Minhas Perguntas

18 de maio de 2016

```
library("dplyr")
library("ggplot2")
library("resample")

REP = 10000

movies <- read.csv("/Users/brunodias/Documents/FPCC2/Lab3/dataset/movies_expanded.csv")
ratings_summ <- read.csv("/Users/brunodias/Documents/FPCC2/Lab3/dataset/ratings_summ.csv")
tags_summ <- read.csv("/Users/brunodias/Documents/FPCC2/Lab3/dataset/tags_summ.csv")
tags <- read.csv("/Users/brunodias/Documents/FPCC2/Lab3/dataset/tags.csv")

movies.me <- merge(movies, ratings_summ, by = "movieId", all.x= TRUE) %>% na.omit
movies.me <- merge(movies.me, tags_summ, by = "movieId", all.x = TRUE) %>% na.omit
```

Perguntas

Foram selecionadas duas perguntas da lista inicial para resolver esse checkpoint.

1. **Existe diferença na variação de quantidade de ratings dos filmes lançados entre 2000-2015 e dos lançados entre 1985-2000?** R: Acredito que a variância na quantidade de ratings será maior em filmes lançados entre 1985-2000 por ter engajamento menor da comunidade (talvez menos pessoas assistiram e fizeram uma avaliação)
2. **Qual a média de ratings dos filmes lançados nos últimos 10 anos?** R: Deve seguir um padrão e ter uma variação das médias pequena em todos os anos.

Análise exploratória

Os dados que iremos explorar fazem parte de um *dataframe* com um conjunto de filmes, seus gêneros e suas avaliações de qualidade. Utilizaremos a amostragem de dados maior disponível na página do site. Essa amostragem contém informações de 34 mil filmes. Abaixo temos o resumo da estrutura dos nossos dados.

- **moviesId**: identificador único do filme
- **title**: título do filme
- **year**: ano do filme
- **NumGenres**: Quantidade total de gêneros do filme.
- **gêneros (vários)**: os gêneros que o filme tem, quando for 0 o gênero não está contido no filme, quando for 1 o mesmo está contido.
- **meanRatings**: média de ratings por filme
- **qtdRatings**: quantidade de ratings por filme

Algumas ajustes e agregações no *dataframe* original foram feitos para facilitar o tratamento. Adicionamos uma coluna com o ano do filme, extraído a partir do titulo do filme no df original, também foram adicionadas colunas da sumarização de informações sobre *ratings* e *tags* de cada filme.

