Untitled

Ana Luzielma
Jaylhane Nunes
Raianny Soares

07/02/2022

Introdução

Contextualização

Motivadas pelo interesse comum em leitura optamos por realizar a análise de um conjuntos de dados envolvendo livros.

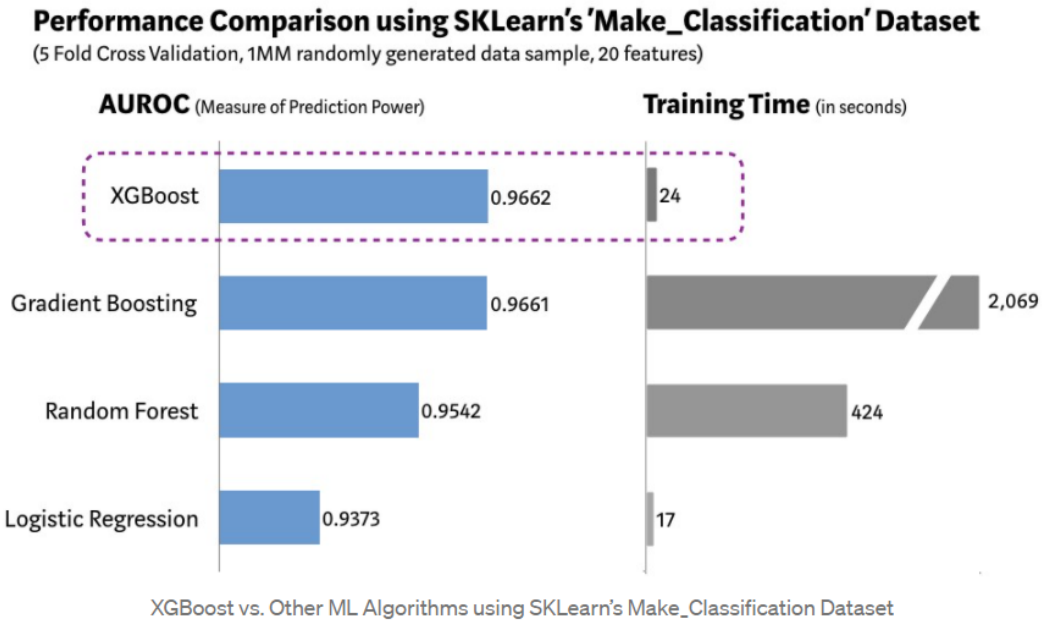
O conjunto de dados selecionado possui **11.131 observações**, foi gerado por meio de **raspagem** de dados na **API** da plataforma **GoodReads** e disponibilizado por ***Soumik*** no site **Kagle**.

Nele é possível encontras as seguintes colunas:

bookID	average_rating	language_code	text_reviews_count
title	isbn	num_pages	publication_date
authors	isbn13	ratings_count	publisher

Chegamos a um consenso que diversos fatores influenciam na satisfação com a leitura e quisemos investigar se, com os dados disponíveis, seria possível obter um modelo que conseguisse predizer se o livro foi considerado: *ruim*, *bom* ou *ótimo* .

Uma Ideia Inicial



Dentre as leituras realizadas sobre as possibilidades de modelos e métodos, percebemos que o XGboost tinha um ótimo desempenho comparado a outros modelos, e que, apesar da variável `average_rating` ser uma variável contínua um método de classificação poderia ser adequado para os nossos objetivos, desde que gerassemos categorias e encaixassemos os intervalos.

