ME115B - Linguagem R

Projeto Final - 1S2022

Décio Miranda Filho, RA: 236087
Felipe Scalabrin Dosso, RA: 236110
Larissa Fazolin, RA: 217395
Nathan Augusto Elias, RA: 236258

Top 50 dos livros mais vendidos da Amazon, de 2009 até 2022

Introdução

Este relatório consiste na análise dos 50 livros mais vendidos da Amazon por ano, durante o período de 2009 até 2022. Com o intuito de entender melhor o sucesso de vendas dos livros em questão, foram levantadas algumas perguntas de interesse, procurando encontrar padrões. Entre elas:

Há relação entre o custo a pagar e a popularidade dos livros? Como a média de preços foi mudando com o decorrer dos anos? No geral, qual gênero de livro é mais caro? Qual gênero é mais popular? Quais autores tiveram o maior número de bestsellers, segundo esses Tops 50? Dentre eles, quantos são homens e quantas são mulheres? Há palavras predominantes nos títulos desses livros? Como a quantidade de reviews e a média de avaliação dos usuários influenciam nesses diferentes aspectos?

Sendo assim, este relatório apresenta análises com foco individual em cada diferente aspecto dos livros em questão - isto é, nas variáveis do conjunto de dados -, além de estudos sobre a relação entre essas informações.

Banco de Dados

O conjunto de dados que aqui será denotado como `bestsellers` pode ser encontrado em [Kaggle: Amazon Top 50 Bestselling Books 2009 - 2022]. Parte destes dados foram coletados por Chris Kachmar e são uma atualização do conjunto obtido por Sooter Saalu (presente em [Kaggle: Amazon Top 50 Bestselling Books 2009 - 2019]), pela técnica de web scrapping.

Sendo do tipo *tibble*, `bestsellers` contém 700 observações e 7 variáveis, cujos nomes são Name, Author, User Rating, Reviews, Price, Year e Genre. A seguir, o que cada coluna representa:

- `Name`: < character >, título do livro;
- `Author`: < character >, nome do autor do livro;
- `User_Rating`: < numeric >, avaliação média do livro pelos usuários da Amazon;
- `Reviews`: < numeric >, quantidade de reviews do livro na Amazon;
- `Price`: < numeric >, preço do livro na Amazon, em dólares;
- `Year`: < numeric >, ano em que o livro esteve no Top 50 dos livros mais vendidos da Amazon;
- `Genre`: < character >, gênero do livro.

Análise Exploratória e/ou Descritiva

(A) Etapa focada nas variáveis categóricas (livros, títulos e autores)

Para dar início a essa análise, foi elaborada a Tabela 1, na qual resumem-se em ordem decrescente de reviews os livros que apresentaram média máxima de avaliação pelos usuários, com suas devidas informações a respeito de seus autores, dos anos em que foram bestsellers, quantidade de reviews, preço e gênero. Em primeira análise, é possível observar que muitos desses livros apareceram mais de uma vez no Top 50, como, por exemplo, "I Love You to the Moon and Back" e "The Very Hungry Caterpillar", de Amelia Hepworth e Eric Carle, respectivamente. Nota-se também que livros referentes à cultura pop estão presentes nesta lista, como "The Deep End (Diary of a Wimpy Kid Book 15)" (conhecido como "O Diário de um Banana", em português), de Jeff Kinney. Em segunda análise, vê-se que houve uma predominância de livros do gênero ficção contra apenas três de não ficção, além de preços que variam de 4 dólares a 17, com a média sendo de 7.4 dólares. É interessante notar que nesta lista há livros desde de 2012 até 2022, com moda no ano de 2020 (totalizando oito de vinte ocorrências).

Tabela 1: Os 20 livros com maior nº de reviews e avaliação máxima.

