Projeto Final A2 - Movies Dataset

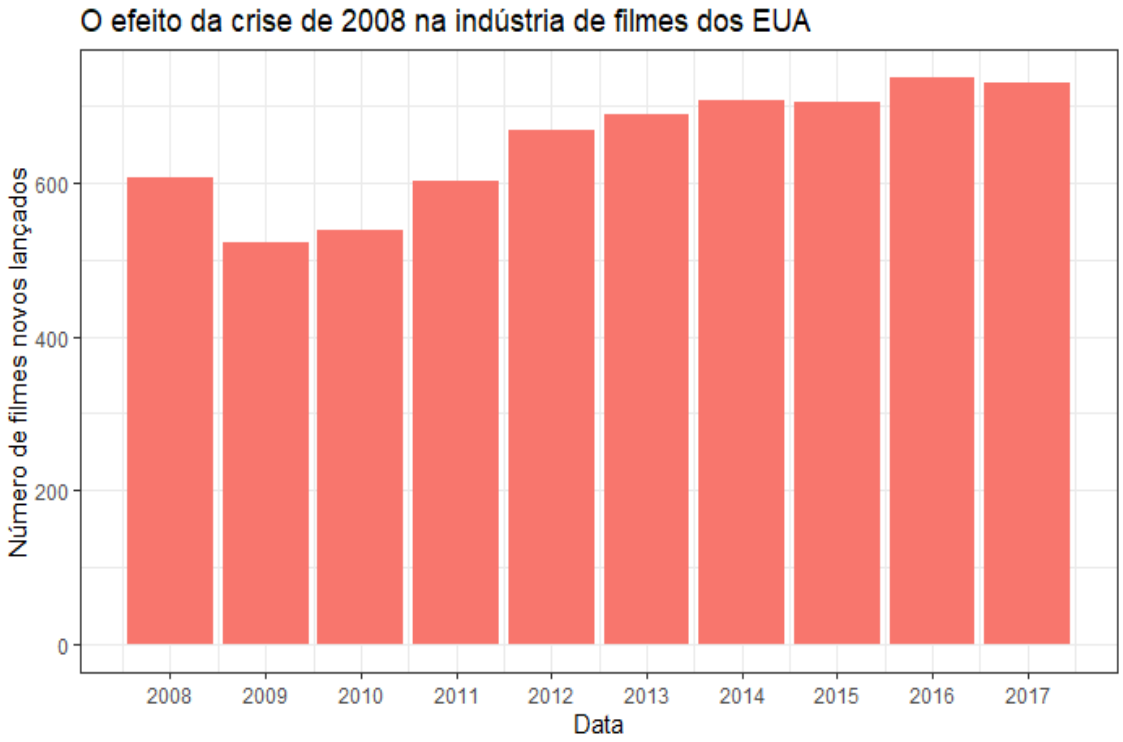
Gabriel Machado
Victor de Almeida Bombarda

Junho 2020

1 Crise de 2008

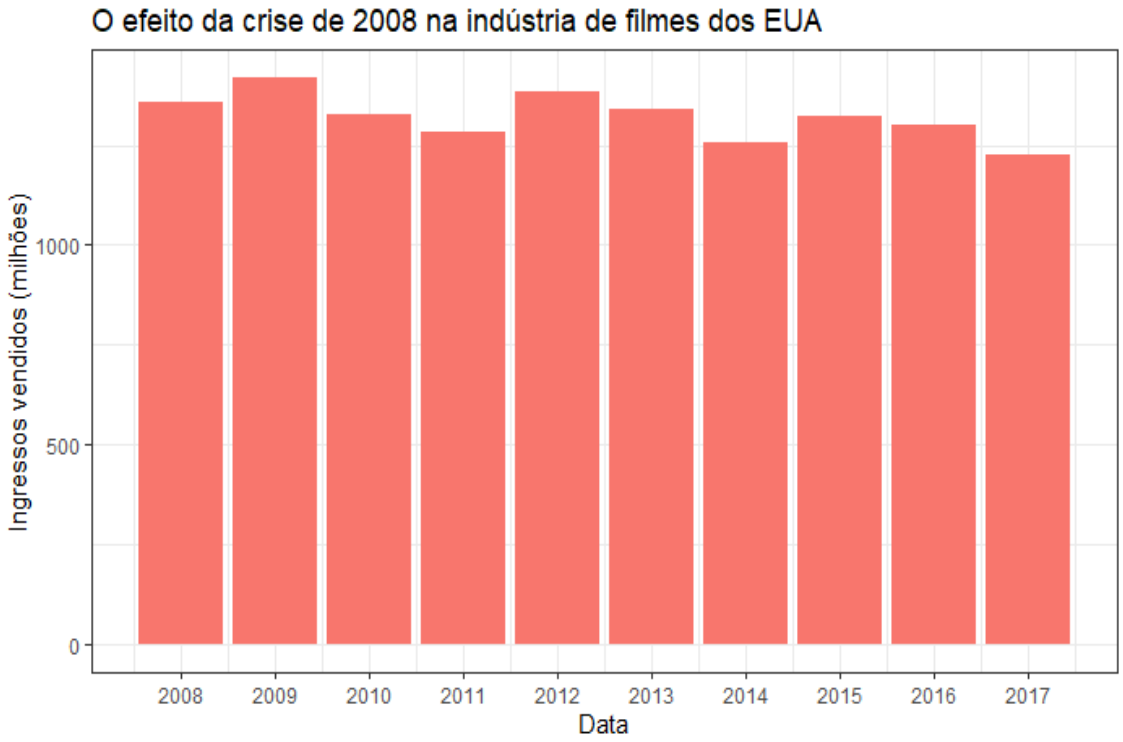
1.1 Quantidade de filmes lançados

O efeito da crise de 2008 na indústria cinematográfica pode ser observado nesse gráfico pela queda de filmes lançados no ano de 2009 e 2010, a crise afetou a produção dos filmes e muitos tiveram de ser cancelados, assim, ela impediu a conclusão de diversos filmes por diversificados fatores econômicos.
É possível observar que foram necessários 4 anos para retomar o mesmo patamar da quantidade de filmes lançados no ano de 2008.