A Inspiração Final

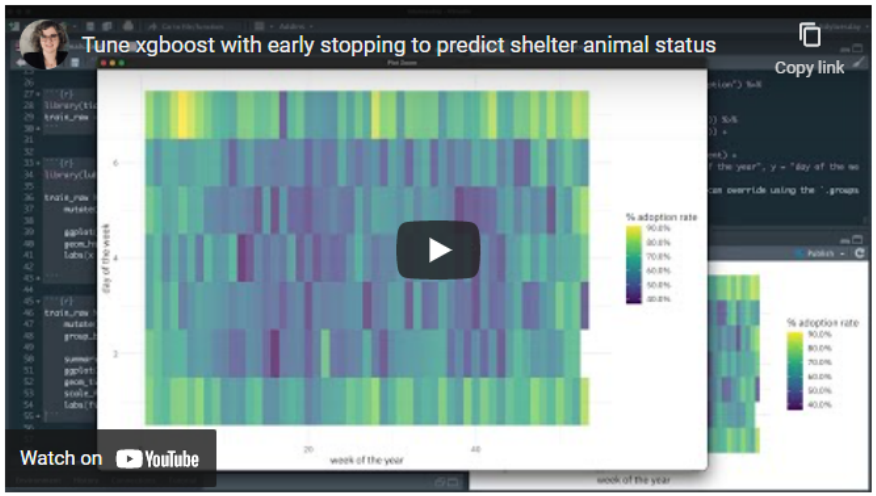
Julia Silge

ABOUT BLOG

Tune xgboost models with early stopping to predict shelter animal status

By Julia Silge in rstats tidymodels
AUGUST 7, 2021

This is the latest in my series of screencasts demonstrating how to use the tidymodels packages, from just getting started to tuning more complex models. I participated in this week's episode of the SLICED playoffs, a competitive data science streaming show, where we competed to predict the status of shelter animals. 😺 I used xgboost's early stopping feature as I competed, so let's walk through how and when to try that out!



Indicamos

Julia Silge

DATA SCIENTIST & SOFTWARE ENGINEER

I'm a tool builder, author, international keynote speaker, and real-world practitioner focusing on data analysis and machine learning. I love making beautiful charts and communicating about technical topics with diverse audiences.

READ LATEST POSTS →

juliasilge.com

- *Agora, ao nosso modelo!*

Desenvolvimento

Engenharia de Dados

A análise exploratória consistiu em:

- Limpeza dos Dados
- Análise Descritiva

Dado os nossos objetivos, percebemos que algumas colunas eram dispensáveis e outras poderiam ser transformadas, de forma que:

Excluídas	Transfomadas	Geradas
bookID	average_rating	book_rating[1]
title	language_code	
isbn	text_reviews_count	prop_text_reviews[2]
authors	publication_date	
isbn13		
publisher		

Nosso **MAIOR** *desafio durante a análise e a modelagem esteve relacionado a essa fase de engenharia de dados*

