção, as novas categorias de estúdio serão:

Categoria de Referência → Indicadora dos principais estúdios, com exceção da Universal (que passa a ser a categoria de referência)

Dummy 1 → Indicadora do estúdio Universal

Dummy 2 → Indicadora de outros estúdios.

**10) Impacto do Elenco ("Elenco"):**

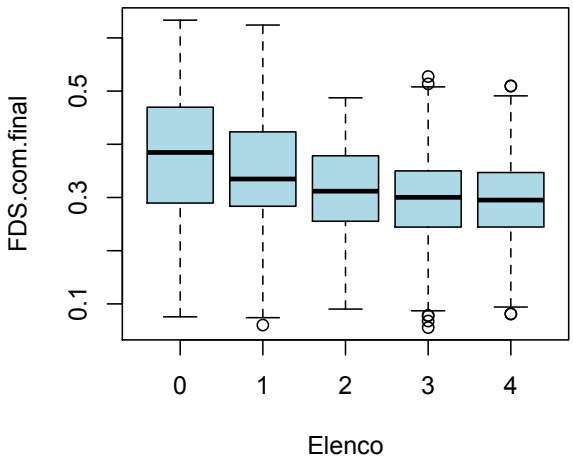
A Tabela 12 apresenta a distribuição de frequência para a variável Elenco. A Figura 20 apresenta a representação gráfica da distribuição dos valores dessa variável utilizando o gráfico da caixa (boxplot).

**Tabela 12 - Distribuição de Frequências dos Filmes Segundo Nota do Elenco**

| Nota do Elenco | Frequência | Frequência Relativa |
|---|---|---|
| 0 ( x < 25 milhões de doláres) | 39 | 5.57% |
| 1 (x entre 25 e 49 milhões de dólares) | 66 | 9.43% |
| 2 (x entre 50 e 74 milhões de dólares) | 149 | 21.29% |
| 3 (x entre 75 e 99 milhões de doláres) | 167 | 23.86% |
| 4 (x > 100 milhões de doláres) | 279 | 39.86% |

* x é a arrecadação média obtida pelos dois principais protagonistas e o diretor do filme.

**Figura 20- Boxplot de Elenco**



A Tabela 12 nos mostra a distribuição dos valores da porcentagem da bilheteria do fina de semana de estréia na bilheteria final segundo cada uma das categorias Elenco. Como somente estamos trabalhando com as 70 maiores bilheterias dos últimos 10 anos (e a média de todos septuagéssimos colocados é de US$ 43.926 milhões), grande parte dos filmes tem como elenco que se encaixa nas categorias dummy acima de 2 (para ser exato, 595 dos 700 dados são da categoria de elenco 2, 3 ou 4). A Tabela 12 corrobora com essa informação.

O boxplot de FDS.com.final segundo as notas de Elenco nos mostra que, à medida que a nota do Elenco aumenta, maior é a tendência de certo filme ter uma longevidade maior de arrecadação, evidenciando-se uma tendência de correlação negativa entre as duas variáveis.

A variável Nota do Elenco é uma variável ordinal, e assim podemos analisar o coeficiente de correlação com a variável dependente. A correlação entre elas é negativa e fraca e equivale a -0.222. A negatividade deste coeficiente é corroborada com a tendência decrescente exposta pelo boxplot.

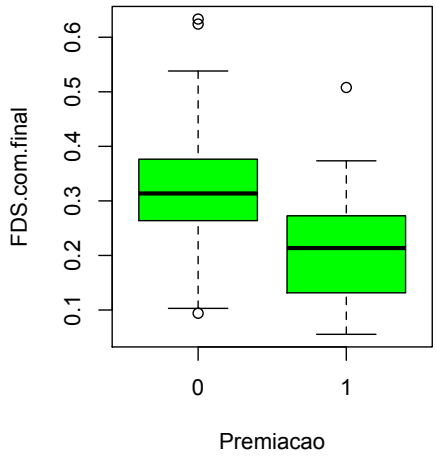
### 11) Indicação à Premiação do Oscar ("Premiacao"):

A Tabela 13 apresenta a distribuição de frequência para a variável Premiação. A Figura 21 apresenta a representação gráfica da distribuição dos valores dessa variável utilizando o gráfico da caixa (boxplot).

**Tabela 13 - Frequência de Premiação**

| Premiação | Frequência | Frequência Relativa |
|---|---|---|
| 0 (Negativo) | 612 | 87.4% |
| 1 (Positivo) | 88 | 12.6% |

**Figura 21- Boxplot de Premiacao**



A Tabela 13 nos mostra que, no nosso banco de dados, temos muito mais filmes que não receberam nenhuma indicação às categorias do Oscar selecionadas como mais relevantes do que filmes que receberam alguma indicação. Podemos também dizer que 12.6% das 70 maiores bilheterias dos últimos 10 anos recebeu a alguma indicação ao Oscar.

O boxplot dos valores de FDS.com.final segundo premiação mostra que todos os três quartis de filmes que receberam alguma indicação importante são menores que na outra categoria de filmes, indicando uma clara separação na distribuição dos valores da variável resposta nos dois grupos de filmes. Filmes que recebem alguma indicação ao Oscar tendem a ter uma bilheteria mais distribuída e menos concentrada no seu final de semana de estréia.

Abaixo, a Tabela 14 traz a distribuição de frequência dos filmes segundo ano de estreia e categoria de premiação.

**Tabela 14- Premiação por Ano**

| Ano de Estréia | 0 (Negativo) | 1 (Positivo) |
|---|---|---|
| 2006 | 60 | 10 |
| 2007 | 60 | 10 |
| 2008 | 65 | 5 |
| 2009 | 63 | 7 |
| 2010 | 61 | 9 |
| 2011 | 64 | 6 |
| 2012 | 59 | 11 |
| 2013 | 61 | 9 |
| 2014 | 61 | 9 |
| 2015 | 58 | 12 |

O ano de 2008 foi o ano que teve menos filmes indicados ao Oscar nas 70 maiores bilheterias, e o ano de 2015 foi o ano que mais teve indicações no top 70 filmes.

Interessante destacar também a relação entre a premiação de um filme e sua temporada de estréia. Lendo artigos sobre o tema, é comum depararmos com o termo "*temporada de filmes do Oscar*": filmes que se especulam com maiores chances de concorrer ao Oscar querem ser lançados no final do ano, talvez por estarem frescos na cabeça do júri da premiação. Para verificar a veracidade desta afirmação, veja a Tabela 15, que traz a distribuição de frequência dos filmes estudados segundo sua temporada de estreia e seu status de indicação ao Oscar.

**Tabela 15 - Premiação por Temporada**

| Temporada de Estréia | 0 (Negativo) | 1 (Positivo) | Total | % Positivo |
|---|---|---|---|---|
| 0 (Janeiro a Abril) | 195 | 6 | 201 | 3.0% |
| 1 (Maio a Julho) | 181 | 17 | 198 | 8.6% |
| 2 (Agosto a Outubro) | 124 | 17 | 141 | 12.1% |
| 3 (Novembro e Dezembro) | 112 | 48 | 160 | 30.0% |

A tabela nos evidencia que novembro e dezembro realmente constituem esta "*temporada de filmes do Oscar*". Filmes que são indicados ao Oscar têm muito mais chances de terem sido lançados nos últimos meses do ano do que no início. Ou então, por outra ótica, filmes que são lançados durante este período do ano têm muito mais chances de serem indicados a uma das categorias selecionadas do Oscar. É possível notar que 30% dos filmes com estréia em novembro ou dezembro acabam tendo alguma indicação, enquanto somente 3% dos filmes lançados de janeiro a abril chegam a ter seu nome na lista de indicados.

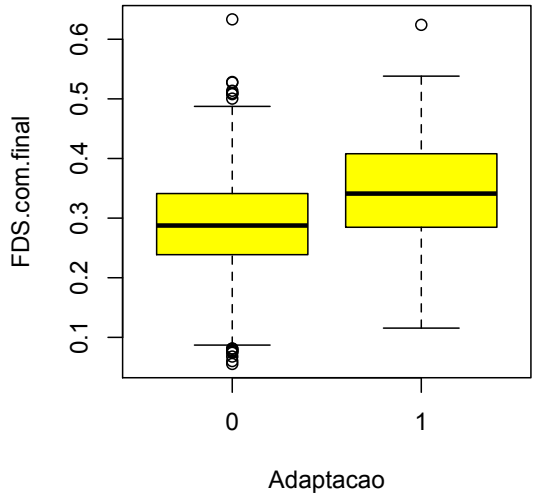
### 12) Adaptação / Continuação ("Adaptacao"):

A Tabela 16 apresenta a distribuição de frequência para a variável Adaptação/Continuação. A Figura 22 apresenta a representação gráfica da distribuição dos valores dessa variável utilizando o gráfico da caixa (boxplot).