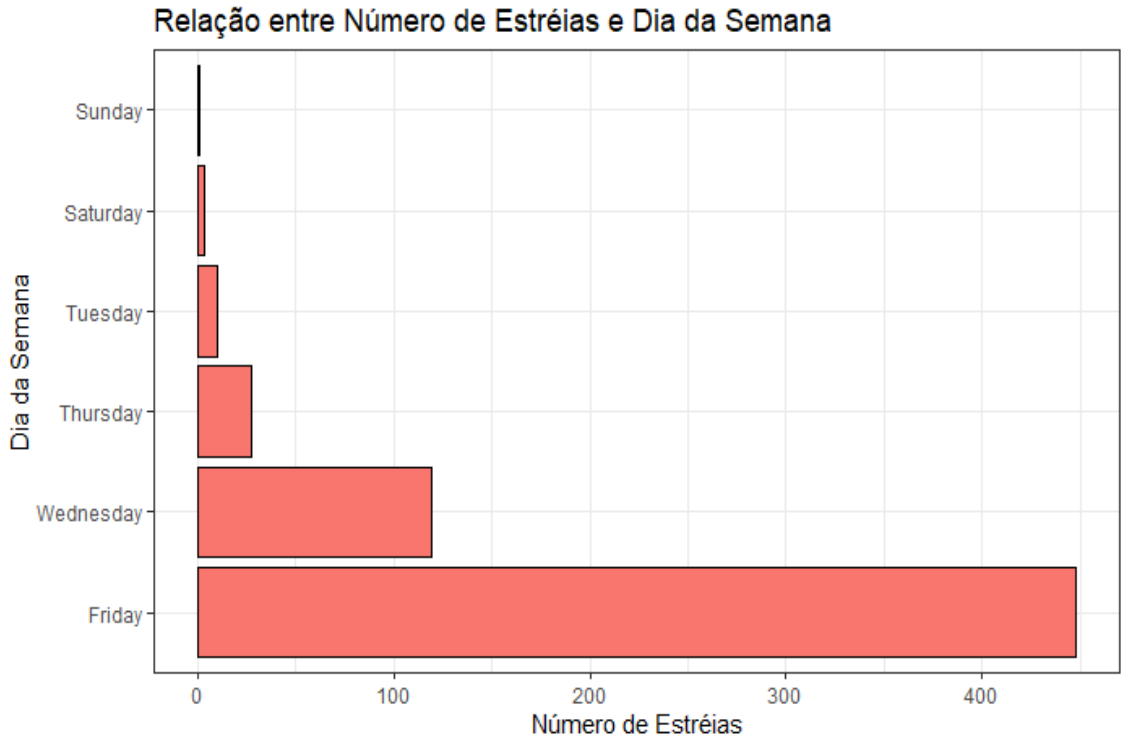
1.2 Fluxo nas salas de cinema

Esse gráfico que mostra que mesmo com o forte impacto que a crise teve na produção de filmes, o público não deixou de ir no cinema, houve até um leve crescimento na quantidade de ingressos vendidos.
Considerando que o valor médio do ingresso para o cinema teve um aumento de 7,18 dólares em 2008 para 7,50 dólares de 2009, é de fácil conclusão que esse aumento na casa de milhões de ingressos vendidos durante o decorrer dos impactos da crise, auxíliou a indústria cinematográfica a retomar suas produções e mostrou que o apreço da população pelo cinema independe dos momentos de crise.



2 Escolha dos dias do lançamento dos filmes

Com esse gráfico é possível fazer uma análise das razões dos dias escolhidos para o lançamento dos filmes. Tradicionalmente, os filmes são lançados às sextas-feiras porque é o último dia útil e o fim de semana ocorre em seguida. Logicamente, faz sentido, pois é nesse momento que a maioria das pessoas visita salas de cinema.
A idéia de lançar um filme na quarta-feira ou em qualquer outro dia é aproveitar ao máximo um feriado que ocorre nesses dias. Como os filmes dependem muito do primeiro final de semana ou dos primeiros dias do lançamento, às vezes faz sentido fazer esta troca e tentar obter o máximo de dias possível para a janela de abertura, essas são estratégias de negócio relativamente recentes e são arriscadas de se fazer, pois há as que já deram muito certo, já outras falharam, assim é possível observar que a maioria esmagadora ainda se concentra nas sextas-feiras.



3 Relação entre Tempo de Duração do filme e Crítica

Para esse gráfico é possível observar que não há uma relação direta facilmente perceptível entre o Tempo de Duração do filme em minutos e a sua nota no site de crítica IMDb, utilizamos o site IMDb para essa análise, pois ele é o mais popular e famoso, assim é o que pussui mais avaliações e é mais o usado.
Com o cálculo da correlação no RStudio é possível constatar as características observadas no parágrafo anterior descritas no valor numérico da correlação que foram colocados na legenda dos gráficos.