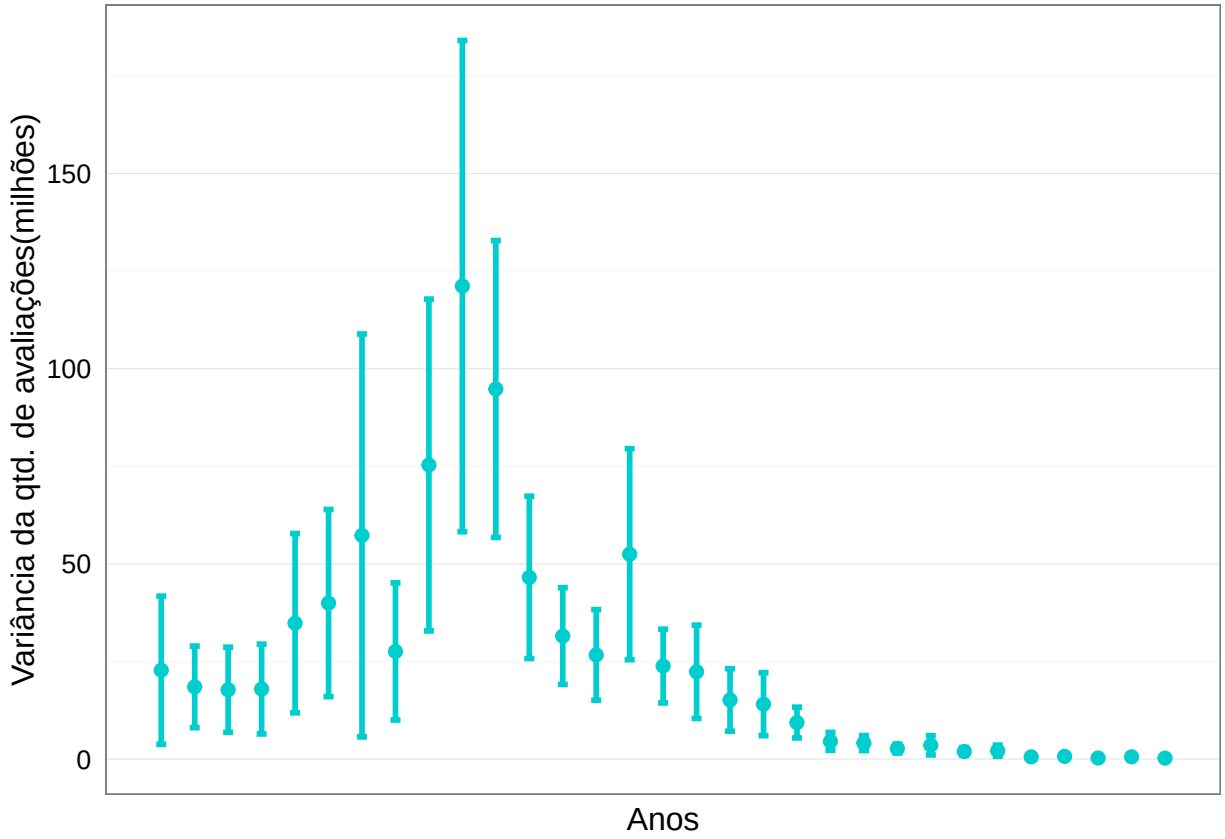
Para realizar essa sumarização foi utilizado a ferramenta feita por um dos alunos que se encontra disponível aqui. Ajudou bastante durante essa análise. Agradeço a disponibilização da ferramenta.

Variação da quantidade de *ratings*

Existe diferença na variação de quantidade de ratings dos filmes lançados entre 2000-2015 e dos lançados entre 1985-2000?

Através de *resample* vamos gerar uma visualização para analisarmos se há diferença entre a variação da quantidade de ratings nesses dois períodos.

```
resampleRatings = function(n) {
  x <- movies.me %>% filter(year == n)
  b = bootstrap(x$qtdRating, var, R = REP)
  CI.percentile(b, probs = c(.025, .975))
}

ratings_plot <- data.frame(year = c(), upper = c(), mean = c(), lower = c())
for (n in 1985:2015) {
  a <- resampleRatings(n)
  ratings_plot <- rbind(ratings_plot, data.frame(
    year = n,
    mean = mean(a),
    lower = a[1],
    upper = a[2]
  ))
}
ratings_plot <- ratings_plot[-1]/1e+06
# Visualização da variação de quantidade de ratings
ratings_plot %>% ggplot(aes(x = 1:nrow(ratings_plot), y = mean)) +
  geom_point(size = 2, color = "cyan3") +
  geom_errorbar(aes(
    ymin = lower,
    ymax = upper), width=.3, size=1, color = "cyan3") +
  xlab("Anos") +
  ylab("Variância da qtd. de avaliações(milhões)") +
  scale_x_continuous(breaks = seq(1985,2015)) +
  theme_bw() +
  theme(axis.ticks=element_blank())
```



Como podemos ver, a suspeita inicial se comprovou no gráfico acima, a variação da quantidade de avaliações é significativamente maior para os filmes entre 80 e 00 do que para os filmes entre 01 e 15. Iremos aplicar a técnica de *bootstrap* para saber exatamente quanto será a diferença de variação. Podemos comprovar abaixo que existe uma diferença clara entre os dois períodos.