**Tabela 16- Frequência de Adaptação / Continuação**

| Adaptação | Frequência | Frequência Relativa |
|---|---|---|
| 0 (Negativo) | 446 | 63.7% |
| 1 (Positivo) | 254 | 36.3% |

**Figura 22- Boxplot de Adaptacao**



No nosso banco de dados, temos mais filmes originais do que filmes que são algum tipo de adaptação ou continuação. Mesmo assim, vale ressaltar que 36.3% não é um número baixo, visto que estamos analisando somente os 70 filmes com maiores bilheterias dos últimos anos. Isto é, mais de um terço dos filmes analisados são fruto de adaptações.

Pelo boxplot de FDS.com.final segundo Adaptacao, podemos concluir que filmes originais tendem a ter uma arrecadação mais distribuída, e menos concentrada no final de semana de estréia. Isto faz sentido, uma vez que filmes que são alguma adaptação ou continuação normalmente são mais aguardados, e tendem a fatigarem sua demanda mais rápido.

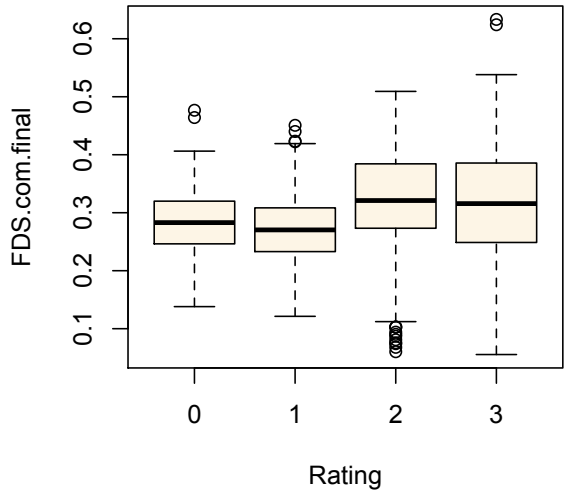
### 13) Classificação Etária ("Rating"):

A Tabela 17 apresenta a distribuição de frequência para a variável Classificação. A Figura 23 apresenta a representação gráfica da distribuição dos valores dessa variável utilizando o gráfico da caixa (boxplot).

**Tabela 17- Frequência de Rating**

| Rating | Frequência | Frequência Relativa |
|---|---|---|
| 0 (G) | 26 | 3.7% |
| 1 (PG) | 161 | 23.0% |
| 2 (PG-13) | 328 | 46.9% |
| 3 (R) | 185 | 26.4% |

**Figura 23- Boxplot de Rating**



Pela Tabela 17, podemos ver que a classificação livre (categoria de referência 0) é, de longe, a menos representativa nas 70 maiores bilheterias de cada um dos últimos 10 anos. Já a classificação PG-13 representa quase metade dos filmes.

O boxplot FDS.com.final segundo Rating nos mostra que a classificação categórica 1 (PG) tem a menor mediana, e o intervalo interquartílico da classificação categórica 3 (R) é o maior entre todas classificações. Portanto, PG é a categoria com maior longevidade de arrecadação e R é a categoria com maior variabilidade dos dados. Entretanto, devido à semelhança entre as caixas das categorias 0 (G) e 1 (PG), e das categorias 2 (PG-13) e 3 (R), e para facilitar a interpretação do nosso modelo de regressão, vamos agrupar estas classificações etárias em somente duas categorias dummy. Desta forma, para nosso modelo de regressão, que será apresentado no próxima sub-seção, as novas categorias de Classificação Etária serão:

Categoria de Referência → Indicadora das classificações etárias G e PG

Dummy 1 → Indicadora das classificações etárias PG-13 e R.

**14) Avaliação dos críticos ("RottenTomatoes"):**

A Tabela 18 apresenta as estatísticas descritivas para a variável RottenTomatoes. As Figuras 24 e 25 apresentam a representação gráfica da distribuição dos valores dessa variável utilizando os gráficos histograma e da caixa (boxplot), respectivamente.

**Tabela 18 - Estatísticas Descritivas de RottenTomatoes**

| Mínimo | 1º Quartil | Mediana | Média | 3º Quartil | Máximo | Variância | Desvio Padrão | Coef. de Variação | Assimetria | Curtose |
|---|---|---|---|---|---|---|---|---|---|---|
| 2.00 | 31.75 | 57.00 | 54.93 | 77.00 | 99.00 | 692.78 | 26.32 | 47.92% | -0.16 | -1.18 |

**Figura 24- Histograma de RottenTomatoes**

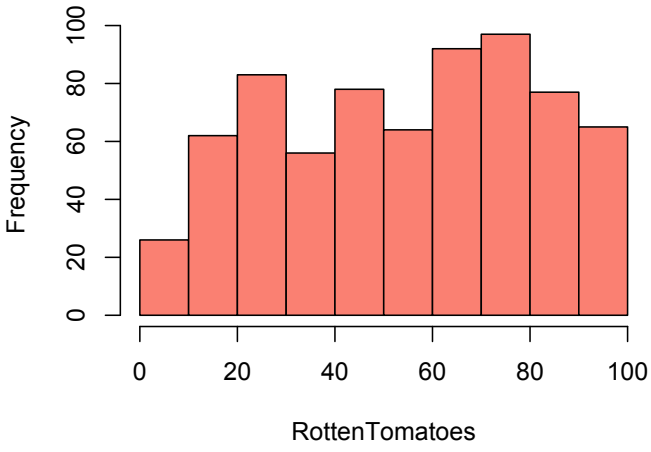
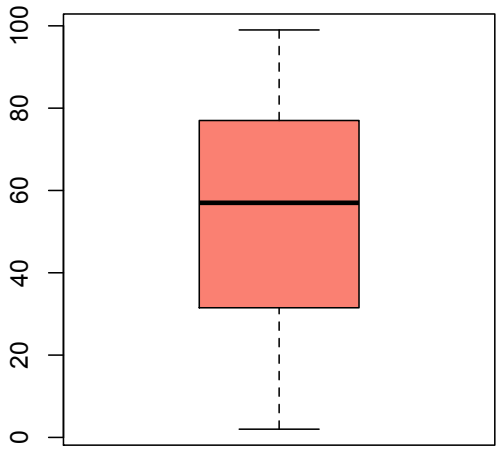


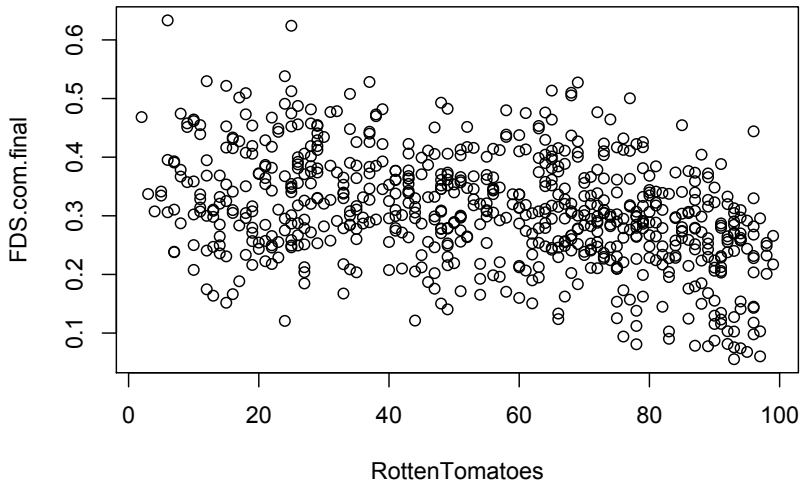

**Figura 25- Boxplot de RottenTomatoes**




As estatísticas descritivas da variável RottenTomatoes, que indica a nota média dos críticos norteamericanos, nos mostra que a pior nota que um filme do nosso banco de dados alcançou foi um 2, enquanto a maior nota foi um 99. A nota mediana é maior que a nota média de todos os dados, e, com um desvio padrão de 26.32, chegamos a um coeficiente de variação de quase 50%. A distribuição é levemente assimétrica a esquerda e negativa, e também é mesocúrtica com um coeficiente de curtose de -1.18. O histograma de RottenTomatoes corrobora com essas informações, e o boxplot nos mostra uma distribuição sem nenhum dado atípico. Entretanto, nota-se que o histograma parece apresentar uma distribuição bimodal, o que neste caso, tornaria o boxplot inapropriado para interpretação.