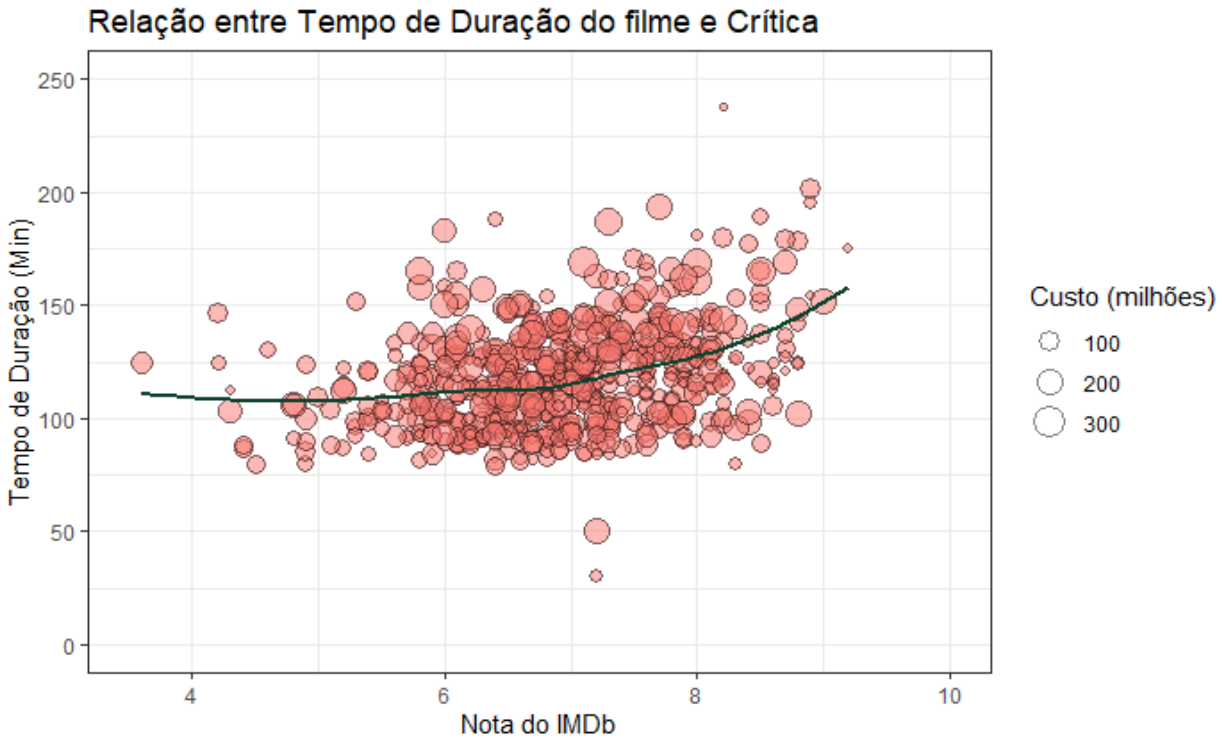


Figure 1: correlação = 0.3166371

3.1 Analisando de acordo com o gênero dos filmes

Quando realizamos a explansão do gráfico com todos os gêneros acumulados, foi de clara constatação que a correlação de certos gêneros como o Drama são os "outliers da correlação principal", apesar da quantidade de filmes de Drama ser substancialmente menor, foi possível observar essa discrepância nas correlações, assim há uma leve relação em certos gêneros com a duração do filme, pois há uma maior complexidade na produção e a história tem mais tempo para ser desenvolvida.
Há uma diferença no foco das críticas de acordo com o gênero, em filmes dramáticos é necessário uma excelente atuação, edição precisa e uma trilha sonora adequada, assim sendo um gênero complexo e difícil de se obter grandes notas pela crítica.
Agora, com os filmes de ação e animação, há muitos fatores que chegam a ofuscar os tais valores desejados nos filmes de Drama, como exemplo são os efeitos especiais, trilhas sonoras manjadas e piadas que cativam o público a sair suficientemente entretido após a sessão de cinema.
Com o cálculo da correlação no RStudio é possível constatar as características observadas no parágrafo anterior descritas no valor numérico da correlação que foram colocados na legenda dos gráficos.