Análise Exploratória

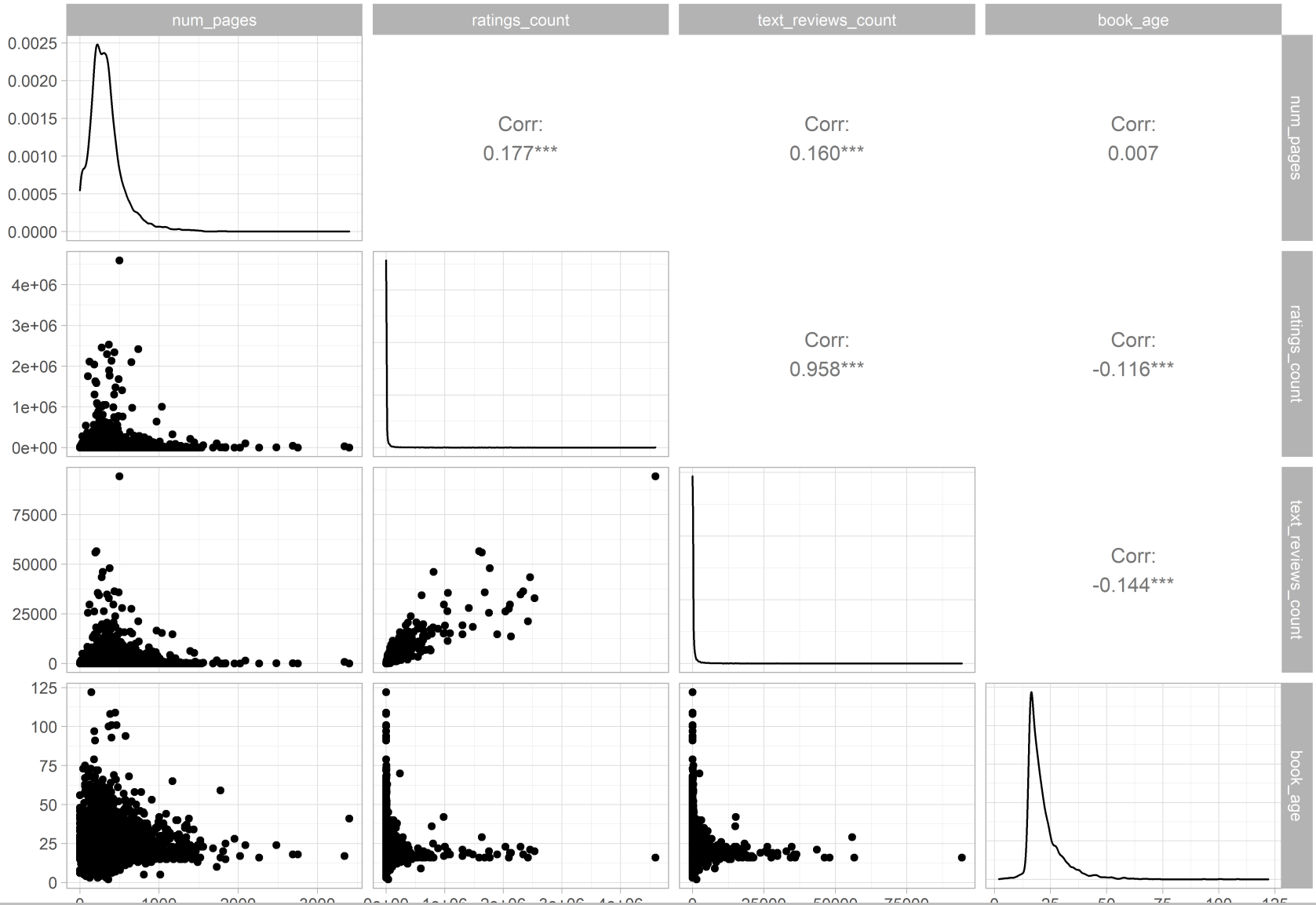
- Para iniciar separamos o conjunto em **treino** e **teste** baseando-nos em 75% das observações e balanceando-as com nossa variável resposta: `book_rating`.

```
set.seed(1904, kind = "Mersenne-Twister", normal.kind = "Inversion")
livros_split <- initial_split(livros, prop = .75, strata = book_rating)
livros_treino <- training(livros_split)
livros_teste <- testing(livros_split)
```

- Em seguida verificamos a dispersão e correlação entre as variáveis numéricas:

```
livros_treino %>%
  select(where(is.numeric)) %>%
  ggpairs(upper = list(continuous = wrap("cor", method = "spearman")))
```

Análise Exploratória



Análise Exploratória

Com o resultado anterior percebemos que:

- Correlação forte entre `text_reviews_count` e `ratings_count`.
- Pensamos em gerar uma variável que considerasse a quantidade de `text_reviews_count`,pois consideramos que isso seria um indicativo importante para nossa resposta e chegamos a uma variável que correspondesse a proporção entre `text_reviews_count/ratings_count`