O coeficiente de correlação entre FDS.com.final e RottenTomatoes é negativo, quase moderado, e equivale a -0.372. A Figura 26 mostra o gráfico de dispersão entre as duas variáveis, onde pode-se ver a relação negativa entre elas.

**Figura 26- Relacao entre RottenTomatoes e FDS.com.final**



### 15) Avaliação do público ("CinemaScore"):

A Tabela 19 apresenta as estatísticas descritivas para a variável CinemaScore. As Figuras 27 e 28 apresentam a representação gráfica da distribuição dos valores dessa variável utilizando os gráficos histograma e da caixa (boxplot), respectivamente.

**Tabela 19 - Estatísticas Descritivas de CinemaScore**

| Mínimo | 1º Quartil | Mediana | Média | 3º Quartil | Máximo | Variância | Desvio Padrão | Coef. de Variação | Assimetria | Curtose |
|---|---|---|---|---|---|---|---|---|---|---|
| 56.00 | 86.00 | 89.00 | 89.94 | 93.00 | 99.00 | 30.29 | 5.50 | 6.12% | -1.08 | 2.34 |

**Figura 27- Histograma de CinemaScore**

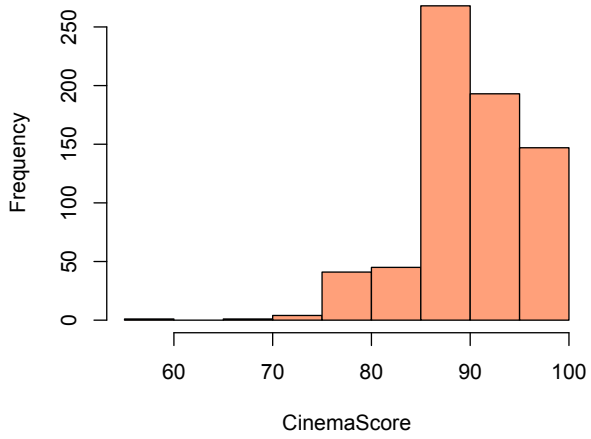
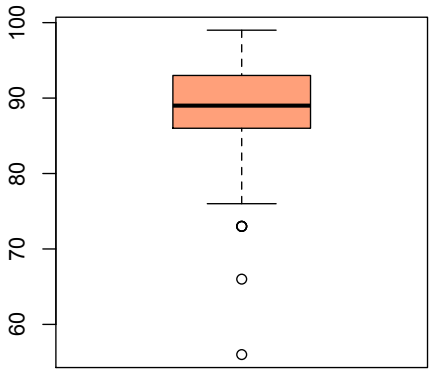



**Figura 28- Boxplot de CinemaScore**

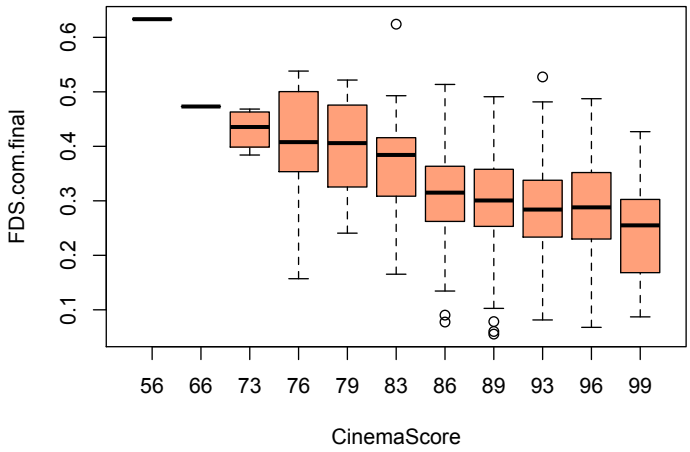



A variável CinemaScore, que indica a nota média do público norteamericano que assistiu a estréia de certo filme, varia de 56 (nota F) a 99 (nota A+). A tabela de estatísticas descritivas nos mostra que a mediana e a média são muito próximas, sendo a mediana levemente inferior. O coeficiente de variação é baixo, e equivale a 6.12%. A distribuição tem assimetria negativa e é leptocúrtica. O histograma e o boxplot acima nos dizem que o público tende a ser generoso com suas notas, já que todas as notas abaixo de 76 são considerados dados atípicos e todos os quartis estão situados na parte superior do boxplot.

O coeficiente de correlação entre FDS.com.final e CinemaScore é negativo, quase moderado, e equivale a -0.371. Curiosamente, note que a correlação de CinemaScore e RottenTomatoes com nossa variável dependente são praticamente as mesmas. A Figura 29 evidencia melhor a relação entre CinemaScore e FDS.com.final.

Em geral, as medianas decrescem à medida que aumenta a nota do público. Isto quer dizer que filmes com maiores notas de CinemaScore tendem a ter uma arrecadação mais distribuída no tempo, e menos concentrada no final de semana de estréia. Podemos então chegar à conclusão que a opinião do público sobre a qualidade de um filme afeta na longevidade de sua arrecadação.

**Figura 29- Relacao entre CinemaScore e FDS.com.final**




## 4.2 MODELOS DE REGRESSÃO:

Nesta seção, iremos analisar e discutir os resultados do modelo de regressão escolhido neste trabalho. De maneira objetiva, a modelagem da regressão será apresentada, seguindo o passo a passo necessário para termos um modelo com um bom ajuste. Os pressupostos básicos serão testados, afim de que nosso modelo não apresente multicolinearidade, que as variáveis estejam corretamente especificadas, e que os erros posam ser considerados normais e homocedásticos. Com a exceção da variável AnoEstreia, e das variáveis BilheteriaFinal, CrescSalas e Premiacao, que foram utilizadas na última seção para estudarmos sua relação com FDS.com.final mas não estão disponíveis no momento da estreia do filme, todas outras 10 variáveis foram testadas como candidatas a explicativas no modelo de regressão. O nosso modelo visa prever a longevidade de arrecadação de um filme, e com a informação sobre sua bilheteria de estréia (também considerada como uma variável explanatória), podemos calcular qual a bilheteria final esperada de certo filme.

Uma condição importante para efetuarmos uma análise de regressão é que a variável dependente seja do tipo continua. Esta condição já é atendida, visto que FDS.com.final é uma variável continua.

A Tabela 20 apresenta os coeficientes de correlação entre a variável resposta e as variáveis quantitativas candidatas a explicativas.

**Tabela 20 - Coeficientes de Correlação das Variáveis quantitativas candidatas a explicativas com a porcentagem da bilheteria de estreia na bilheteria final (Variável Resposta)**

| Variável | Coeficiente de Correlação |
|---|---|
| SalasEstreia | 0.252 |
| MesEstreia * | -0.310 |
| FDS.Estreia | 0.273 |
| Elenco | -0.222 |
| Críticos - RottenTomatoes | -0.372 |
| Público - CinemaScore | -0.371 |

*Correlação entre MesEstreia via os pesos mensais porcentuais de referência mencionados na seção anterior e FDS.com.final (a ótica dos meses de estréia via temporadas se provou insignificante e inferior à ótica dos pesos mensais, e por isso, não foi incluída na análise de regressão).

Diante do exposto, podemos agora seguir para os resultados.

### 4.2.1 *Análise e Resultado do Modelo:*

**1º passo (MODELO A):**

Na modelagem de regressão apresentada a seguir, foi adotada a abordagem de saída uma-a-uma. O primeiro passo a fazer é ajustar um modelo completo, com todas as variáveis explicativas, e retirando aquelas estatisticamente insignificantes. Ao optar por esta abordagem, a presença de multicolinearidade deve ser avaliada concomitamente, pois ela pode afetar os testes T das variáveis explicativas. Com alta multicolinearidade, os valores das estatísticas de teste T ficam próximos de zero devido às variâncias infladas, levando a falsas conclusões. A forma utilizada para detectar multicolinearidade é calcular os fatores de inflação da variância (VIF) das variáveis: se alguma variável apresentar um VIF maior que 5, esta variável pode

apresentar problemas de multicolinearidade. A Tabela 21 mostra os valores de VIFs para cada uma das 11 variáveis explicativas.