Name	Author	Reviews	Price	Year	Genre
A Promised Land	Barack Obama	121109	16	2020	Non Fiction
I Love You to the Moon and Back	Amelia Hepworth	51188	4	2020	Fiction
I Love You to the Moon and Back	Amelia Hepworth	51188	4	2021	Fiction
I Love You to the Moon and Back	Amelia Hepworth	51188	4	2022	Fiction
The Very Hungry Caterpillar	Eric Carle	47260	5	2020	Fiction
The Very Hungry Caterpillar	Eric Carle	47260	5	2021	Fiction
The Very Hungry Caterpillar	Eric Carle	47260	5	2022	Fiction
Dog Man: Grime and Punishment: A Graphic Novel (Dog Man #9): From the Creator of Captain Underpants (9)	Dav Pilkey	41021	6	2020	Fiction
Brown Bear, Brown Bear, What Do You See?	Bill Martin Jr.	38969	5	2020	Fiction
Brown Bear, Brown Bear, What Do You See?	Bill Martin Jr.	38969	5	2021	Fiction
Brown Bear, Brown Bear, What Do You See?	Bill Martin Jr.	38969	5	2022	Fiction
The Deep End (Diary of a Wimpy Kid Book 15)	Jeff Kinney	38674	7	2020	Fiction
Oh, the Places You'll Go!	Dr. Seuss	35287	9	2021	Fiction
Chicka Chicka Boom Boom (Board Book)	Bill Martin Jr.	30145	5	2020	Fiction
American Marxism	Mark R. Levin	29510	14	2021	Non Fiction
Magnolia Table, Volume 2: A Collection of Recipes for Gathering	Joanna Gaines	24352	17	2020	Non Fiction
Oh, the Places You'll Go!	Dr. Seuss	21834	8	2012	Fiction
Oh, the Places You'll Go!	Dr. Seuss	21834	8	2013	Fiction
Oh, the Places You'll Go!	Dr. Seuss	21834	8	2014	Fiction
Oh, the Places You'll Go!	Dr. Seuss	21834	8	2015	Fiction

Em seguida, foram questionados possíveis padrões nos títulos dos livros presentes no conjunto de dados aqui chamado de `bestsellers`. Foram reunidas todas as palavras e suas frequências, conforme a Figura 1. Vale ressaltar que palavras de algumas classes foram removidas (conjunções, pronomes, artigos, etc) pois, apesar de frequentes, não apresentam informações relevantes para essa análise. Esperava-se que as palavras mais utilizadas nos títulos tornassem explícitas algumas motivações populares dos leitores e, de fato, a alta ocorrência de palavras como *Love* (amor), *Life* (vida) e *Kids* (crianças) apoia essa hipótese. No entanto, ainda mais comum foram meta-palavras, i.e., palavras referentes ao livro em si, como *Book* (livro), *Novel* (romance), *Guide* (guia) e *Edition* (edição).

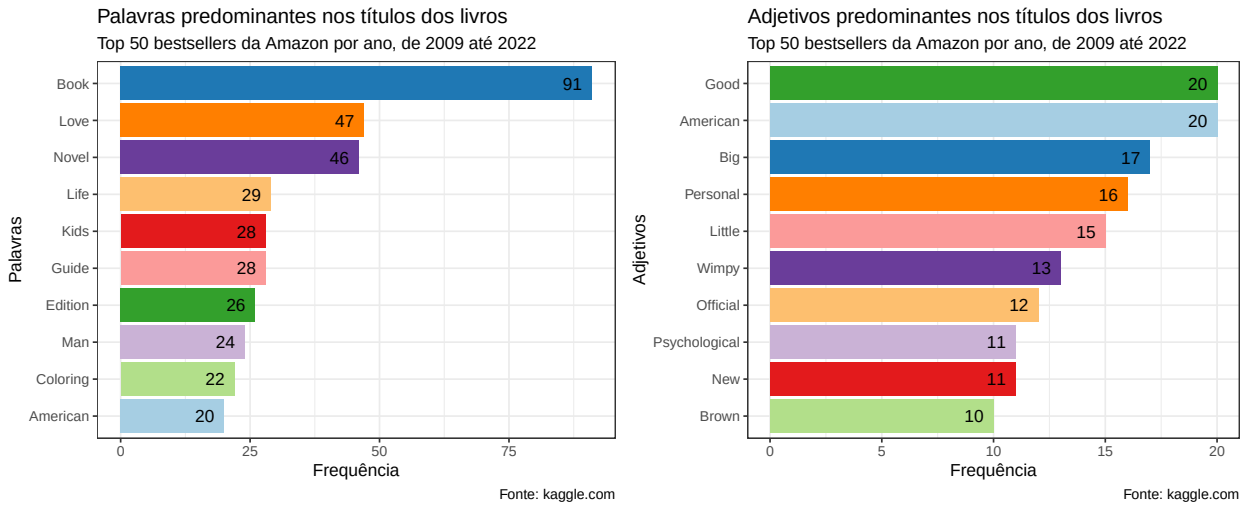

Figura 1: Gráfico de frequência das 10 palavras e adjetivos mais utilizados nos títulos dos bestsellers da Amazon, de 2009 até 2022.