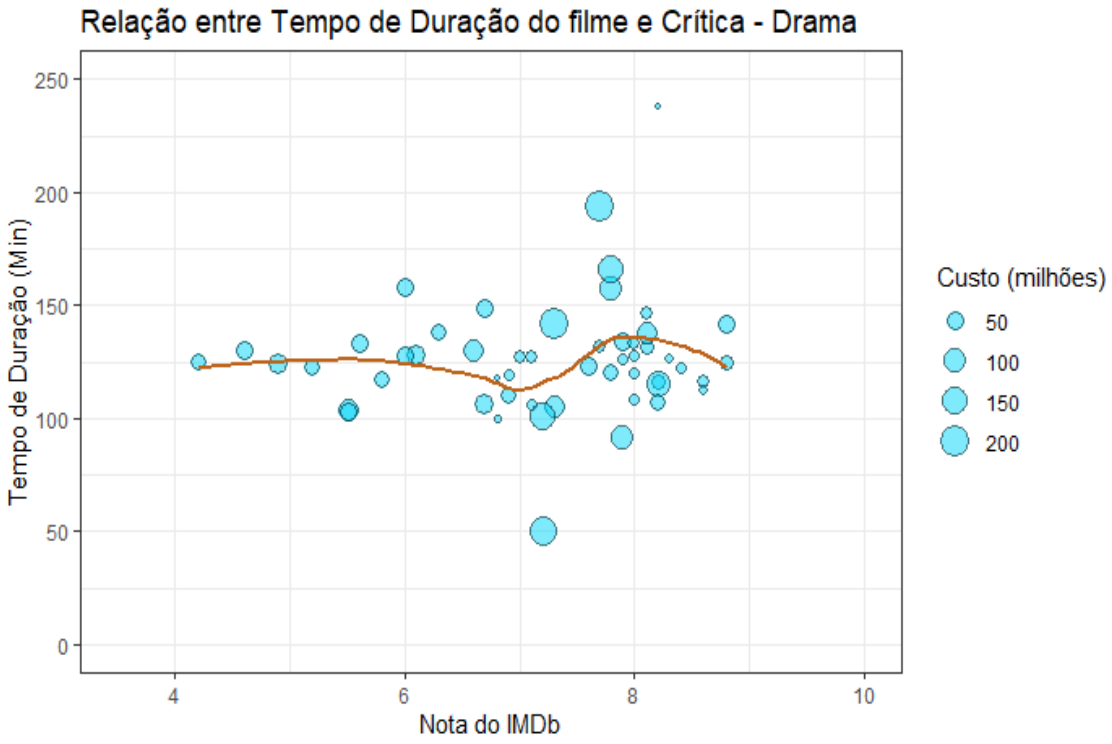


Figure 2: correlação = 0.07139958

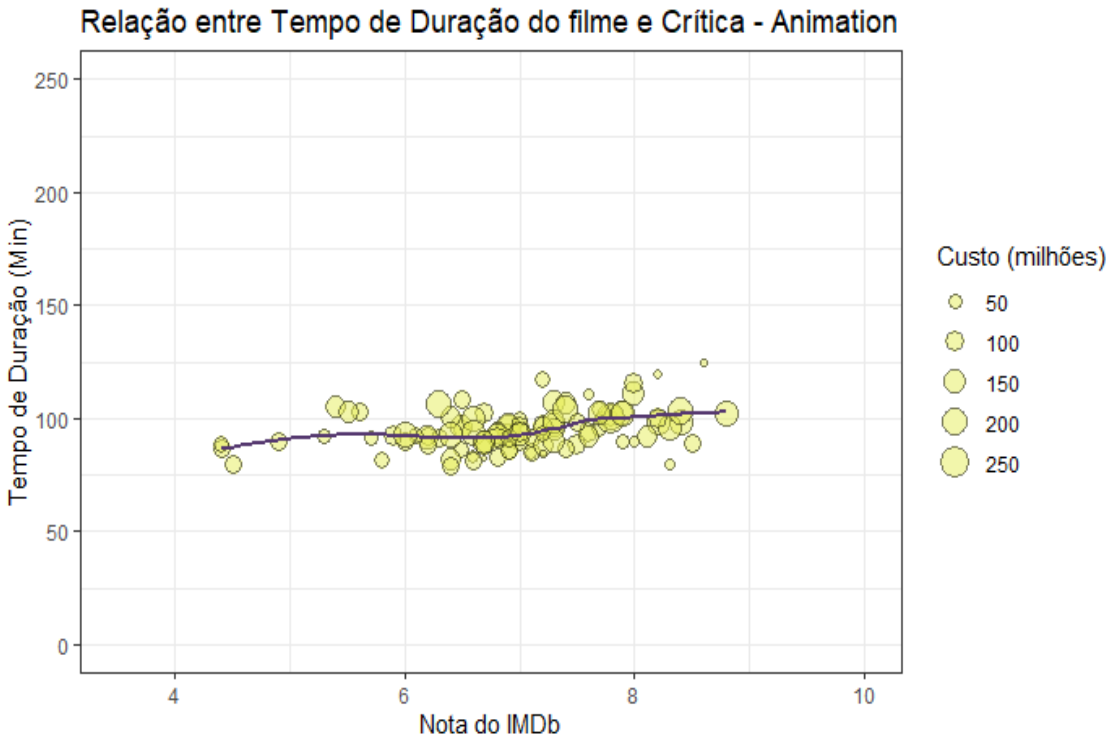


Figure 3: correlação = 0.3704476

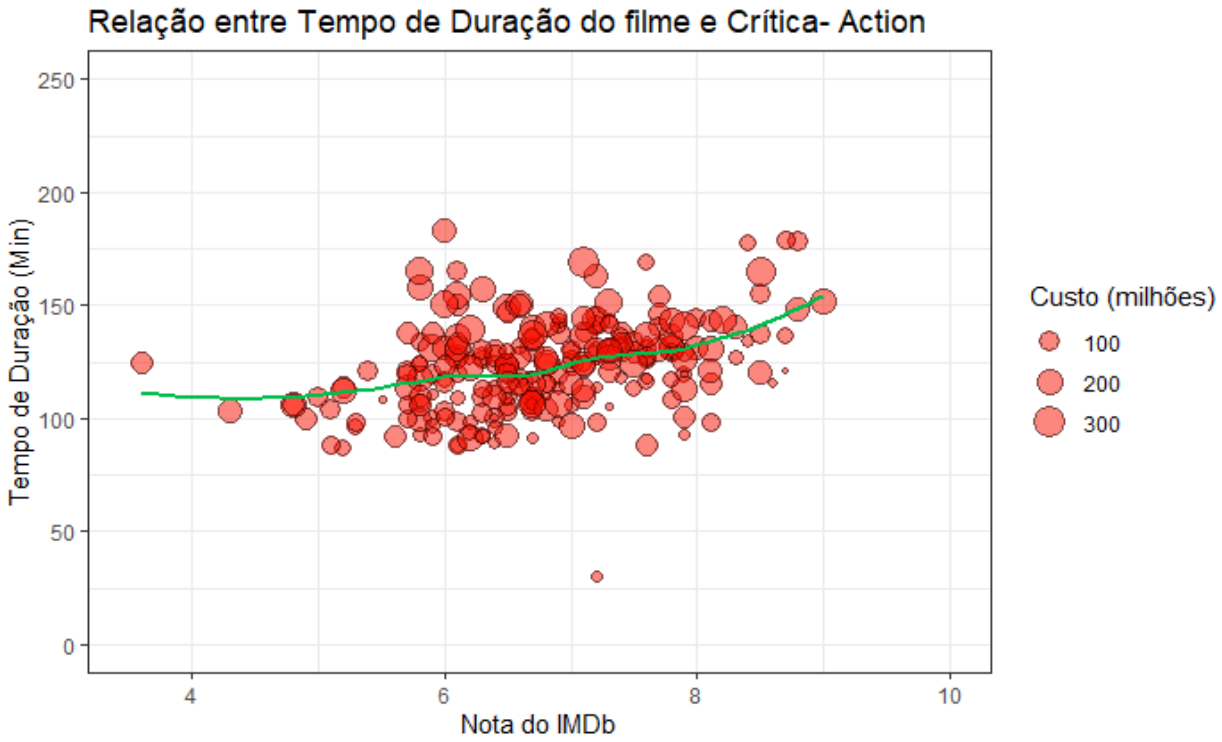


Figure 4: correlação = 0.3861769

4 Relação entre os dois sites de crítica

Com a análise desse gráfico é possível verificar essa linha de tendência quase y = x/2, o que representa uma forte relação entre os dois sites de crítica, há alguns outliers que distorcem a linha de tendência, porém ainda há uma forte relação. Com o cálculo da correlação cálculada no RStudio é possível constatar as características observadas no paragráfo anterior descrita

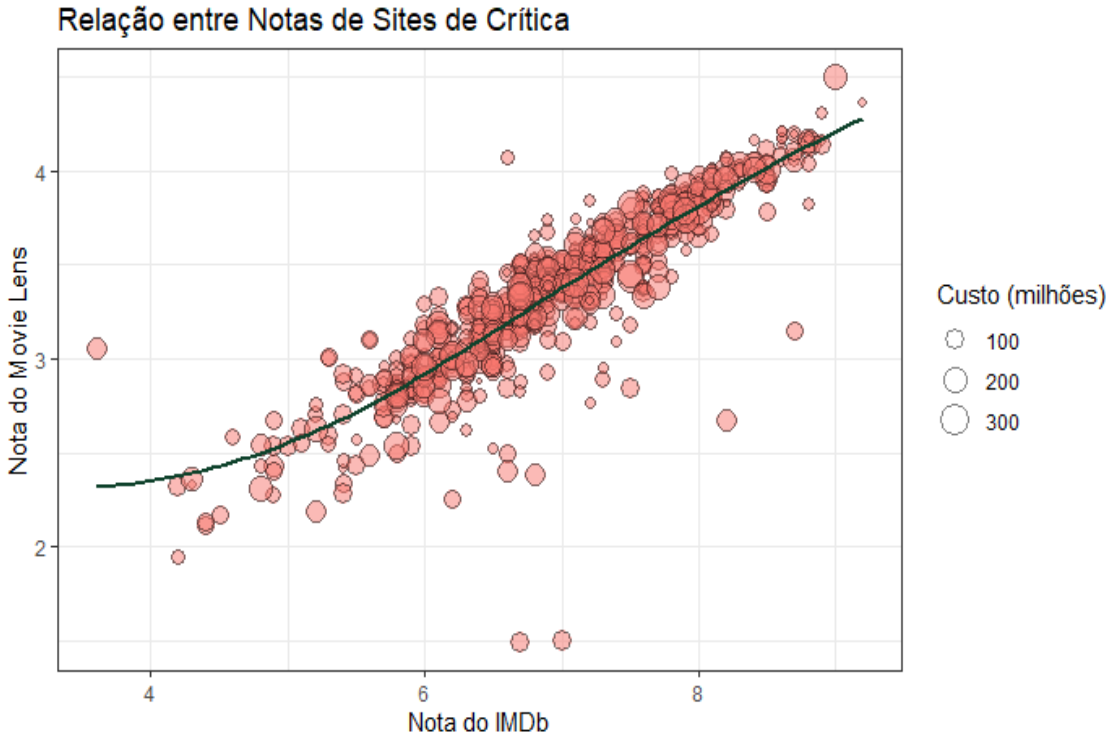


Figure 5: correlação = 0.8798356

5 Relação entre gêneros e a crítica

Com a análise desses boxplots é possível observar que não há relação entre os gêneros dos filmes e a sua crítica, apesar dos valores das críticas serem diferentes para cada gênero, há uma média muito parecida entre todos os boxplots, a distância interquartil mostra as altas variações nos filmes de Ação e Animação. Além disso é possível observar as características dos filmes de Drama já citadas anteriormente, como a tendência de se ter mais outliers negativos do que positivos. Com o cálculo do R quadrado foi possível constatar que realmente não há uma ligação tendenciosa etre o gênero dos filmes e a crítica de ambos os sites utilizados.