**Tabela 21 – Fator de Inflação da Variância**

| Variável | GVIF | Graus de liberdade | GVIF^(1/(2*Df)) |
|---|---|---|---|
| Mês de Estreia | 1.185 | 1 | 1.086 |
| Gênero | 6.633 | 5 | 1.231 |
| Bilheteria do Fim de semana de Estreia | 2.266 | 1 | 1.474 |
| Studio | 1.916 | 6 | 1.068 |
| Nota do Elenco | 1.723 | 4 | 1.061 |
| Adaptacao | 1.457 | 1 | 1.186 |
| Rating | 3.393 | 3 | 1.494 |
| Nota dos Críticos (RottenTomatoes) | 1.575 | 1 | 1.181 |
| Nota do Público (CinemaScore) | 1.772 | 1 | 1.280 |
| Número de Salas de Estreia | 2.659 | 1 | 1.582 |

Como mostra a Tabela 21, nenhuma variável apresentou valor alto de VIF quando ajustado pelos graus de liberdade, logo, podemos considerar que não há presença de multicolineariedade severa. Podemos então seguir adiante testando um modelo com todas 10 variáveis explicativas. Chamaremos este modelo de "Modelo A.0". A Tabela 22 mostra a Análise de Variância do Modelo A.

**Tabela 22 - ANOVA Modelo A**

| Variável | Graus de liberdade | Soma dos Quadrados | Médias dos Quadrados | Teste F | Pr(>F) |
|---|---|---|---|---|---|
| MesEstreia | 1 | 0.5804 | 0.5804 | 143.4733 | < 0.001 |
| Genero | 4 | 1.1682 | 0.2921 | 57.3733 | < 0.001 |
| FDS.Estreia | 1 | 0.4210 | 0.4209 | 104.1270 | < 0.001 |
| Studio | 2 | 0.0611 | 0.0305 | 3.6540 | 0.0005 |
| Elenco | 4 | 0.1352 | 0.0338 | 8.8684 | < 0.001 |
| Adaptacao | 1 | 0.1037 | 0.1037 | 23.3972 | < 0.001 |
| Rating | 1 | 0.0216 | 0.0216 | 2.3263 | 0.0212 |
| RottenTomatoes | 1 | 0.5434 | 0.5334 | 131.46381 | < 0.001 |
| CinemaScore | 1 | 0.2336 | 0.2336 | 55.3838 | < 0.001 |
| SalasEstreia | 1 | 0.0485 | 0.0485 | 12.1377 | 0.0005 |
| Residuals | 682 | 2.7696 | 0.0041 | | |

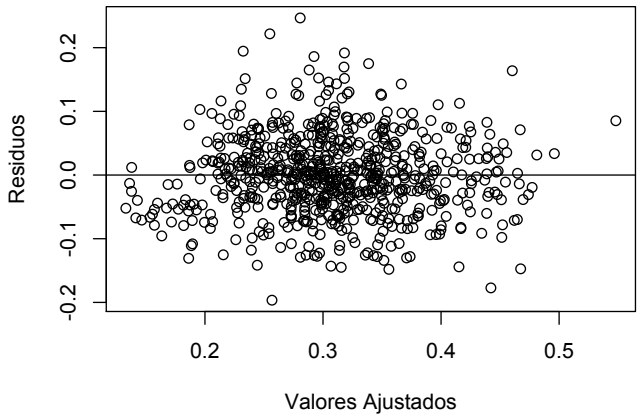
Erro Padrão dos Resíduos: 0.06373 com 682 de graus de liberdade

$R^2$: 0.5453 *$R^2$ ajutado: 0.5340*

A tabela ANOVA acima nos mostra que todas as variáveis possuem p-valores para o teste F sequencial abaixo do nível de significância de 5%, e, portanto, devem permanecer no modelo. O coeficiente de determinação ajustado do Modelo A é moderado e equivale a 53.4%. Isto quer dizer que, neste modelo, 53.4% da variância da nossa variável resposta FDS.com.final é explicada pelas variáveis explicativas.

Em seguinda, queremos checar se o pressuposto da homocedasticidade dos erros é atendido pelo Modelo A. A Figura 30 nos mostra a relação entre os valores ajustados do modelo e os resíduos.

**Figura 30- Residuos vs Valores Ajustados**

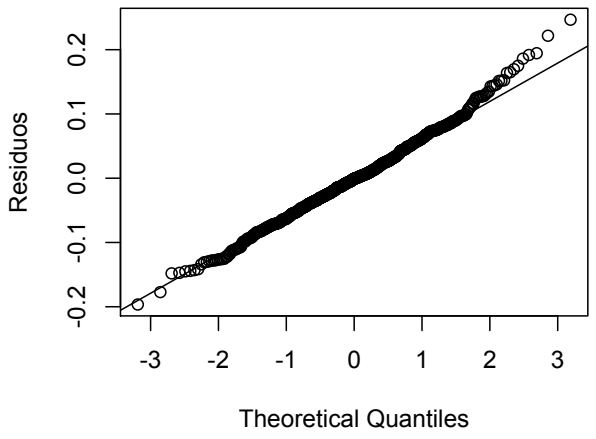



A Figura 30 nos mostra que os resíduos não apresentam nenhuma tendência. Os pontos estão razoalmente bem distribuídos pelo gráfico, e, portanto, podemos dizer que os erros são homocedásticos, isto é, possuem variância constante e este pressuposto é atendido.

Outro pressuposto a respeito dos erros é que eles possuem uma distribuição normal. Para isto, examinamos um gráfico de normalidade (Figura 31) e efetuamos o teste de normalidade Shapiro-Wilk.

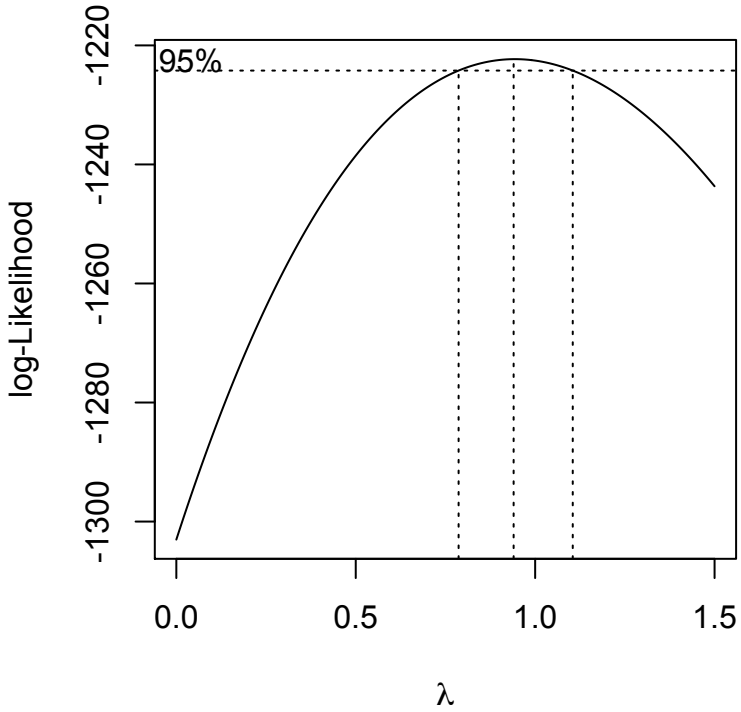
**Figura 31- Grafico de Normalidade Q-Q**

Shapiro-Wilk normality test:
W = 0.99415, p-value = 0.008289



A hipótese nula do teste de normalidade Shapiro-Wilk é de que os erros são normais. Como o p-valor do nosso teste foi de 0.008289, podemos rejeitar a hipótese nula ao nível de significância de 5%. Assim, não podemos afirmar que os erros seguem uma distribuição normal e esse pressuposto não foi atendido. A Figura 31 corrobora com essa informação, uma vez que os pontos não estão bem distribuídos em torno da reta diagonal da figura. Diante disso, o ajuste do modelo de regressão linear na sua forma original é inadequado. A fim de solucionar este problema da não-normalidade dos erros, vamos tentar realizar uma transformação na variável resposta. O procedimento utilizado foi a transformação de Box Cox, cujo resultado é apresentado graficamente na Figura 33.