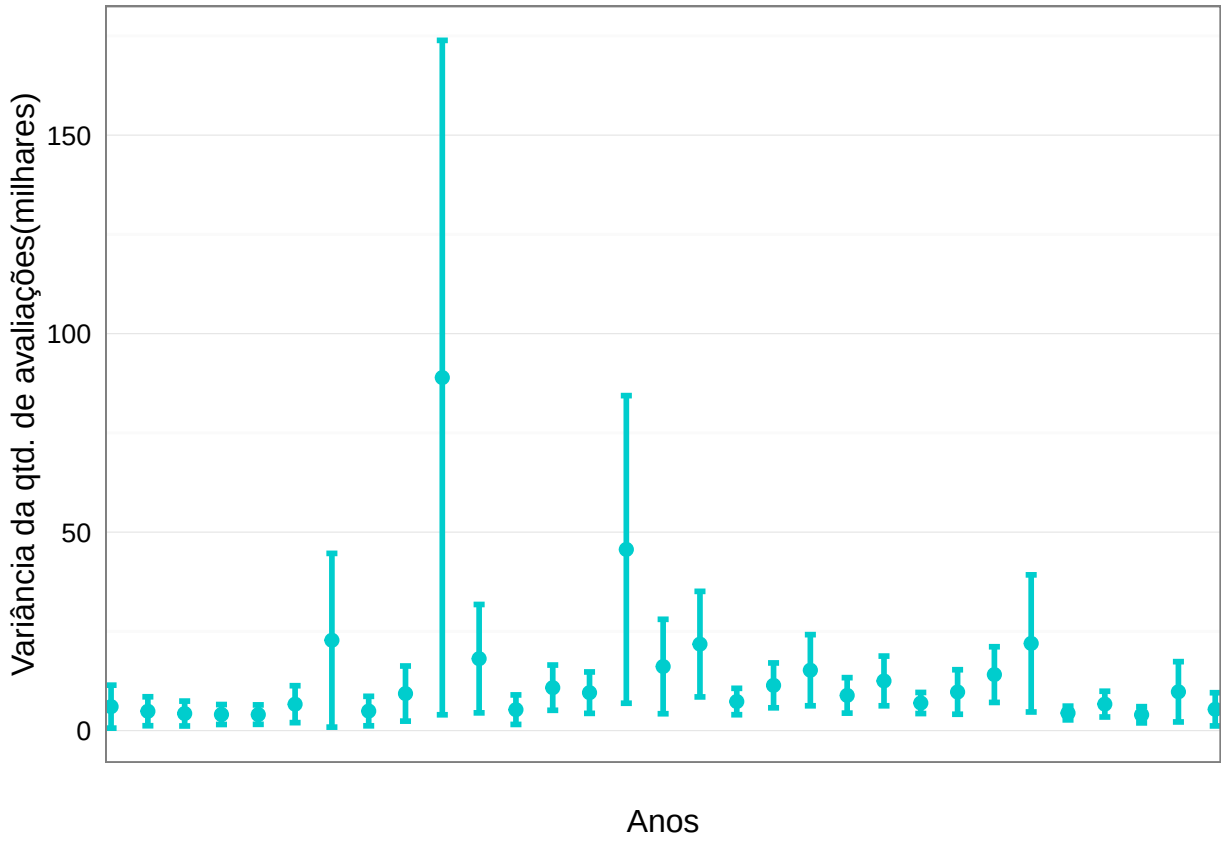
```
ratings_85.00 <- movies.me %>% filter(year >= 1985 & year <= 2000)
ratings_01.15 <- movies.me %>% filter(year >= 2001 & year <= 2015)

diff_ratings <- bootstrap2(
  data=ratings_85.00$qtdRatings,
  data2=ratings_01.15$qtdRatings,
  statistic = var)
CI.percentile(diff_ratings, probs = c(.025, .975))
```

```
FALSE                       2.5%    97.5%
FALSE var: data-data2 31583838 47008238
```