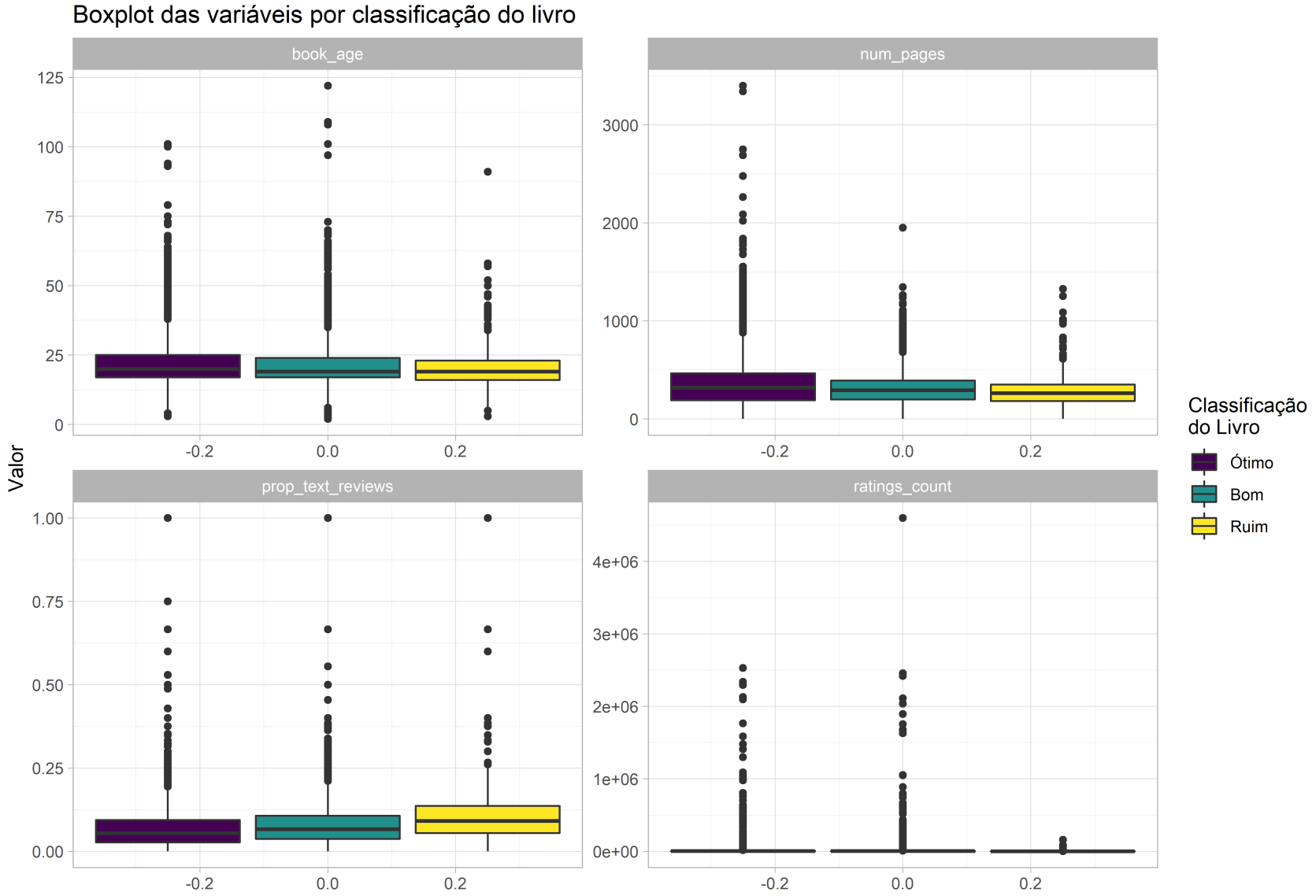
```
livros_treino <- livros_treino %>%
  mutate(prop_text_reviews = text_reviews_count / ratings_count) %>%
  select(-text_reviews_count)

cor(livros_treino$prop_text_reviews,livros_treino$ratings_count,
    use = "complete", method = "spearman")
```