**Figura 33 – Transformação de Box Cox**




O gráfico da transformação de Box Cox, indicado pela Figura 33, nos diz que uma boa transformação para nossa variável resposta seria:

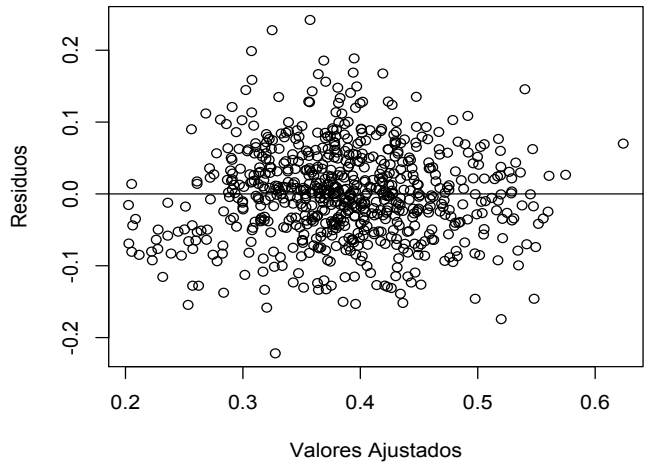
**FDS.com.finalTrans = (FDS.com.final^0.8)**

Desta forma, agora podemos tentar um novo modelo com a variável resposta "FDS.com.finalTrans", ao invés de "FDS.com.final". Após esta transformação, espera-se que os erros passam a ser normais.

**2º passo (MODELO B):**

Com nossa variável resposta transformada e nove das variáveis explicativas, foi ajustado um novo modelo de regressão. As Figuras 34 e 35 mostram a análise de resíduos desse novo modelo.

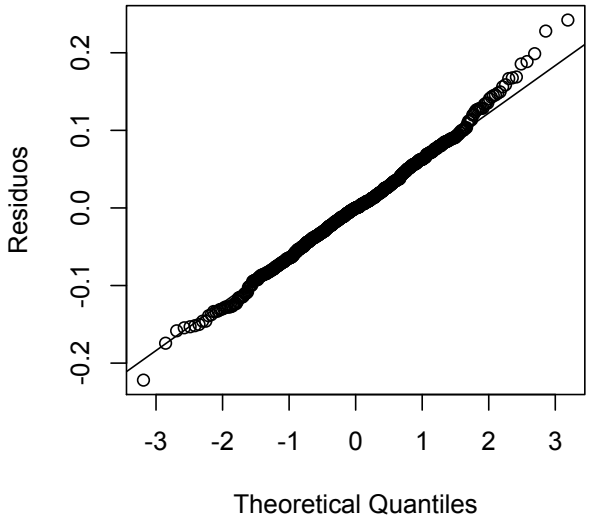
**Figura 34- Residuos vs Valores Ajustados**




A Figura 34 nos diz que os erros atendem o pressuposto de homocedasticidade, já que os pontos aparecem bem distribuídos no gráfico.

**Figura 35- Grafico de Normalidade Q-Q**

Shapiro-Wilk normality test
W = 0.9961, p-value = 0.08033

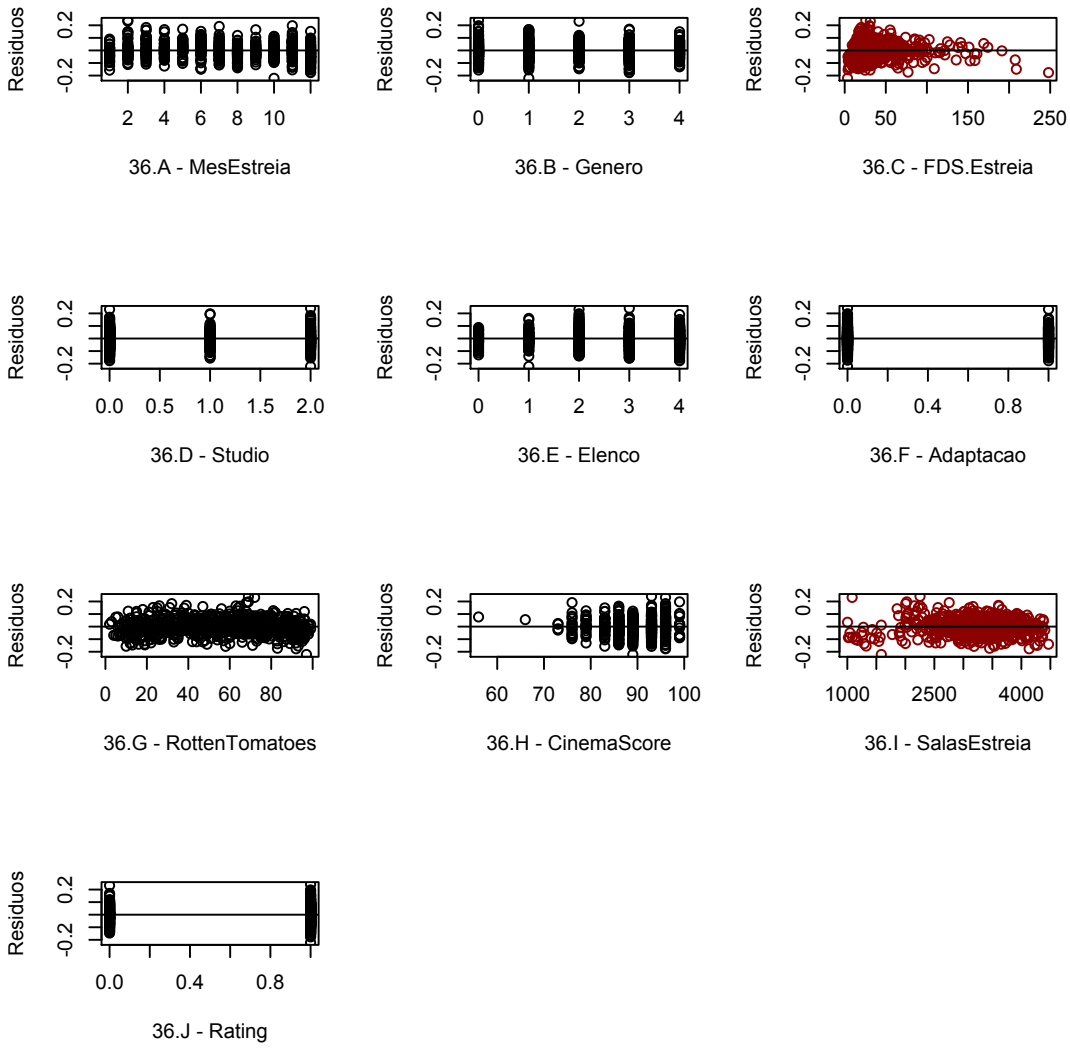



A Figura 35 e o p-valor do teste de normalidade Shapiro-Wilk nos diz que a distribuição dos erros atendem o pressuposto da normalidade. A hipótese nula do teste

de Shapiro-Wilk, que diz que os erros seguem uma distribuição normal, é aceita a um nível de significância de 5%.

Com os pressupostos da homocedasticidade e da normalidade dos erros atendidos, precisamos checar se não há erro de especificação nas variáveis explicativas do modelo. Para tal, precisamos analisar os gráficos dos resíduos com cada uma das nove variáveis explicativas. Assim como no gráfico de resíduos versus valores ajustados, queremos que nossos gráficos apresentem erros bem espalhados. Caso o gráfico apresente alguma tendência, também será necessário fazer transformações nas variáveis explicativas. A Figura 36 mostra todas as relações entre os resíduos e as variáveis explicativas.

**Figura 36 - Gráficos de Resíduos vs Variáveis Explicativas**




As variáveis Mês de Estréia, Gênero, Studio, Elenco, Adaptação, RottenTomatoes, Rating e CinemaScore apresentam resíduos bem espalhados, e,

portanto, estão corretamente especificadas. Entretanto, não podemos dizer o mesmo de Bilheteria no final de semana de estreia (FDS.Estreia) e Salas de Estréia. Estas duas variáveis precisarão de transformação para que fiquem na forma correta e o modelo fique adequado. Sem estas transformações, nosso modelo ainda não está completo. Após varias tentativas, foram efetuadas as seguintas transformações:

**SalasEstreiaTrans = (SalasEstreia^3)**

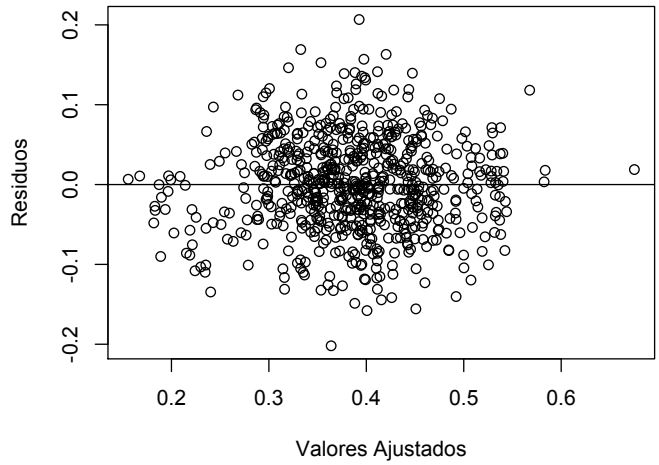
**FDS.EstreiaTrans = log(FDS.Estreia)**

Com essas duas transformações e com a transformação já efetuada sobre a variável resposta, podemos testar um novo modelo. Agora, espera-se que todos os pressupostos sejam atendidos, e finalmente tenhamos um modelo final.