Do lado direito da Figura 1 temos um gráfico semelhante, mas dessa vez limitado apenas aos adjetivos. Com variância de 13.611 para a distribuição das frequências dos adjetivos contra 457.656 das palavras sem restrição de classe, nota-se que não houve adjetivos muito mais frequentes que os outros, como foi o caso da palavra *Book* no gráfico da esquerda, que apareceu 91 vezes. Dado que a sede da Amazon se situa em Seattle, Washington, nos EUA e que um dos adjetivos com maior número de ocorrências trata-se de *American* (americano), é possível que haja uma concentração maior de usuários da plataforma na América do Norte.

Figura 2: Gráfico de barras com a proporção de sexo dos 20 autores com mais bestsellers na Amazon, de 2009 até 2022.

Para o último item desta etapa, o foco se deu nos criadores dessas obras. A partir de um levantamento dos 20 autores com mais bestsellers na Amazon, dentre o Top 50 de 2009 até 2022, foram computadas as porcentagens segundo o sexo dos escritores em questão, como mostra a Figura 2. Percebe-se que existe uma forte predominância de homens, totalizando 14 autores do sexo masculino contra 5 do sexo feminino. A categoria `Outro`, com 1 ocorrência, refere-se à American Psychological Association.

(B) Etapa focada nas variáveis numéricas (avaliações, reviews e preços)

Pela Figura 3, nota-se que a distribuição das avaliações médias dos usuários para os bestsellers possui uma assimetria à esquerda. Ou seja, há uma concentração maior de livros com rating elevado, o que era esperado dado que tratam-se dos livros com maior sucesso de vendas na plataforma da Amazon. Ademais, com média 4.64 (linha tracejada vermelha) e mediana 4.7 (linha tracejada azul), percebe-se que, mesmo com a presença de alguns outliers com avaliação abaixo de 4.0, a diferença entre essas medidas centrais ainda foi pequena. Quanto às medidas de dispersão, as avaliações dos usuários tem variância de 0.048, desvio padrão de 0.219 e amplitude de 1.6, com a nota mínima sendo 3.3 e a máxima 4.9. Sob a ótica de gênero dos livros, ambas as distribuições são bem próximas.

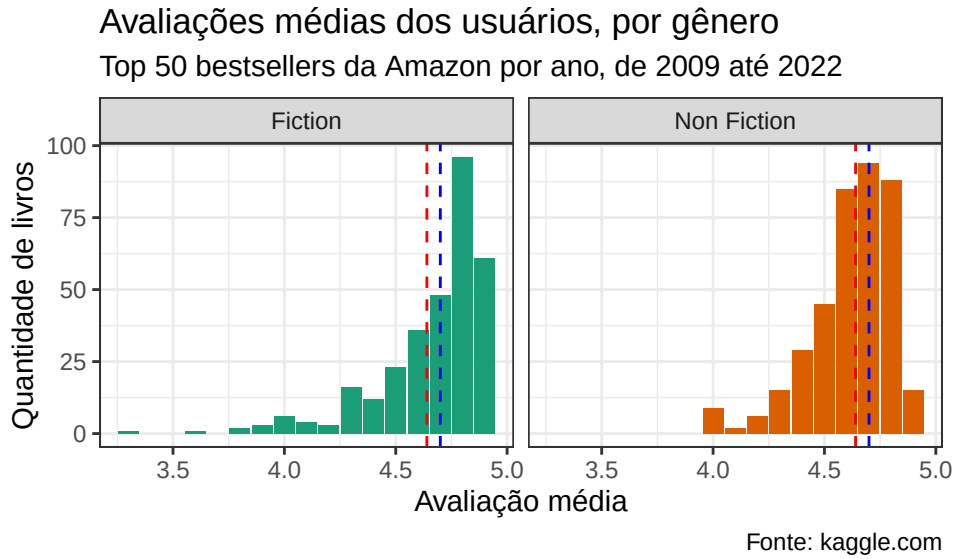

Figura 3: Gráfico de barras das avaliações médias dos usuários, por gênero de livro dos bestsellers da Amazon, de 2009 até 2022 (com linha tracejada vermelha e azul representando a média e a mediana, respectivamente).