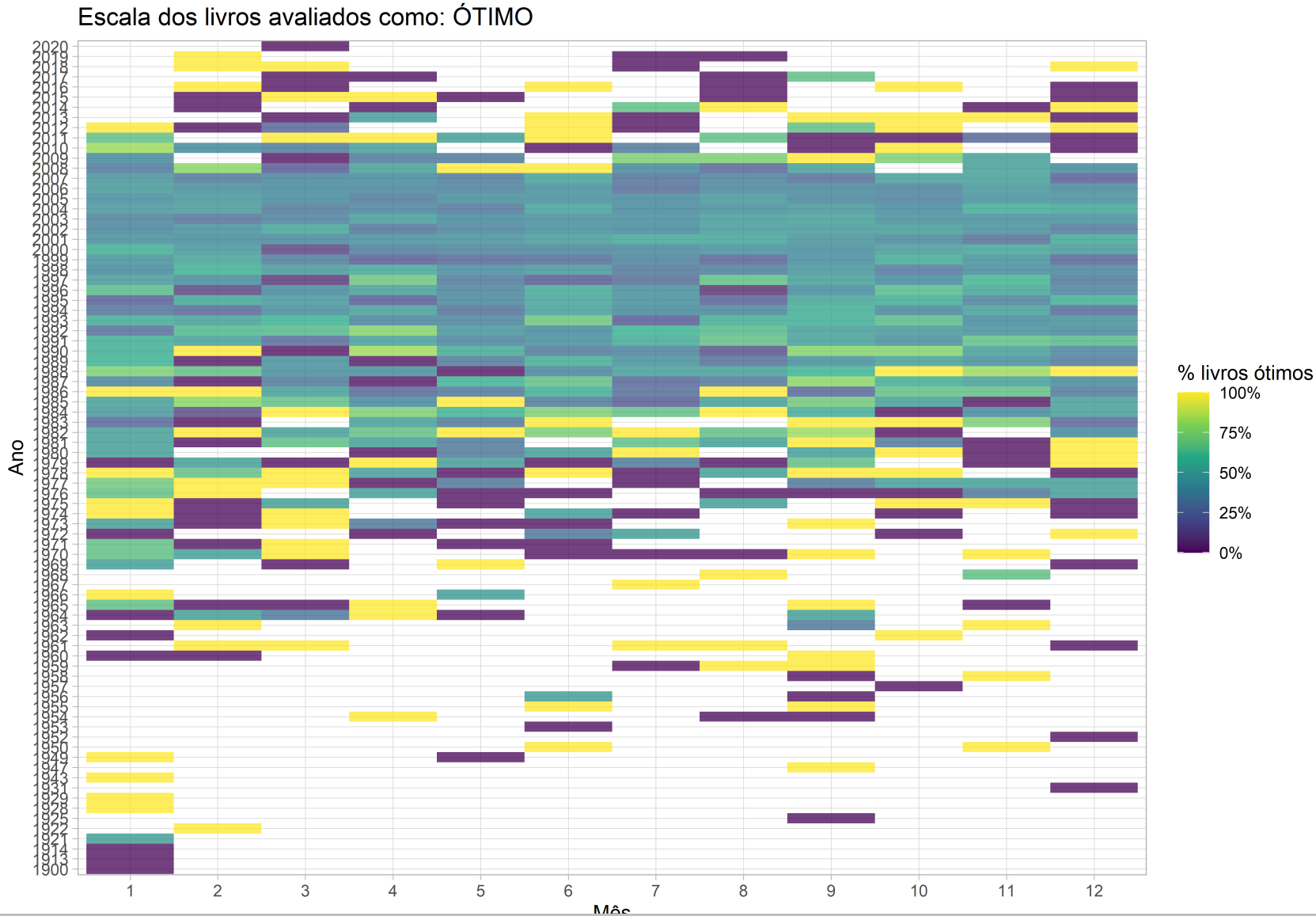
```
[1] -0.3605444
```

- A correlação entre `prop_text_reviewse` `ratings_count` não indicou multicolinearidade, então prosseguimos com essa variável como preditora.