**3º passo (MODELO C):**

As figuras 37, 38 e 39 mostram os gráficos utilizados na análise de resíduos do modelo de regressão ajustado com a variável respostas transformada e também duas das variáveis explicativas transformadas (Modelo C).

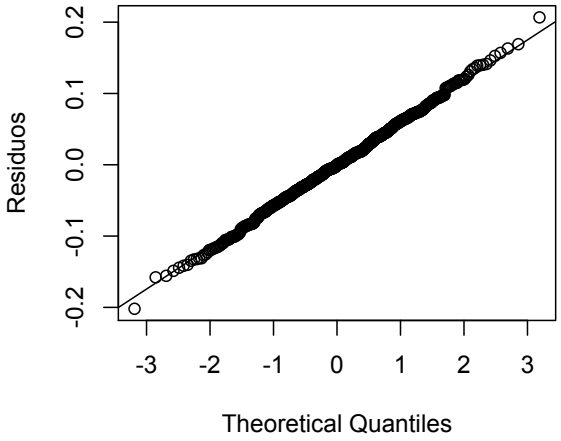
**Figura 37- Residuos vs Valores Ajustados**




O pressuposto da homocedasticidade, ou variância constante dos erros, é confirmado pelos pontos espalhados na Figura 37, que mostra a relação entre os resíduos e os valores ajustados.

**Figura 38- Grafico de Normalidade Q-Q**

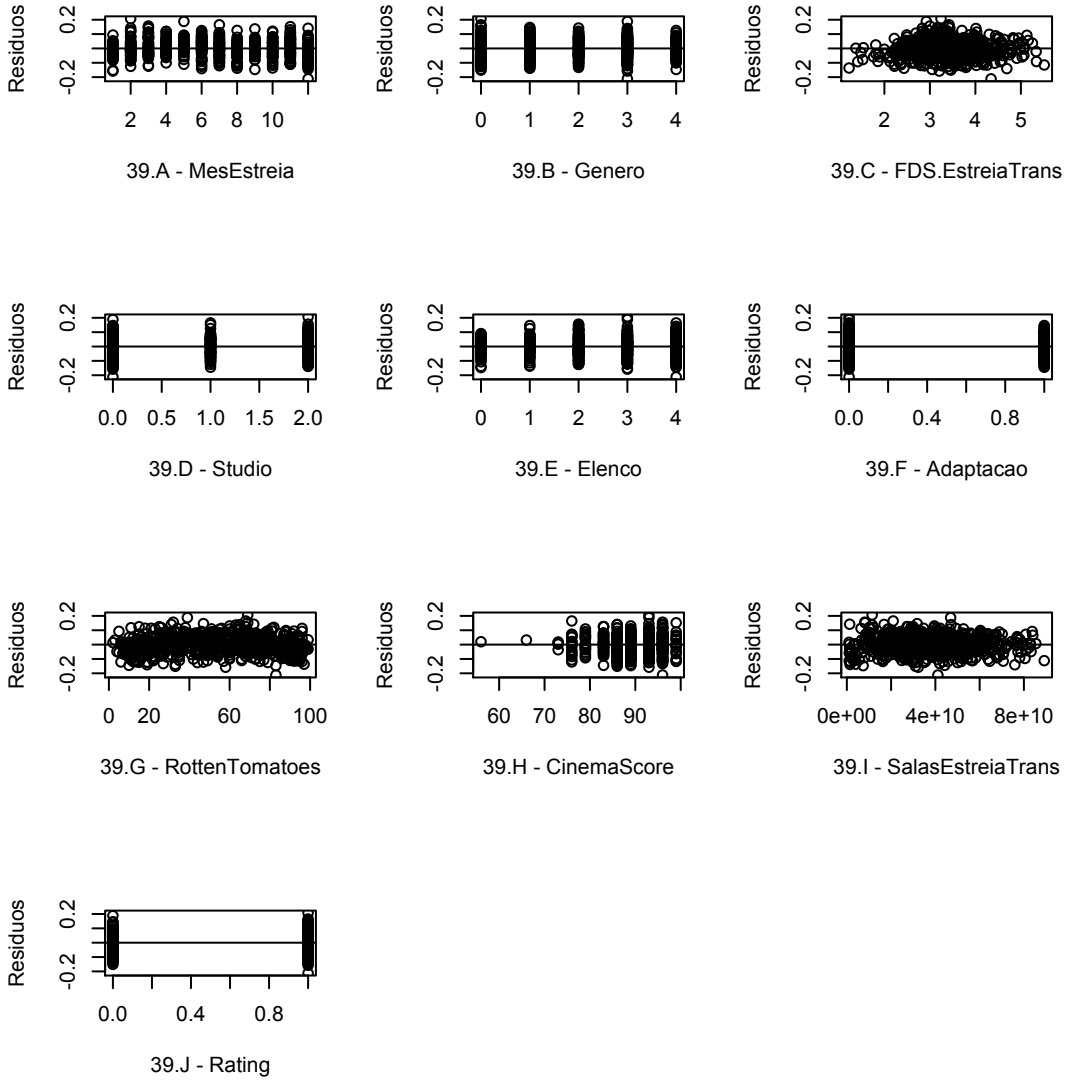
Shapiro-Wilk normality test
W = 0.99896, p-value = 0.966




O pressuposto da normalidade dos erros é confirmado pelo gráfico de normalidade, apresentado pela Figura 38, e principalmente pelo alto p-valor do teste de normalidade Shapiro-Wilk, levando à não rejeição da hipótese de normalidade dos erros.

Por fim, a Figura 39 mostra os gráficos de dispersão dos resíduos versus todas as variáveis explicativas incluídas no Modelo C.

**Figura 39 – Gráficos de Resíduos vs Variáveis Explicativas**



Diferentemente do Modelo B, agora podemos dizer que todas variáveis explicativas estão corretamente especificadas. As transformações feitas nas variáveis Bilheteria do final de semana de estreia (FDS.Estreia) e número de Salas de Estréia (SalasEstreia) foram satisfatórias, e agora, a análise de resíduos apresenta gráficos que nos permitem dizer que o ajuste do Modelo C aos dados satisfaz os pressupostos para um bom ajuste de um modelo de regressão linear.

A Tabela 25 apresenta a ANOVA para o modelo C, modelo final com as variáveis transformada:

**Tabela 23 - ANOVA Modelo C**

| Variável | Graus de liberdade | Soma dos Quadrados | Médias dos Quadrados | Teste F | Pr(>F) |
|---|---|---|---|---|---|
| MesEstreia | 1 | 0.6308 | 0.6302 | 171.3868 | < 0.001 |
| Genero | 4 | 1.2356 | 0.3088 | 84.1138 | < 0.001 |
| FDS.EstreiaTrans | 1 | 0.9588 | 0.9667 | 261.0993 | < 0.001 |
| Studio | 2 | 0.075 | 0.0375 | 10.2191 | < 0.001 |
| Elenco | 4 | 0.1892 | 0.0473 | 12.8843 | 0.0079 |
| Adapatacao | 1 | 0.026 | 0.026 | 7.0857 | < 0.001 |
| RottenTomatoes | 1 | 0.5451 | 0.5451 | 148.441 | < 0.001 |
| CinemaScore | 1 | 0.2338 | 0.2338 | 63.6733 | < 0.001 |
| Rating | 1 | 0.0097 | 0.0097 | 2.6414 | 0.0104 |
| SalasEstreiaTrans | 1 | 0.0253 | 0.0253 | 6.8959 | 0.0088 |
| Residuals | 682 | 2.493 | 0.0037 | | |

Erro Padrão dos Resíduos: 0.0606 com 682 de graus de liberdade

$R^2$: 0.6107 *$R^2$ ajutado: 0.601*

Todas as variáveis explicativas são estatisticamente significantes. O coeficiente de determinação ajustado melhorou, passando de 53.4% (modelo A) para 60.1%, isto é, houve um incremento na explicação da variância da porcentagem da bilheteria final representada pela bilheteria do final de semana de estréia (na escala transformada) pelas variáveis explicativas do modelo C.