Além disso, segue na Figura 4 as principais informações referentes aos quartis da distribuição das avaliações dos usuários. Vê-se que em quase todos os anos a mediana dos livros de ficção foi superior aos de não ficção, exceto em 2022. Visualmente nota-se também um ligeiro aumento geral nas medianas e decréscimo na dispersão das notas a partir da segunda metade do período computado. Houve presença de outliers em múltiplos anos.

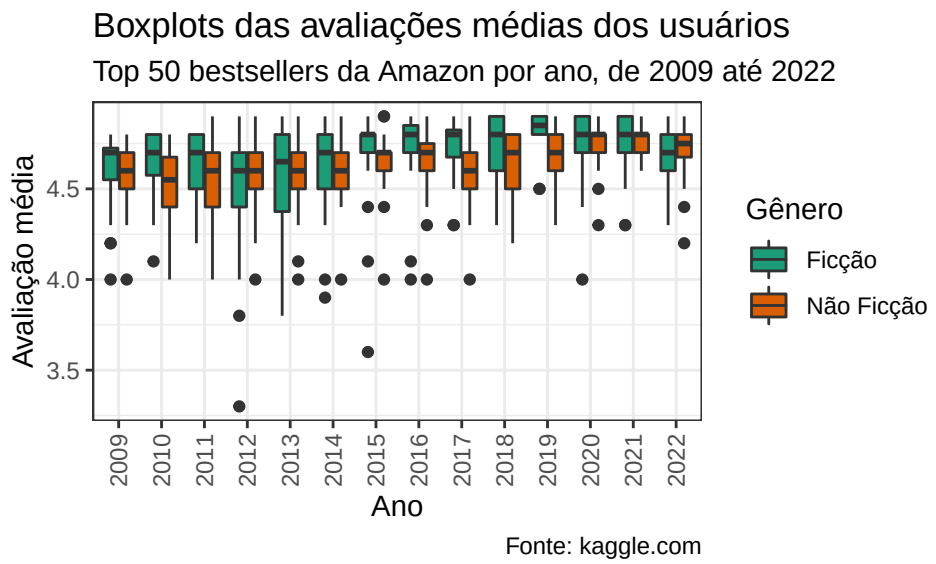

Figura 4: Boxplots das avaliações médias dos usuários, por gênero de livro e ano dos bestsellers da Amazon, de 2009 até 2022.

Quanto ao estudo da quantidade de reviews por gênero, tem-se que o primeiro quartil, a mediana e o terceiro quartil para os livros de ficção foram, respectivamente, de 7122, 1.444×10^4 e 2.758×10^4; já para os de não ficção, foram de 3490, 7910.5 e 1.958×10^4. Essas informações estão melhor apresentadas na Figura 5, onde nota-se também uma grande quantidade de outliers. Para o gênero de ficção, identifica-se uma grande concentração de outliers acima de 50 mil reviews, enquanto que para o outro gênero essa proporção é ligeiramente menor.