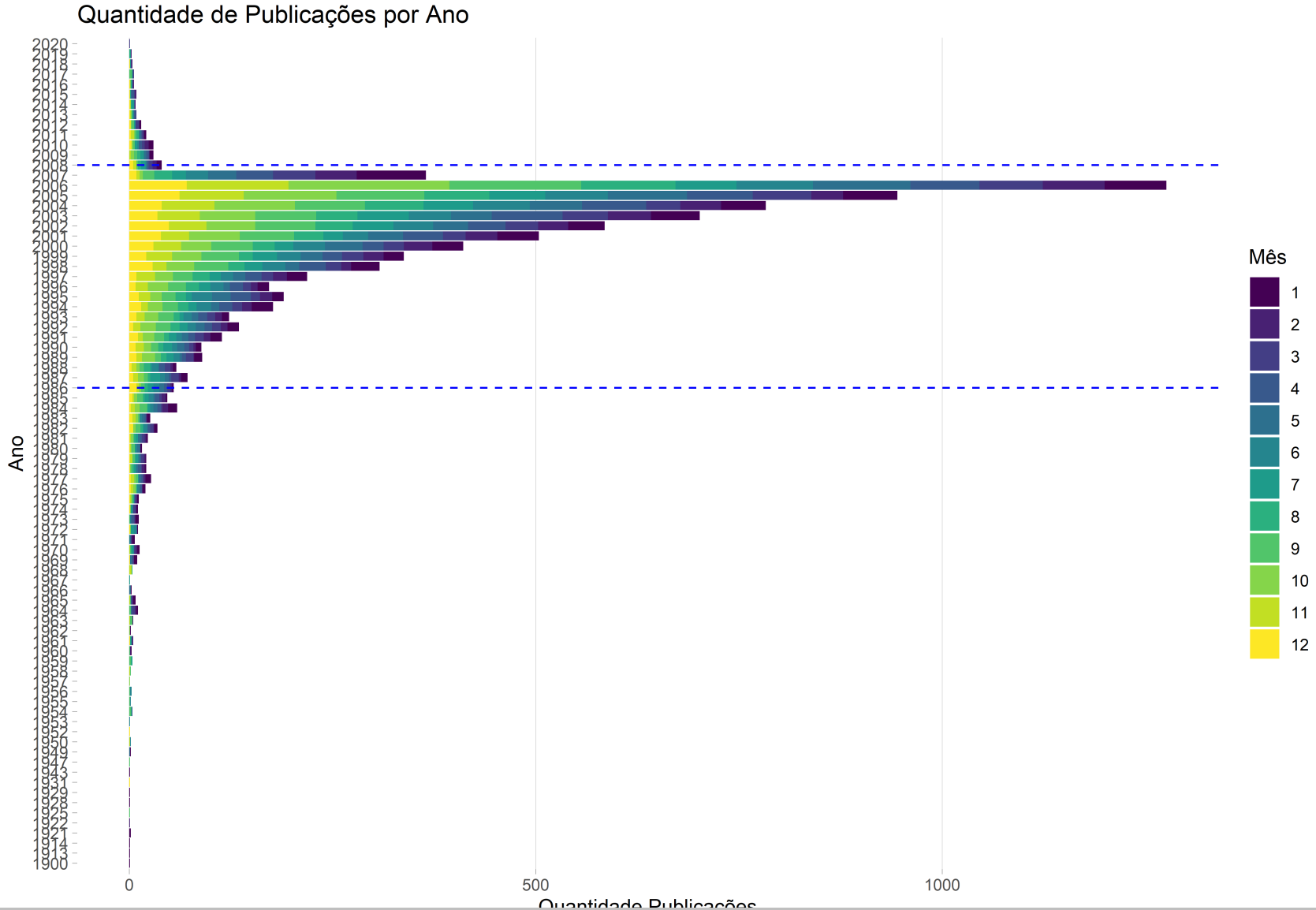
Continuando a Análise Exploratória e Descritiva



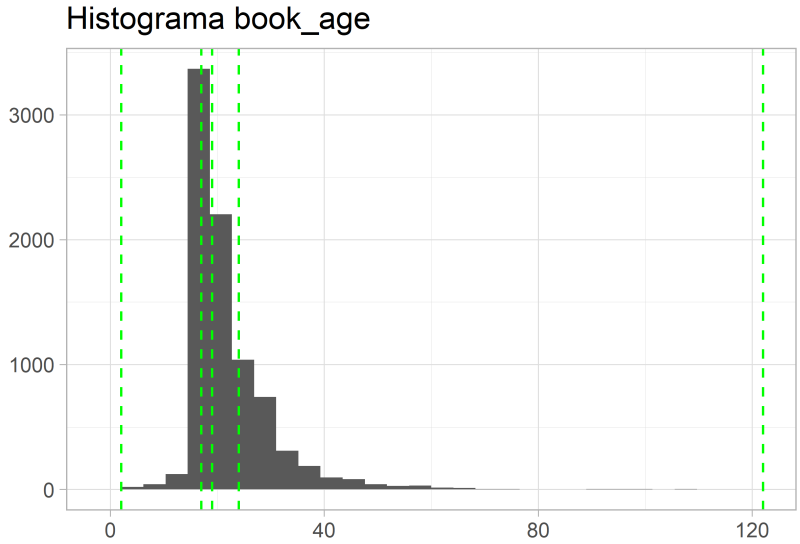
Continuando a Análise Exploratória e Descritiva