Diante do exposto, o Modelo C é o nosso modelo de regressão final. Todas as hipóteses básicas para um modelo bem ajustado foram atendidas, e o coeficiente de determinação ajustado de 60.1% é satisfatório. Esta interpretação do coeficiente de determinação é somente possível devido à presença do intercepto no nosso modelo. A Tabela 26 nos mostra que o intercepto é estatisticamente significante e permanece

no Modelo C. Seu p-valor da estatística do teste T é muito baixo, rejeitando a hipótese nula do teste T de que seu coeficiente é igual a zero.

**Tabela 24 - Significância do Intercepto**

| | Valor Estimado | Desvio Padrão | Teste T | Pr(>\|t\|) |
|---|---|---|---|---|
| Intercepto | 0.6173 | 0.0469 | 13.149 | < 0.001 |

### 4.2.2 *Interpretação do Modelo:*

O modelo de regressão final é apresentado abaixo e os seus coeficientes são apresentados na Tabela 27.

$$FDS.com.finalTrans = Intercepto + ß_1 * MesEstreia + ß_2 * Genero + ß_3 * FDS.EstreiaTrans + ß_4 * Studio + ß_5 * Adaptacao + ß_6 * Elenco + ß_7 * RottenTomatoes + ß_8 * CinemaScore + ß_9 * SalasEstreiaTrans + ß_{10} * Rating + \varepsilon$$

**Tabela 25 – Coeficientes ßetas do Modelo C**

| Termo | Coeficientes | Termo | Coeficientes |
|---|---|---|---|
| Intercepto | 0.6173 | Rating (G e PG) | ---------- |
| MesEstreia | -0.8022 | Rating1 (PG-13 e R) | -0.0109 |
| Gênero de referência (Comédia) | ---------- | | |
| Genero1 (Drama) | -0.0094 | | |
| Genero2 (Animação/Família) | -0.0069 | Adaptacao (sim) | 0.0162 |
| Genero3 (Ficção / Fantasia, Ação/Aventura) | 0.0172 | Elenco0 | ---------- |
| Genero4 (Terror/Suspense) | 0.0345 | Elenco1 | 0.0165 |
| | | Elenco2 | 0.006 |
| FDS.EstreiaTrans | 0.0833 | Elenco3 | -0.0136 |
| Studio de referencia (Buena Vista, Sony, Paramount, Fox e Warner) | ---------- | Elenco4 | -0.0238 |
| Studio1 (Universal) | 0.0139 | RottenTomatoes | -0.00089 |
| Studio2 (Outros Estúdios) | 0.0208 | CinemaScore | -0.0042 |
| | | SalasEstreiaTrans | -6.36E-13 |

O intercepto, que equivale a 0.6173, é o valor esperado para FDS.com.finalTrans quando todas variáveis explicativas são iguais a zero ou se referem às categorias de referência. Neste caso, ele não deve ser interpretado, pois

representa uma extrapolação, uma vez que nem todas as variáveis explicativas podem assumir o valor 0, como Mês de estreia, por exemplo. Os coeficientes angulares ß's representam a variação esperada em FDS.com.finalTrans, quando a variável explicativa respectiva àquele ß aumenta uma unidade. As interpretações individuais de alguns coefientes são:

- MesEstreia: considerando todas outras variáveis constantes, para cada unidade porcentual de peso de referência mensal a mais que certo mês recebe, espera-se que FDS.com.finalTrans diminua, uma vez que o sinal do seu coeficiente é negativo. Isto é, o sinal negativo nos diz que um aumento na variável explicativa provocaria uma redução na variável resposta.

- Gênero: Drama é o gênero que apresenta menor coeficiente, e Terror/Suspense, é o que apresenta maior coeficiente, todos em relação ao gênero de referência, Comédia. Os gêneros Drama e Animação/Família apresentam sinais de coeficientes negativos, e, por isto, diminuem os valores da variável resposta em comparação com Comédia. Já os outros gêneros aumentam os valores da variável resposta.

- FDS.EstreiaTrans: considerando todas outras variáveis constantes, para cada unidade de FDS.EstreiaTrans a mais, FDS.com.finalTrans aumenta, uma vez que o sinal do seu coeficiente é positivo.

- Studio: Universal e Outros Estúdios apresentam variação positiva na variável resposta em relação aos estúdios de referência. Isto quer dizer que Universal e Outros Estúdios aumentam o valor da variável resposta, em relação a categoria de referência, que engloba todos outros principais estúdios, sendo eles Buena Vista, Sony, Paramount, Fox e Warner.

- Rating: as classificações etárias PG-13 e R apresentam variação negativa na variável resposta em relação as classificações de referência G e PG.

- Adaptação: o coeficiente de para a variável Adaptação, que indica quando um filme foi indicado ao Oscar, é positivo, indicando que esta característica aumenta o valore da variável resposta, considerando todas outras variáveis do modelo constantes.

- Elenco: enquanto as classes de elenco 1 e 2 possuem coeficientes positivos, e, portanto, aumentam a média da variável resposta em relação a categoria de referência de Elenco, as classes de elenco 3 e 4 diminuem essa média devido a seus coeficientes serem negativos.

- RottenTomatoes: considerando todas outras variáveis constantes, para cada ponto de RottenTomatoes a mais, FDS.com.finalTrans diminui, uma vez que o valor do coeficiente de RottenTomatoes tem sinal negativo.

- CinemaScore: considerando todas outras variáveis constantes, para cada ponto de CinemaScore a mais, FDS.com.finalTrans diminui, uma vez que o valor do coeficiente de CinemaScore tem sinal negativo. Note que o efeito de um ponto adicional de CinemaScore na redução do valor da variável resposta é maior que o efeito de um ponto adicional na nota de RottenTomatoes, indicando que a nota do público produz um efeito negativo maior sobre a variável resposta do modelo do que a nota dos críticos.

- SalasEstreia: considerando todas outras variáveis constantes, para cada sala de cinema a mais que o filme é exibido em sua estréia, a variável resposta tem seu valor reduzido, devido ao valor negativo do coeficiente de Salas de Estréia.

Devido ao fato de a variável resposta do nosso modelo ser a porcentagem que a bilheteria arrecadada no final de semana de estréia representa na bilheteria final, elevada à potência 0.8, para estimarmos a bilheteria final, precisamos multiplicar os valores preditos no modelo pela bilheteria arrecadada no final de semana de estréia. Mas antes disso, como nossa variável resposta sofreu uma transformação, precisamos fazer uma transformação inversa. Como fizemos "FDS.com.finalTrans = (FDS.com.final^0.8)", precisamos agora elevar os resultados por 1.25 para retirar essa transformação:

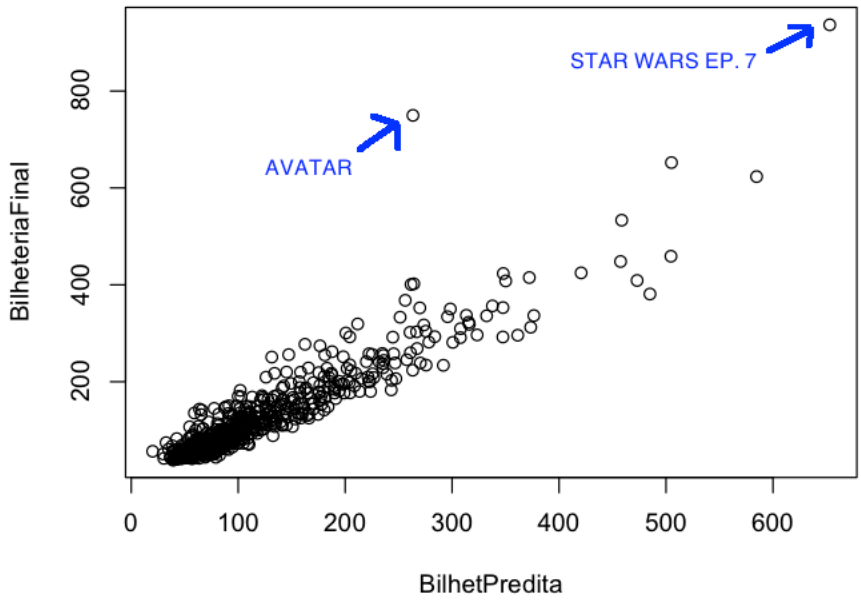
Preditos Transformados= (Preditos pelo Modelo C)^1.25

Com os valores dos preditos transformados, que equivale a valores preditos para FDS.com.final, podemos calcular qual a bilheteria final estimada pelo nosso modelo dividindo a bilheteria arrecadada no final de semana de estréia pelos próprios preditos transformado:

Bilheterial Final Predita = (Bilheteria do Final de Semana de Estréia) / Preditos Transformados

A Figura 40 nos mostra o gráfico de dispersão entre as bilheterias finais preditas pelo nosso modelo e as reais bilheterias finais dos 700 filmes considerados no nosso banco de dados.