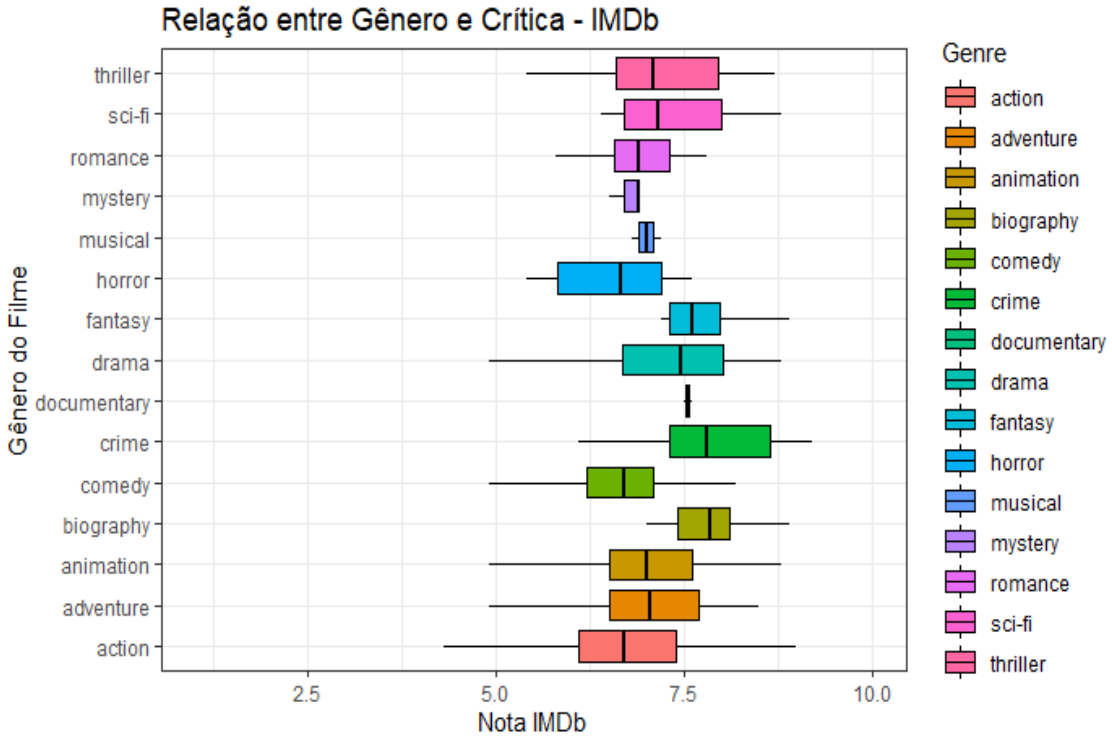


Figure 6: R quadrado = 0.09306632

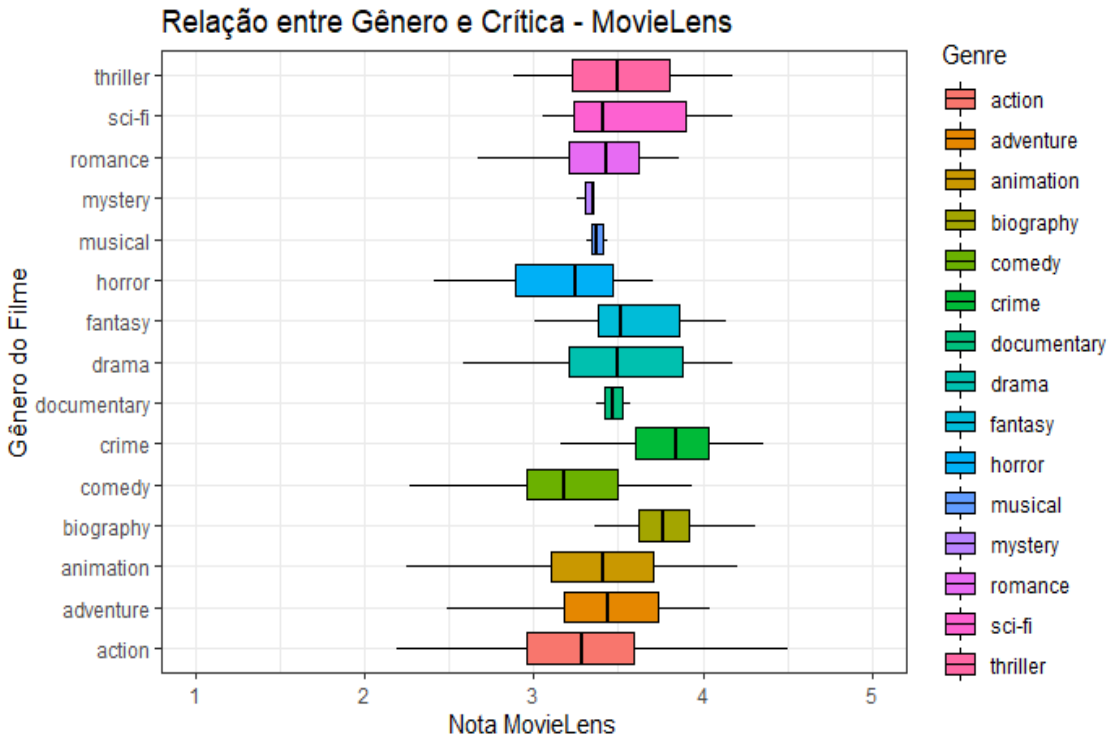


Figure 7: R quadrado = 0.0815231

6 Códigos do Projeto A2 Movies Dataset

```
---
title: "data_filmes_A2"
author: "Gabriel Machado"
date: "18/06/2020"
output: html_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

