Projeto IMDB

Natanael Magalhães Cardoso*

Questão 01

A Tabela 1 mostra o número de filmes, duração média, nota média, máxima e mínima, desvio padrão das notas e soma de likes para cada faixa de orçamento.

Table 1: Sumário 1

Faixa de orçamento	Número de filmes	Duração média	Nota média	Nota máxima	Nota mínima	Desvio padrão da nota	Soma de likes
<10mi	423	99.5461	6.381087	9.2	2.0	1.1611694	3021869
10mi-20mi	271	104.9446	6.332841	9.0	1.6	1.1986068	3242027
20mi-30mi	211	108.2322	6.199052	9.3	2.7	1.2513989	2956386
30mi-40mi	163	113.0675	6.342331	8.6	3.7	0.9989436	2424807
40mi-50mi	110	114.8909	6.370909	8.5	4.1	0.9922031	1874404
50mi-60mi	96	119.4271	6.388542	8.8	2.4	1.0831601	1699363
>60mi	363	119.2066	6.602479	9.0	2.2	1.0986277	12117851

Questão 02

Como pode ser visto na Figura 1, a duração média dos filmes tende a aumentar com o aumento do orçamento.

Questão 03

A maior nota média foi obtida para filmes produzidos com orçamento superior a 60mi. Já a maior nota total foi obtida na faixa de orçamento de 20mi-30mi. E, a menor nota total foi obtida para faixa de orçamento de 10mi-20mi. O filme *Justin Bieber: Never Say Never* obteve a menor nota e o filme *The Shawshank Redemption* obteve a maior nota.

Questão 04

As faixas 30mi-40mi e 40mi-50mi apresentam menor variação nas notas, pois têm o menor desvio padrão, como mostra a Tabela 1.

*nUSP: 8914122. Usando dados para o segundo grupo (datas entre 2000-2010)

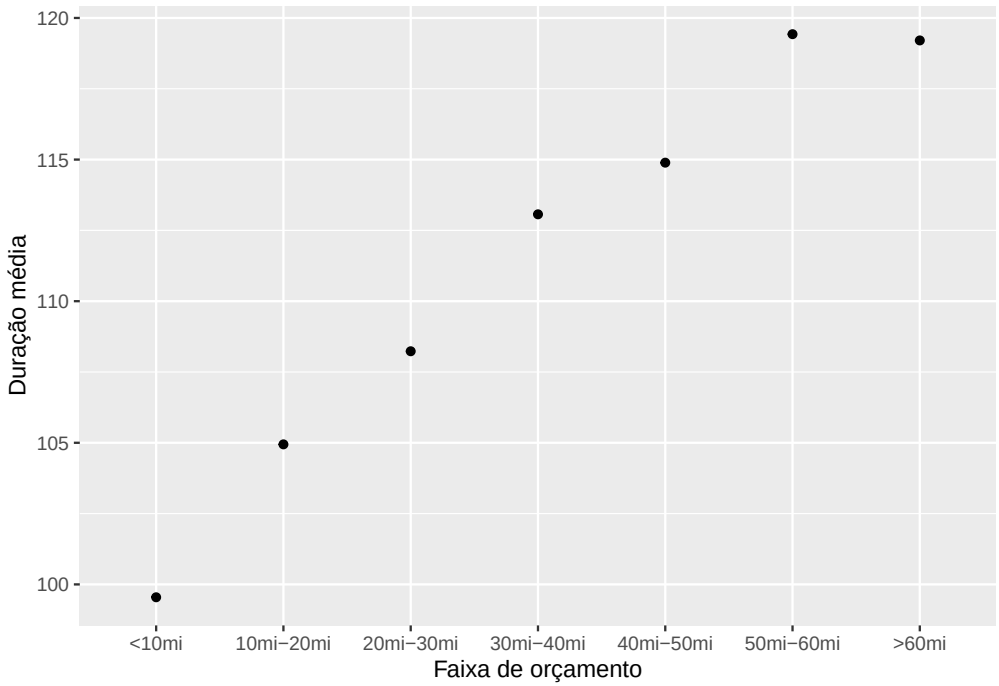

Figure 1: Duração média para cada faixa de orçamento

Questão 05

A Tabela 2 mostra o número de filmes, duração média, nota média, máxima e mínima, desvio padrão das notas e soma de likes para cada faixa de orçamento para filmes lançados entre 2000 e 2010.