**Figura 40 - BilheteriaFinal Predita vs BilheteriaFinal**




A Figura 40 nos mostra uma forte relação linear positiva entre os valores reais da Bilheteria Final e os valores preditos para ela pelo modelo de regressão. O coeficiente de correlação linear de Pearson entre esses valores é forte, positivo e equivale a 0.931. Entretanto, temos dois filmes que se destacam na Figura 40. São eles os filmes Avatar, e Star Wars Episodio 7. Estes são os dois filmes que mais arrecadaram bilheteria no mercado norteamericano e desafiaram qualquer previsão de arrecadação total. Quando excluídos esses dois filmes, o coeficiente de correlação entre os valores reais e os valores preditos para Bilheteria Final aumentou, passando para 0.946. Esta forte correlação é um indicador de que nosso modelo, além de estatisticamente significante, pode ser utilizado para prever satisfatoriamente a bilheteria final arrecadada por um filme.

## 4.3 *LIMITAÇÕES DE ESTUDO:*

Os fatores que não temos como controlar, ou fatores sobre os quais não temos informações completas e confiáveis, limitaram nosso estudo sobre bilheteria arrecadada no mercado norteamericano. Alguns desses fatores são:

- Inflação: a inflação dos preços médios dos ingressos de cinema não foi considerada neste estudo. Embora os Estados Unidos tenham baixos índices inflacionários, em 10 anos, o preço do ingresso de cinema aumentou. O preço médio anual dos ingressos poderia ser uma variável explicativa de um modelo de previsão de bilheteria. Sua significância seria testada para avaliar sua relevância estatística em um modelo de previsão.
- Tamanho do banco de dados: o banco de dados montado neste trabalho conta somente com 700 filmes, sendo 70 filmes para cada um dos últimos 10 anos completos. Isto pode ser visto como uma limitação do estudo, uma vez que poderíamos testar modelos com um maior banco de dados.
- Orçamento e Custos de Marketing: as informações disponíveis na internet sobre orçamento e custos de marketing dos filmes são bastante limitadas, sendo quase impossível ter dados confiáveis sobre todos os filmes incluídos no banco de dados. Essas informações provavelmente seriam de muita relevância para nosso estudo de previsão de bilheteria, mas sua falta de transparência impediu que estes dados de entrarem no nosso banco.
- Exibições totais: alguns filmes têm pouco mais de uma hora de duração, enquanto outros possuem quase três horas de duração. Este tempo de duração de um filme afeta quantas exibições uma sala de cinema pode disponibilizar ao consumidor. Filmes mais longos podem arrecadar menos simplesmente por serem exibidos em menos seções, mesmo que em muitas salas, e vice-versa para filmes muito curtos.
- Bilheteria Arrecadada Internacionalmente: o presente trabalho trabalhou somente com as bilheterias arrecadas em solo norteamericano, mas investidores também se interessam pelo total arrecadado nos outros países. Uma limitação deste estudo foi não ter acesso aos dados de bilheteria arrecadada em todos outros países. Mesmo assim, tal tipo de dado é muito

controverso, uma vez que a própria cotação das moedas estrangeiras afeta na arrecadação internacional.

- *Twitter Buzz*: nos últimos anos, a quantidade de menções de um filme no Twitter está sendo utilizada para prever a sua bilheteria e seu "boca a boca". Este número de menções no Twitter pode indicar uma longevidade maior de arrecadação, uma vez que os próprios usuários do Twitter fazem propaganda dos filmes. Por outro lado, o conteúdo das menções também pode ser relevantes, pois um filme com muitas menções negativas pode ter sua bilheteria final arrecadada prejudicada. Os dados acerca do *Twitter Buzz* podem ser relevantes em um modelo de previsão de bilheteria, mas por ser uma característica relativamente recente, somente filmes dos últimos anos teriam essa informação disponível na internet.

Existem dezenas de outras características que separam um filme dos outros, contudo, os fatores citados acima são os principais fatores que podem ajudar a prever a bilheteria final de um filme baseado um modelo de regressão que não conte com essas limitações.

# 5. CONCLUSÕES

Este trabalho preveu a porcentagem que a arrecadação do final de semana de estréia representa na arrecadação final de um filme, por meio de uma análise de regressão com nove variáveis explicativas e um banco com dados de 700 filmes.

Chegou-se a um modelo de previsão que atende todos os pressupostos básicos de uma análise de regressão, e com um coeficiente de determinação ajustado de 60.1%. Das 10 possiveis variáveis candidatas a explicativas, todas se provaram estatisticamente significantes e foram incluídas no modelo, sendo elas o mês de estréia, o gênero, a bilheteria arrecadada no final de semana de estréia, o estúdio, se o filme é uma adaptação, o elenco, a classificação etária, as notas dos críticos e do público, e o número de salas de estréia. A alta correlação entre os valores preditos pelo modelo final e os valores reais da bilheteria final arrecadada nos indica que temos um bom modelo em mãos. Comparando com o modelo apresentado por Jeffrey Simonoff em seu artigo "*Predicting Total Movie Grosses After One Week",* onde a variável resposta foi a bilheteria final arrecadada e as variáveis explicativas foram a arrecadação no final de semana de estréia e a nota dos críticos RottenTomatoes, nosso modelo apresentou um coeficiente de determinação ajustado inferior (92.04% contra 60.1%). Embora ambos modelos sejam significantes, nosso modelo conta com mais dados, informações e características a respeito de um filme.

Na análise exploratória dos dados, analisamos o comportamento individual de 15 variáveis, e vimos como cada uma se relaciona com a variável resposta do modelo de regressão, porcentagem da bilheteria final representada pela bilheteria do final de semana de estreia. Em destaque, tivemos a variável Premiacao, que embora não incluída no modelo de regressão, nos indicou ser relevante para prever a longevidade da arrecadação. Sua análise nos mostrou que filmes indicados às principais categorias do Oscar tendem a ter uma arrecadação mais bem distribuída do que filmes que não tiveram nenhuma indicação. Vimos também que o ano de estréia foi a variável com menor correlação com a variável resposta, e a nota média dada pelos críticos apresentou a maior correlação com a variável resposta.

Diante do exposto, conclui-se que é possível fazer uma estimativa de quanto um filme pode arrecadar baseado em um modelo de regressão utilizando as variáveis aqui apresentadas. Entretanto, este estudo contém algumas limitações, e recomenda-se

para um estudo futuro, modelar uma regressão com ainda mais dados e informações sobre as características de um filme, como por exemplo seu custo de produção e de marketing, o número de ingressos vendidos, entre outros. Também recomenda-se uma análise de regressão não linear dos dados; trabalhos futuros poderiam testar modelos gama de regressão, e desta forma, se possível, até incluir a variável Bilheteria Final diretamente no modelo de regressão.

# REFERÊNCIAS

GUJARATI, D.; PORTER, D. Econometria Básica. Quinta edição. São Paulo: McGraw Hill, 2008.

STATISTA – The Statistics Portal. Statistics and Facts About the Film Industry. Disponível em: < https://www.statista.com/topics/964/film/>. Acesso em: Outubro de 2016.

SIMONOFF, J. Predicting total movie grosses after one week, 2015. Disponível em: < http://people.stern.nyu.edu/jsimonof/classes/2301/pdf/movies.pdf>. Acesso em: Outubro de 2016.

BOX OFFICE MOJO. Disponível em: <www.boxofficemojo.com>. Acesso em: Agosto de 2016.

CINEMA SCORE. Disponível em: <www.cinemascore.com>. Acesso em: Agosto de 2016.

ROTTEN TOMATOES. Disponível em: <www.rottentomatoes.com>. Acesso em: Agosto de 2016.