Tendo em vista essa análise surge outra dúvida, será que existe semelhança na tendência que o gráfico anterior mostra, quando analisamos a variação da quantidade de tags que os filmes tem? Segue abaixo a análise desse questionamento.

```
resampleTags = function(n) {
  x <- movies.me %>% filter(year == n)
  b = bootstrap(x$qtdTags, var, R = REP)
  CI.percentile(b, probs = c(.025, .975))
}
```

```r
tags_plot <- data.frame(year = c(), upper = c(), mean = c(), lower = c())
for (n in 1985:2015) {
  a <- resampleTags(n)
  tags_plot <- rbind(tags_plot, data.frame(
    year = n,
    mean = mean(a),
    lower = a[1],
    upper = a[2]
  ))
}
tags_plot <- tags_plot[-1]/1000
# Visualização da variação de quantidade de ratings
tags_plot %>% ggplot(aes(x = 1:nrow(tags_plot), y = mean)) +
  geom_point(size = 2, color = "cyan3") +
  geom_errorbar(aes(
    ymin = lower,
    ymax = upper), width=.3, size=1, color = "cyan3") +
  xlab("Anos") +
  ylab("Variância da qtd. de avaliações(milhares)") +
  scale_x_discrete(limits = c(1985:2000)) +
  theme_bw() +
  theme(axis.ticks=element_blank())
```



Com essa visualização podemos afirmar que a organização dos dados não se assemelha de maneira alguma com nossa avaliação anterior da quantidade de ratings.

Porém, não podemos afirmar nada sobre a tendência dos dados ao longo dos anos, eles não parecem seguir

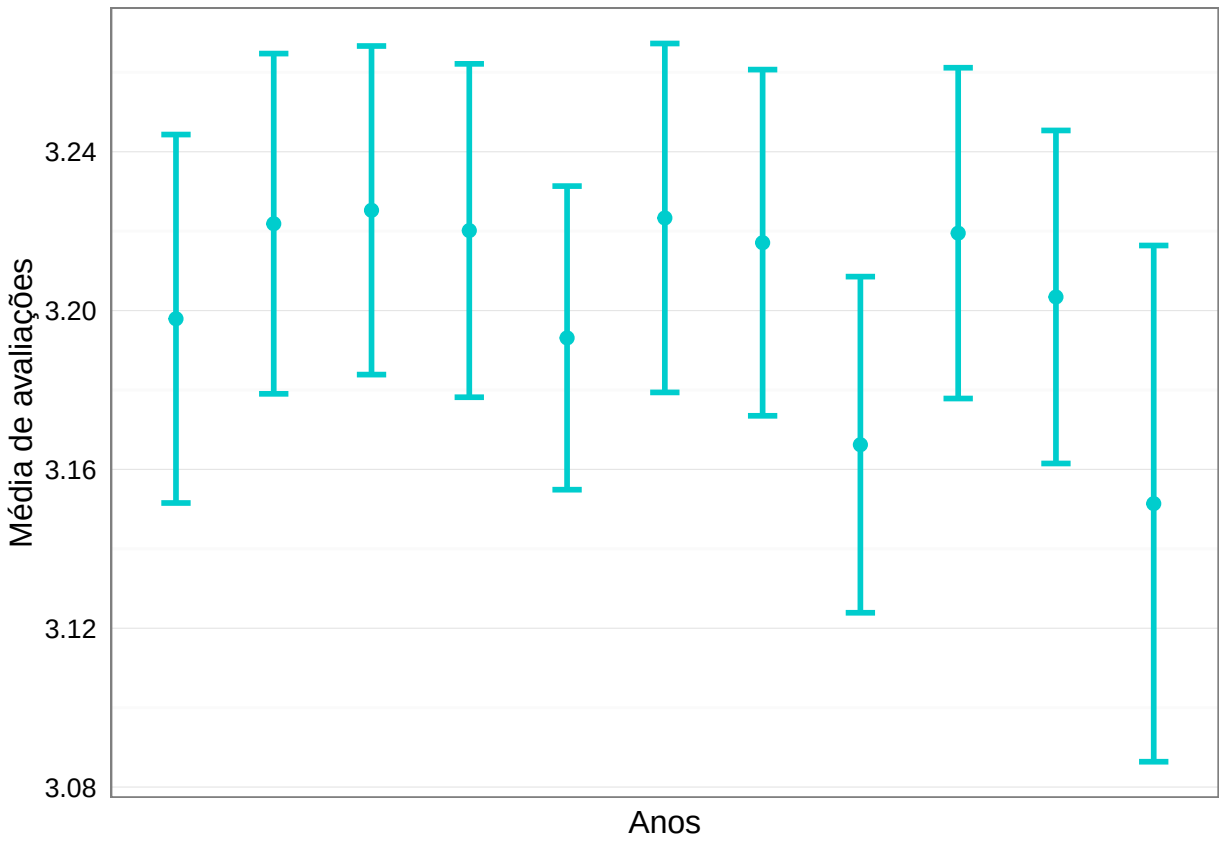
nenhum tipo de tendência ascendente ou descendente.