Table 2: Sumário 2

Faixa de orçamento	Número de filmes	Duração média	Nota média	Nota máxima	Nota mínima	Desvio padrão da nota	Soma de likes
<10mi	181	97.98895	6.225967	8.5	2.0	1.160957	454171
10mi-20mi	111	100.84685	5.936937	8.2	3.3	1.036772	383317
20mi-30mi	97	102.82474	5.896907	8.3	2.7	1.273238	489447
30mi-40mi	66	103.28788	5.990909	8.2	4.0	1.042147	377849
40mi-50mi	41	107.26829	6.082927	8.5	4.4	1.107678	240075
50mi-60mi	36	119.36111	6.252778	8.2	2.4	1.256182	259869
>60mi	100	116.57000	6.393000	9.0	2.2	1.441187	797787

Questão 06

Comparando a Tabela 1 com a Tabela 2, é possivel notar que a tendência do aumento da duração média com o aumento do orçamento continua presente na amostra de filmes lançados entre 2000 e 2010. Já a duração média dos filmes lançados nesta amostra é menor que a duração média dos filmes lançados em todo intervalo de tempo analisado para todas as faixas de orçamento. Além disso, é notado as faixas de orçamento para os filmes com nota máxima e mínima não são as mesmas. Na amostra de filmes lançados entre 2000 e 2010, a nota máxima é obtida para filmes lançados com orçamento superior a 60mi e a nota mínima é obtida para filmes lançados com orçamento inferior a 10mi. Já a maior nota média continua para filmes lançados com orçamento maior que 60mi.

Questão 07

O número de classes foi determinado por $K \approx \sqrt{n} \approx \sqrt{632} \approx 25$, onde n é o número de elementos da amostra. A medida de assimetria, A, é calculada usando a equação de Assimetria de Pearson $A = \frac{\bar{x} - m_0}{S_x} = 0,1815$, onde $\bar{x}$ é a média amostral, m_0 é a moda amostral e S_x é o desvio padrão amostral. Como $0,15 < |A| < 1$, pode-se dizer que esta distribuição é moderadamente assimétrica. Isto pode ser confirmado na Figura 2, onde percebe-se que o histograma é ligeiramente assimétrico.

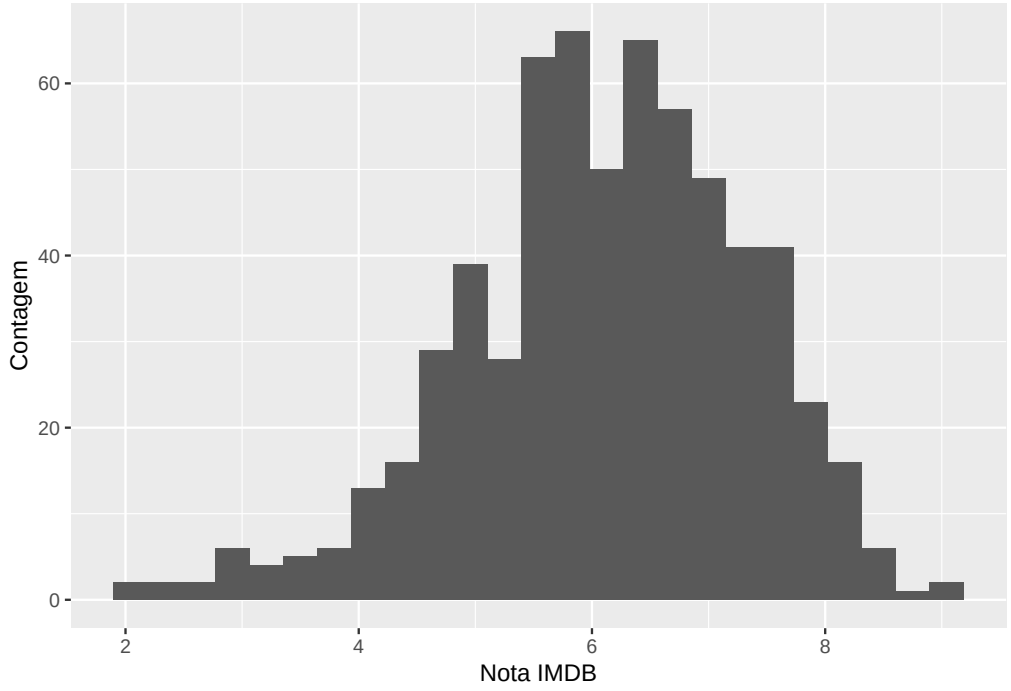

Figure 2: Histograma das Notas no IMDB para a amostra.