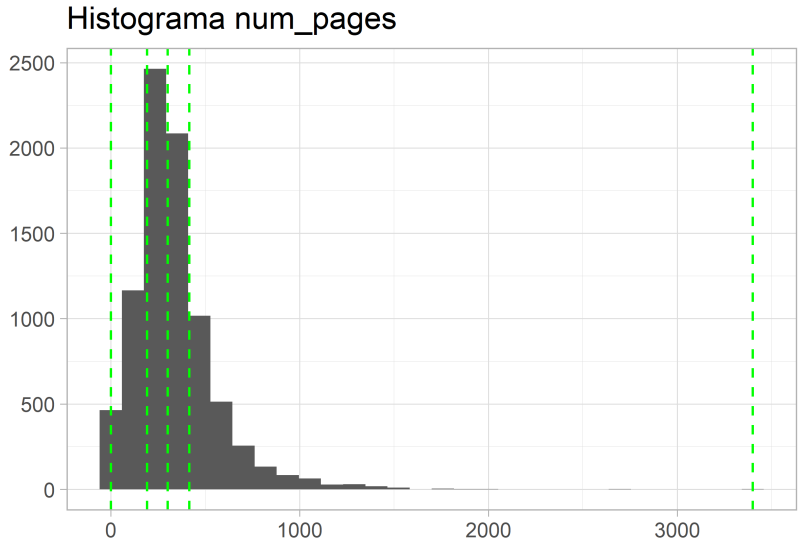


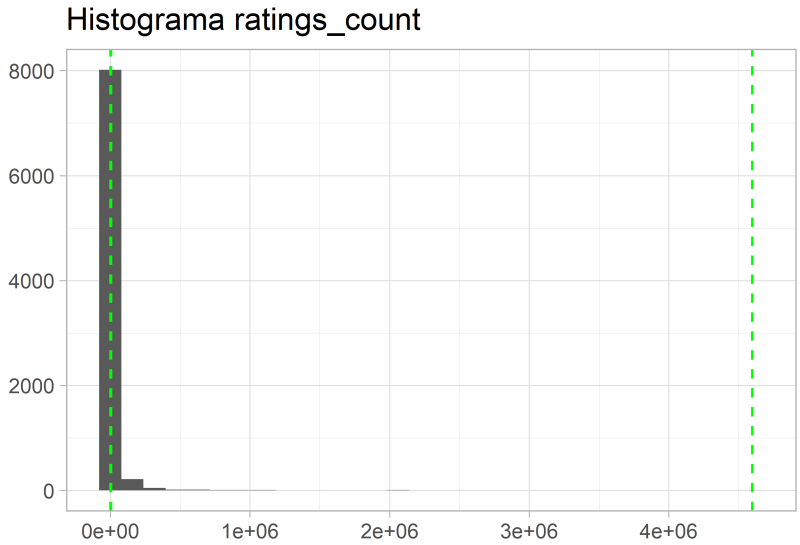
Mais gráficos =)