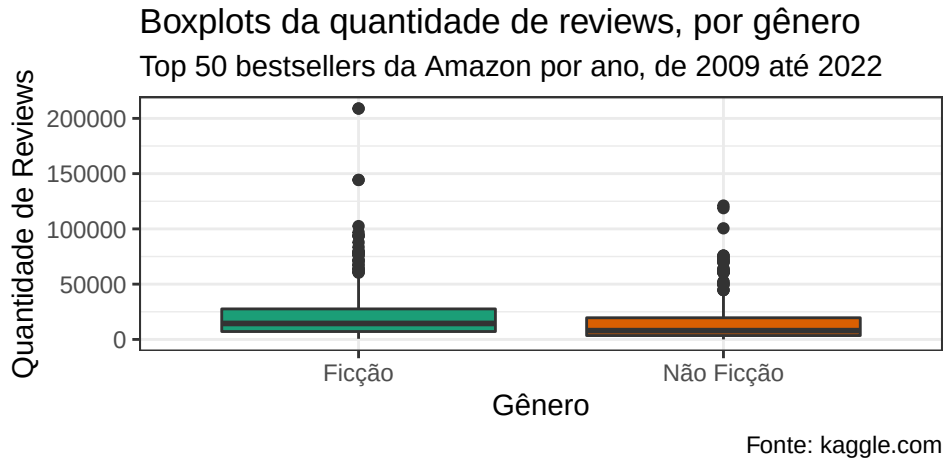

Figura 5: Boxplots da quantidade de reviews, por gênero de livro dos bestsellers da Amazon, de 2009 até 2022.

Agora, de modo a estudar a variação na média dos preços conforme o passar dos anos, foi plotado o gráfico que se vê na Figura 6, por gênero de livro. A partir dele, percebe-se que os preços variaram intensamente dentro do espaço de 13 anos, especialmente no período de 2009 até 2014. Logo após, identifica-se uma aproximação entre a média de preços de ambos os gêneros, além de uma redução geral. Além disso, vale destacar que durante quase toda a série histórica aqui estudada, a média nos preços dos livros de não ficção

foi maior do que a dos de ficção.

A hipótese inicial do grupo era de que a média dos preços aumentasse de maneira quase linear conforme se aproximasse do ano atual (2022); no entanto, os dados apontam uma média geral menor na segunda metade do período aqui tratado. Tendo a Amazon uma plataforma digital, é possível que isso tenha ocorrido devido à difusão e ao aumento da acessibilidade de eBooks, isto é, livros digitais que costumam custar mais barato que suas versões físicas.

Sem separação de gêneros, a média geral dos preços foi de 12.7, com mediana de 11, variância de 98.31, desvio de 9.915 e amplitude de 105. Os anos com as maiores médias foram 2013 e 2014.

Figura 6: Gráfico de linhas da variação na média dos preços ao passar dos anos, por gênero de livro dos bestsellers da Amazon, de 2009 até 2022.

Por fim, foram realizados alguns estudos quanto à relação entre essas diferentes variáveis. Pela Figura 7, nota-se uma tendência à diminuição na quantidade de reviews para livros de preço maior. Apesar de todos os livros neste conjunto de dados serem um sucesso de vendas, é possível que, por alguns deles apresentarem um custo maior, o público de leitores com poder aquisitivo para adquirí-los seja mais limitado e, por consequência, menos pessoas são capazes de opinar sobre os conteúdos dos livros em questão através de uma review. No entanto, com correlação de -0.115 entre essas variáveis, sua relação de dependência é fraca.

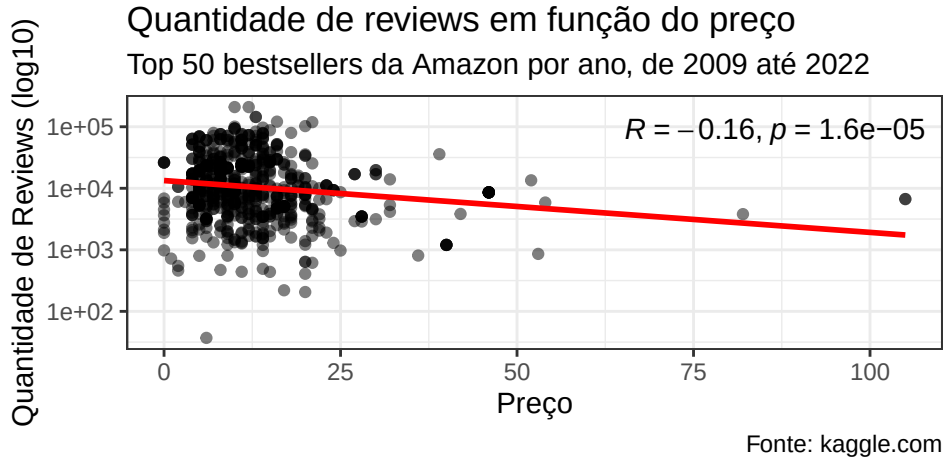

Figura 7: Gráfico de relação entre a quantidade de reviews e o preço dos livros bestsellers da Amazon, de 2009 até 2022.