Questão 08

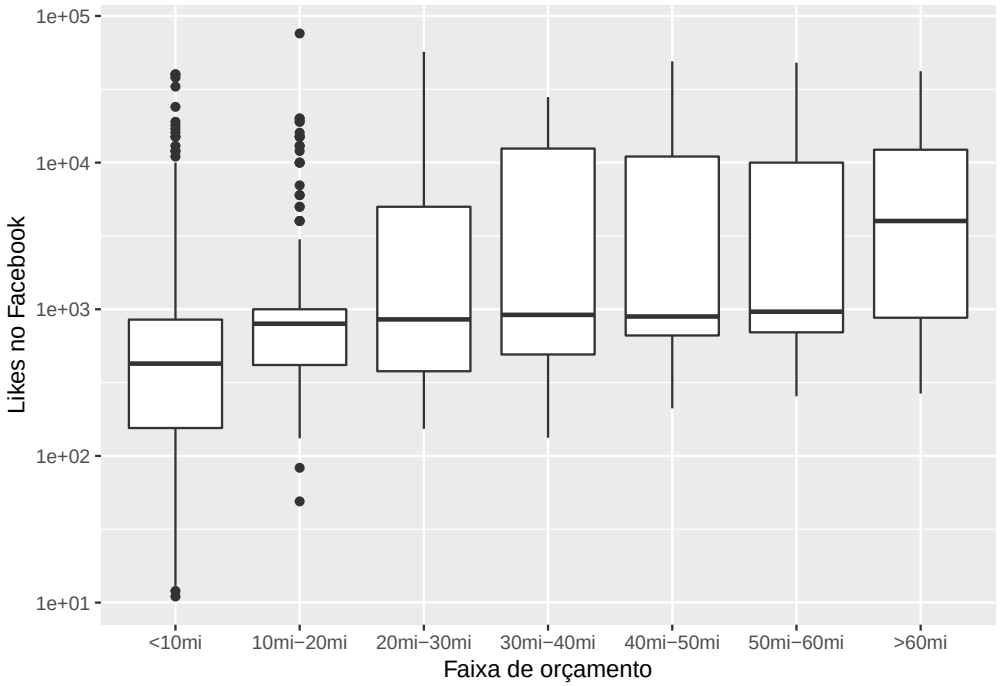

Figure 3: Boxplot do número de likes para cada faixa de orçamento.

Questão 09

Pelo gráfico da Figura 3, é possível notar que a mediana do número de likes no facebook aumenta com o aumento da faixa de orçamento. Como as caixas representam o intervalo interquartil (IQR) entre o 25° e o 75° percentil, é notável que o IQR também tende a ser maior com o aumento da faixa de orçamento. O aumento do IQR junto com a diminuição do *whisker* mostra que a maioria dos valores da distribuição tende a se concentrar entre os percentis 25 e 75 de acordo com o aumento da faixa de orçamento. Isso também é notado nos *outliers*, que só aparecem nas menores faixas de orçamento.