```{r}
moviesx <- read.csv("filmes_eua.csv")
moviesy <- read.csv("industria_filmes_eua.csv")

head(moviesx)

moviesx$Day.of.Week <- as.factor(moviesx$Day.of.Week)

```

```{r}
library(ggplot2)
moviesx <- read.csv("moviesx.txt")

moviesx$Day.of.Week <- factor(moviesx$Day.of.Week, levels = c("
    Tuesday", "Wednesday", "Thursday", "Friday","Saturday", "Sunday
    "))
```

```{r}
library(ggplot2)
library(readr)

industria_filmes_eua <- read_csv("industria_filmes_eua.csv")

moviesx <- read.csv("moviesx.txt")

moviesx$Day.of.Week <- factor(moviesx$Day.of.Week, levels = c("
    Friday", "Wednesday","Thursday","Tuesday","Saturday", "Sunday")
    )

moviesx$Price.min.relation <-  moviesx$Budget...mill./moviesx$
    Runtime..min.

```

```{r}
#n o utilizado pois n o fazia mas sentido essa vizualiza  o,
    n o indica nada relevante
# 2 - Rela  o entre or amento e Profit
ggplot(moviesx, aes(x=Budget...mill., y = Adjusted.Gross...mill.))
    +
  geom_point(size = 1) +  scale_y_discrete(breaks = c("911","100","
      1000","2000","2,363.60")) + theme_bw()
```