Média de ratings dos filmes

Qual a média de ratings dos filmes lançados nos últimos 10 anos?

Para responder essa questão precisamos analisar os filmes e seus ratings dentre os anos de 2005 e 2015. É importante informar que desconsideramos o ano de 2016 por ser o ano corrente dessa análise e não termos uma amostragem significativa de dados deste ano.

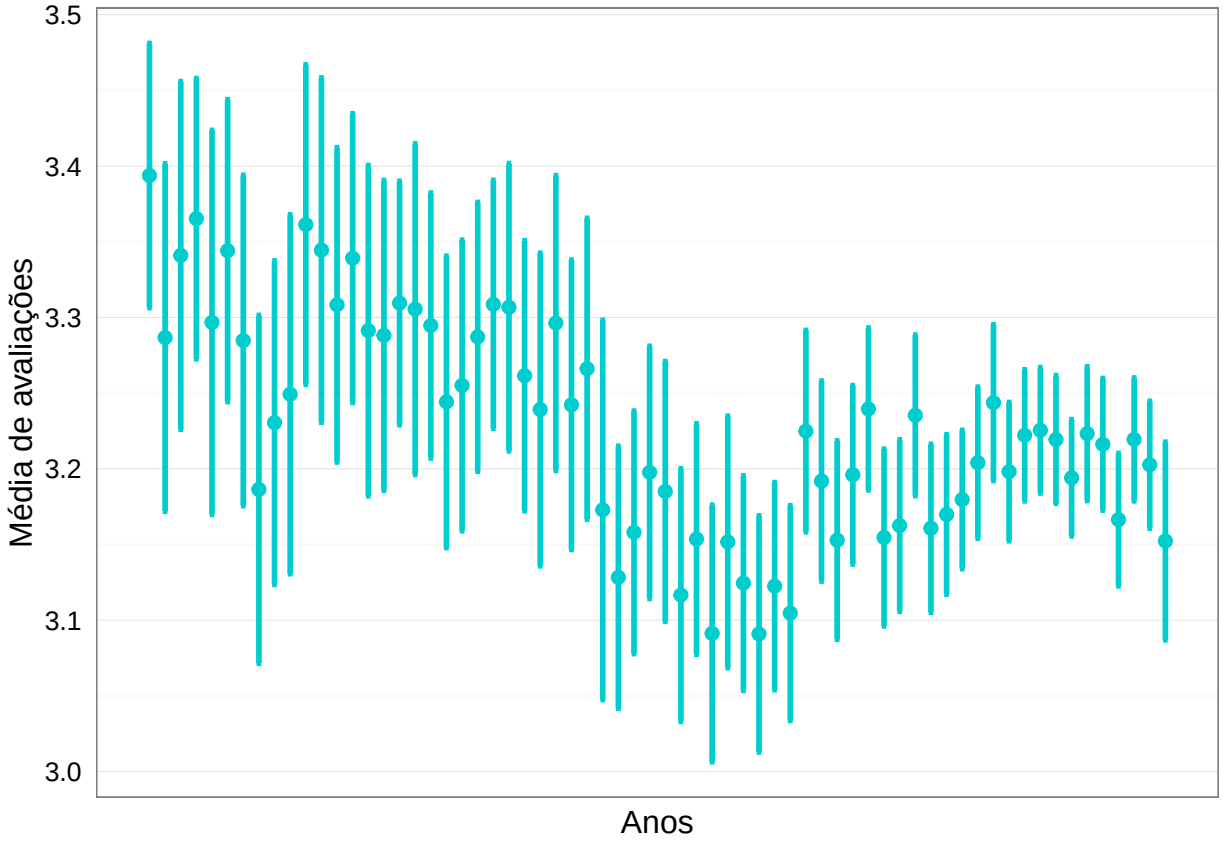
```
resampleRatingMean = function(n) {
  x <- movies.me %>% filter(year == n)
  b = bootstrap(x$meanRatings, mean, R = REP)
  CI.percentile(b, probs = c(.025, .975))
}

ratingAll_plot <- data.frame(year = c(), upper = c(), mean = c(), lower = c())
for (n in 2005:2015) {
  a <- resampleRatingMean(n)
  ratingAll_plot <- rbind(ratingAll_plot, data.frame(
    year = n,
    mean = mean(a),
    lower = a[1],
    upper = a[2]
  ))
}

ratingAll_plot %>% ggplot(aes(x = 1:nrow(ratingAll_plot), y = mean)) +
  geom_point(size = 2, color = "cyan3") +
  geom_errorbar(aes(
    ymin = lower,
    ymax = upper), width=.3, size=1, color = "cyan3") +
  xlab("Anos") +
  ylab("Média de avaliações") +
  scale_x_continuous(breaks = seq(2005,2015)) +
  theme_bw() +
  theme(axis.ticks=element_blank())
```