Questão 10

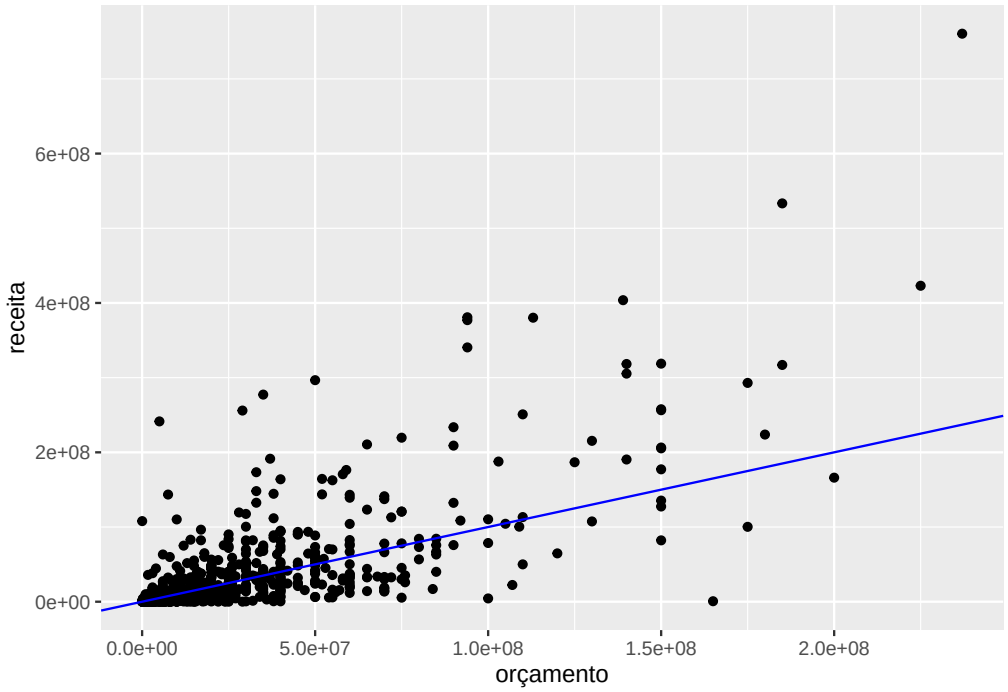

Figure 4: Gráfico de dispersão da receita em função do orçamento.

Pela análise do gráfico de dispersão da Figura 4, é possível notar que os dados estão são mais concentrados para menores valores de orçamento e receita. E, a medida que se aumenta o orçamento, a disperção dos dados também aumenta. A reta azul mostra o limiar *orçamento* = *receita*, pontos acima dela mostram produções que tiveram lucro (receita > orçamento), enquanto que pontos abaixo mostram produções que não tiveram lucro.

Códigos

Bibliotecas

```
library(tidyverse)
library(readr)
library(ggplot2)
```

Questão 01

```
imdb <- read_rds("imdb2.rds")
imdb$estreia = fct_reorder(
  imdb$estreia,
  imdb$ordem,
  min
)
imdb$"Faixa de orçamento" = fct_reorder(
  imdb$"Faixa de orçamento",
  imdb$"ordem orçamento",
  min
)

sum1 <- imdb %>%
  group_by(`Faixa de orçamento`) %>%
  summarise(
    "Número de filmes"=n(),
    "Duração média"=mean(duracao),
    "Nota média"=mean(nota_imdb),
    "Nota máxima"=max(nota_imdb),
    "Nota mínima"=min(nota_imdb),
    "Desvio padrão da nota"=sd(nota_imdb),
    "Soma de likes"=sum(likes_facebook)
  )
sum1
```

Questão 02

```
ggplot(sum1, aes(x=`Faixa de orçamento`, y=`Duração média`)) + geom_point()
```

Questão 03

```
imdb %>% slice_min(nota_imdb)
imdb %>% slice_max(nota_imdb)
```

Questão 05

```
df_g2 <- imdb %>% filter(estreia=="2000-2010")

sum2 <- df_g2 %>%
  group_by(`Faixa de orçamento`) %>%
  summarise(
    "Número de filmes"=n(),
```

```
    "Duração média"=mean(duracao),
    "Nota média"=mean(nota_imdb),
    "Nota máxima"=max(nota_imdb),
    "Nota mínima"=min(nota_imdb),
    "Desvio padrão da nota"=sd(nota_imdb),
    "Soma de likes"=sum(likes_facebook)
  )
sum2
```

Questão 07

```
compute_mode <- function(v) {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}

compute_asymmetry <- function(d) {
  (mean(d) - compute_mode(d)) / sd(d)
}

compute_asymmetry(df_g2$nota_imdb)
```

```
ggplot(df_g2, aes(x=nota_imdb)) +
  geom_histogram(bins=25) +
  labs(x="Nota IMDB", y="Contagem")
```

Questão 08

```
ggplot(df_g2, aes(
    x=fct_reorder(`Faixa de orçamento`, `ordem orçamento`, min),
    y=likes_facebook)
  ) +
  geom_boxplot() +
  scale_y_log10() +
  labs(x="Faixa de orçamento", y="Likes no Facebook")
```

Questão 10

```
ggplot(df_g2, aes(x=`orcamento`, y=`receita`)) +
  geom_point() +
  geom_abline(intercept=0, slope=1, color="blue") +
  labs(x="orçamento")
```