Na Figura 8, por outro lado, é estudada a avaliação média dos usuários em função da quantidade de reviews dos bestsellers. Observa-se que há uma maior concentração de pontos em torno da nota 4.5, justificada pela popularidade dos livros neste conjunto de dados. Contudo, pelas retas de modelo linear dispostas nesse gráfico, é notável um sensível decréscimo na quantidade de reviews conforme aumento da avaliação para o gênero de ficção (ou seja, não necessariamente os livros com maior quantidade de reviews são aqueles com as melhores avaliações) e o fenômeno oposto para o de não ficção.

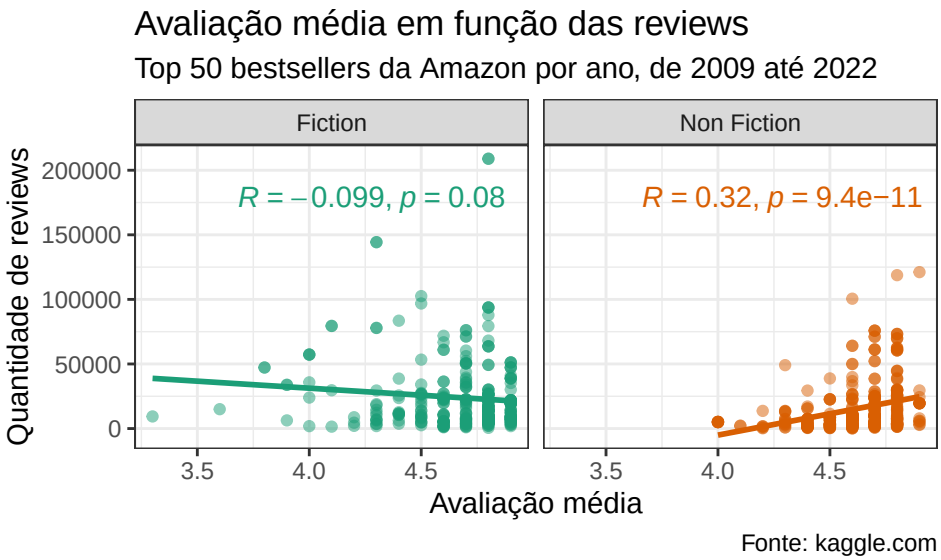

Figura 8: Gráfico de relação entre a avaliação média dos usuários e a quantidade de reviews, por gênero dos bestsellers da Amazon, de 2009 até 2022.

Considerações Finais

Em síntese, observou-se que muitos livros e autores apareceram na lista dos 50 bestsellers em múltiplos anos. Dentre estes livros, grande parte deles era do gênero ficção e a média de preços destes livros foi de 7.4 dólares, o que parece um valor razoável. Quanto a estes autores mais frequentes, foi visto que a sua maioria era do sexo masculino.

Agora, em uma análise geral dos livros, notou-se que algumas palavras foram bem mais frequentes que outras nos títulos, como é o caso de *Book* (livro) e *Love* (amor). Outras palavras frequentes, como *Kids* (crianças) e *American* (americano) podem refletir algo sobre as motivações e a identidade do público mais comum de leitores.

Quanto às variáveis numéricas, notou-se que as avaliações dos usuários foram, em média, mais altas para os livros do gênero ficção. No geral, a quantidade de reviews e as avaliações dos usuários não apresentaram indícios fortes de relação, assim como a quantidade de reviews e o preço. Percebeu-se, no entanto, que os livros de não ficção costumam ser mais caros e que a média dos preços foi menor na segunda metade dos anos analisados, podendo ser um indicativo da maior acessibilidade e popularidade de eBooks na plataforma digital da Amazon.

Bibliografia

KACHMAR, Chris. **Amazon Top 50 Bestselling Books 2009 - 2022**. Disponível em https://www.kaggle.com/datasets/chriskachmar/amazon-top-50-bestselling-books-2009-2022?select=bestsellers_with_categories_2022_03_27.csv. Acesso em: 11 jul. 2022.

WICKHAM, Hadley et al. **ggplot2: elegant graphics for data analysis**. 3. ed. Disponível em https://ggplot2-book.org/index.html. Acesso em: 11 jul. 2022.