```{r}
#finalizado
#Rela  o entre dia da semana e dia da estr ia
```
```

```
53
54 ggplot(industria_filmes_eua, aes(x=industria_filmes_eua$Ano,y=
       industria_filmes_eua$Numero_de_novos_filmes_lancados)) + geom_
       col(aes(fill = FALSE)) +  scale_x_continuous(breaks = c
       (2008,2009,2010,2011,2012,2013,2014,2015,2016,2017),labels = c
       (2008,2009,2010,2011,2012,2013,2014,2015,2016,2017)) + labs(
       title="O efeito da crise de 2008 na ind stria de filmes dos
       EUA",x="Data",y="N mero de filmes novos lan ados") + guides(
       fill=FALSE) + theme_bw()
55 ```
56
57 ```{r}
58 #finalizado
59 #Rela  o entre dia da semana e dia da estr ia
60
61 ggplot(industria_filmes_eua, aes(x=industria_filmes_eua$Ano,y=
       industria_filmes_eua$`Ingressos_vendidos(1000)`/1000)) + geom_
       col(aes(fill = FALSE)) +  scale_x_continuous(breaks = c
       (2008,2009,2010,2011,2012,2013,2014,2015,2016,2017),labels = c
       (2008,2009,2010,2011,2012,2013,2014,2015,2016,2017)) + labs(
       title="O efeito da crise de 2008 na ind stria de filmes dos
       EUA",x="Data",y="Ingressos vendidos (milh es)") + guides(fill=
       FALSE) + theme_bw()
62 ```
63
64 ```{r}
65 #finalizado
66 #Rela  o entre dia da semana e dia da estr ia
67 ggplot(moviesx, aes(x=Day.of.Week, fill = FALSE)) + geom_histogram(
       stat="count", color = "black") +
68   labs(x = "Dia da Semana", y = "N mero de Estr ias", title = "
         Rela  o entre N mero de Estr ias e Dia da Semana") +
         guides(fill = FALSE) + coord_flip() + theme_bw()
69 ```
70
71 ```{r}
72 #finalizado
73 #rela  o tempo e cr tica - todos
74 ggplot(moviesx, aes(x=IMDb.Rating, y= Runtime..min., size = Budget
       ...mill., color = Genre)) +
75   geom_point(shape = 21, color = "black", aes(fill = FALSE), alpha
         = 0.5)+ xlim(3.5, 10) +
76   ylim(0,250) +
77   labs(x = "Nota do IMDb", y = "Tempo de Dura  o (Min)", title =
         "Rela  o entre Tempo de Dura  o do filme e Cr tica",size
          = "Custo (milh es)") + geom_smooth(method = "loess", se=
         FALSE,colour = "#0F422C",show.legend = FALSE) + guides(fill =
          FALSE) + theme_bw()
78
79 c7 <- moviesx$IMDb.Rating
80 c8 <- moviesx$Runtime..min.
81 #n o foi poss vel usar cor.test obtendo o valor exato de p pois
       haviam empates
82 #correla  o
83 #usammos o method spearman pois h  uma grnade quantidade de
       observa  es
84 cor(c7,c8, method = "spearman")
85 ```
86
87 ```{r}
88 #finalizado
89 #ANIMATION
```

```
ggplot(moviesx[moviesx$Genre =="animation",], aes(x=IMDb.Rating, y=
    Runtime..min., size = Budget...mill.)) +
  geom_point(shape = 21, color = "black", fill = "#E6ED58", alpha =
      0.5) + xlim(3.5, 10) +
  ylim(0,250) +
  labs(x = "Nota do IMDb", y = "Tempo de Dura  o (Min)", title =
      "Rela  o entre Tempo de Dura  o do filme e Cr tica -
      Animation",size = "Custo (milh es)")+ geom_smooth(method = "
      loess", se=FALSE,colour = "#563970",show.legend = FALSE) +
      theme_bw()

c5 <- moviesx[moviesx$Genre =="animation",]$IMDb.Rating
c6 <- moviesx[moviesx$Genre =="animation",]$Runtime..min.
#n o foi poss vel usar cor.test obtendo o valor exato de p pois
    haviam empates
#correla  o
cor(c5,c6, method = "spearman")
```

```{r}
#finalizado
#ACTION
#utilizamos o method loess no geom_smooth pois ele cria uma curva
    de tend ncia para um an lise mais precisa do que a reta do
    method lm
ggplot(moviesx[moviesx$Genre =="action",], aes(x=IMDb.Rating, y=
    Runtime..min., size = Budget...mill.)) +
  geom_point(shape = 21, color = "black", fill = "#FA0F00", alpha =
      0.5) + xlim(3.5, 10) +
  ylim(0,250) +
  labs(x = "Nota do IMDb", y = "Tempo de Dura  o (Min)", title =
      "Rela  o entre Tempo de Dura  o do filme e Cr tica -
      Action",size="Custo (milh es)") + geom_smooth(method = "
      loess", se=FALSE,colour = "#01BD4B",show.legend = FALSE)+
      theme_bw()

c3 <- moviesx[moviesx$Genre =="action",]$IMDb.Rating
c4 <- moviesx[moviesx$Genre =="action",]$Runtime..min.
#n o foi poss vel usar cor.test obtendo o valor exato de p pois
    haviam empates
#correla  o
cor(c3,c4, method = "spearman")
```

```{r}
#finalizado
#DRAMA
ggplot(moviesx[moviesx$Genre =="drama",], aes(x=IMDb.Rating, y=
    Runtime..min., size = Budget...mill.)) +
  geom_point(shape = 21, color = "black", fill = "#00D6FA", alpha =
      0.5) + xlim(3.5, 10) +
  ylim(0,250) +
  labs(x = "Nota do IMDb", y = "Tempo de Dura  o (Min)", title =
      "Rela  o entre Tempo de Dura  o do filme e Cr tica -
      Drama", size="Custo (milh es)") + geom_smooth(method = "
      loess", se=FALSE,colour = "#BD641C",show.legend = FALSE) +
      theme_bw()
c1 <- moviesx[moviesx$Genre =="drama",]$IMDb.Rating
c2 <- moviesx[moviesx$Genre =="drama",]$Runtime..min.
#n o foi poss vel usar cor.test pois p tinha valores com empates
#correla  o
cor(c1,c2, method = "spearman")
```

```
```

```{r}
#### 4 - Compara o entre pre o/min.
ggplot(moviesx, aes(x=Genre, y= Price.min.relation, fill = Genre)) +
      labs(y="Rela o Pre o/min",x="G nero do Filme",legend="
     G nero") + geom_boxplot(outlier.shape = NA, color = "black")
     + coord_flip() + guides(fill = FALSE) + theme_bw()
#R quadrado - Nominal x Num rica
summary(lm(moviesx$Price.min.relation~moviesx$Genre))$r.squared
```

```{r}
#finalizado
# 5 - Rela o entre Runtime e IMDb?
#Rela o entre IMDb e MovieLens? H uma linerialidade? Sim
ggplot(moviesx, aes(x=IMDb.Rating, y= MovieLens.Rating,xlim(0,10),
     ylim(0,5), size = Budget... mill., color = Genre)) +
  geom_point(shape = 21, color = "black", aes(fill = FALSE), alpha
       = 0.5) +
   labs(x = "Nota do IMDb", y = "Nota do Movie Lens", title = "
       Rela o entre Notas de Sites de Cr tica",size = "Custo (
       milh es)") + guides(fill = FALSE) + geom_smooth(method = "
       loess", se=FALSE,colour = "#0F422C",show.legend = FALSE)+
       theme_bw()
#correla o qualitativo x qualitativo = Ordinal x Ordinal
#alto n vel de correla o, assim, claro que a forma de
     an lise dos filmes semelhante nos dois sites de cr tica
cor(moviesx$IMDb.Rating, moviesx$MovieLens.Rating)
```

```{r}
#finalizado
#rela o entre g nero e IMDB rating
ggplot(moviesx, aes(x=Genre, y= IMDb.Rating, fill = Genre)) + geom
     _boxplot(outlier.shape = NA,color = "black") +
   labs(x = "G nero do Filme", y = "Nota IMDb", title = "Rela o
       entre G nero e Cr tica - IMDb") + ylim(c(1,10)) + coord_
       flip() + theme_bw()
#R quadrado - Nominal x Num rica
#levemente mais parcial do que o MovieLens, por m os valores ainda
      s o baixos e pouco representativos
summary(lm(moviesx$IMDb.Rating~moviesx$Genre))$r.squared
```

```{r}
#finalizado
#rela o entre g nero e IMDB rating
ggplot(moviesx, aes(x=Genre, y= MovieLens.Rating, fill = Genre)) +
      geom_boxplot(outlier.shape = NA,color = "black") +
   labs(x = "G nero do Filme", y = "Nota MovieLens", title = "
       Rela o entre G nero e Cr tica - MovieLens") + ylim(c
       (1,5)) + coord_flip() + theme_bw()
#R quadrado - Nominal x Num rica
summary(lm(moviesx$MovieLens.Rating~moviesx$Genre))$r.squared
```

Listing 1: Código fonte em R