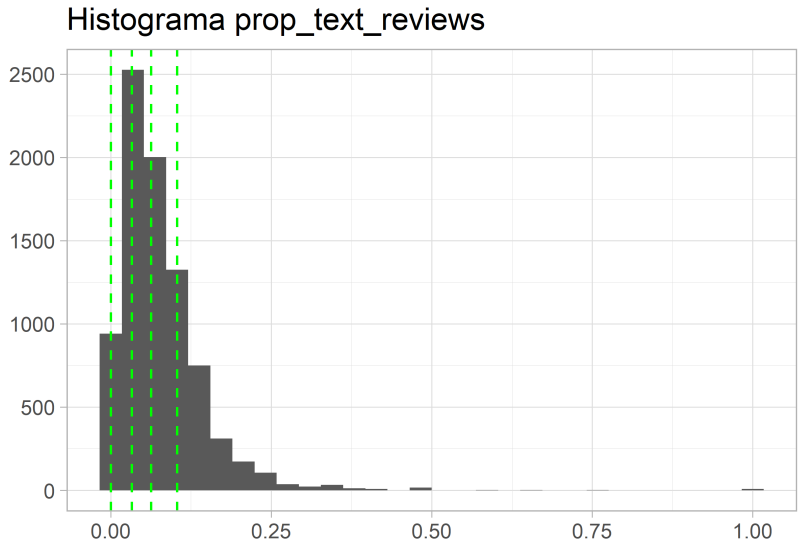


Definindo Filtros

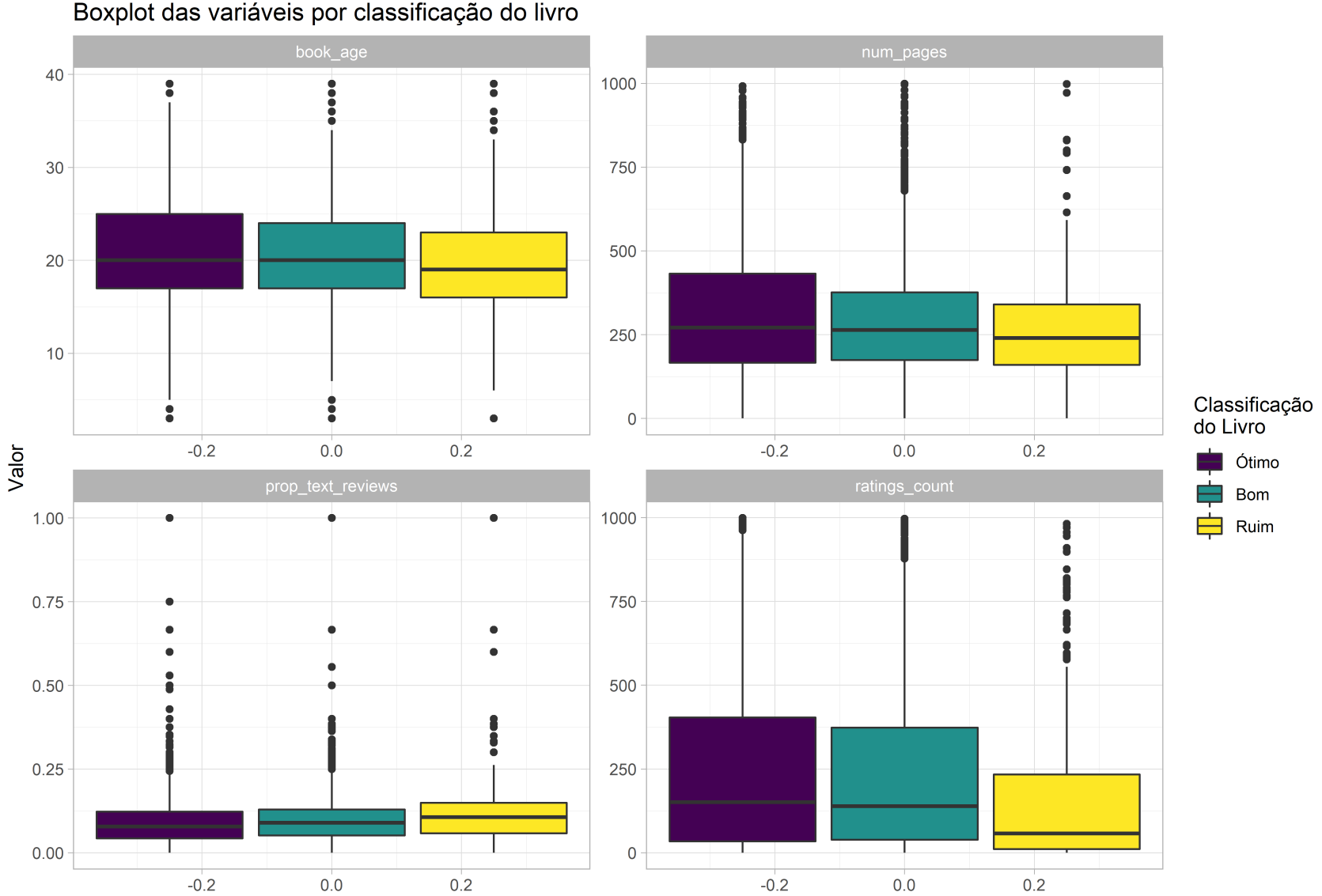








Novo boxplot com filtros aplicados



A modelagem