Como previsto inicialmente temos uma diferença muito pequena das médias ao longo dos últimos anos. Podemos perceber que, em geral, os filmes tem médias de avaliações entre 3.1 e 3.4. Isso quer dizer que em uma escala de 0 a 5, os filmes agradam seus espectadores na maioria das vezes.

Com isso, podemos levantar uma questão interessante, será que ao longo das décadas temos uma evolução nesse quadro, ou sempre segue dentro dessa faixa de avaliação? Vamos analisar agora a partir da década de 50 até os dias atuais.

```
ratingMean_plot <- data.frame(year = c(), upper = c(), mean = c(), lower = c())
for (n in 1950:2015) {
  a <- resampleRatingMean(n)
  ratingMean_plot <- rbind(ratingMean_plot, data.frame(
    year = n,
    mean = mean(a),
    lower = a[1],
    upper = a[2]
  ))
}
ratingMean_plot %>% ggplot(aes(x = 1:nrow(ratingMean_plot), y = mean)) +
  geom_point(size = 2, color = "cyan3") +
  geom_errorbar(aes(
    ymin = lower,
    ymax = upper), width=.3, size=1, color = "cyan3") +
  xlab("Anos") +
  ylab("Média de avaliações") +
  scale_x_continuous(breaks = seq(2005,2015)) +
  theme_bw() +
  theme(axis.ticks=element_blank())
```



Podemos perceber que existe uma tendência ao longo dos anos de diminuição nas notas dos filmes. Mas será que isso nos aponta que os filmes vem perdendo qualidade ao longo dos anos? Ou até que a tendência ao longo dos anos é de diminuição dessa qualidade? Será que em 2050 teremos filmes com avaliações piores?

Supomos que não, pois podemos notar que as barras de **intervalo de confiança** dos últimos anos tendem a ser menores do que as dos anos anteriores e quanto mais antigo os filmes, mais imprecisos tendem a ser seus intervalos de confiança, ou seja, maiores são suas barras.

Aparentemente isso nos mostra que há uma amostragem menor de avaliações a medida que os filmes tendem a ser mais antigos, provavelmente se tivéssemos exatamente a mesma amostragem de avaliações por filme ao longo dos anos, a média se aproximaria de modo que seria possível afirmar que não há diferença na qualidade dos filmes ao longo das décadas.