

Democratizing Data Science in Portuguese: A Community-Led Translation of "R for Data Science" (2nd Edition)

Cesar A. Galvao, Beatriz Milz, Riva Quiroga, Ariana Cabral, Alberson da Silva Miranda, Marcus Antonio Cardoso Ramalho, Maximilian Canez Fernandes, José Carlos Barbosa, Bruno Miotto, Halian Vilela, Eric Scopinho, Ana Carolina Moreno, Luana Antunes Alexandre, Magno Tairone, Debora Silva, Arthur de Carvalho e Silva, Ían Muliterno, Carolina Musso, Jeanne Franco

Keywords: teaching, learning, book, volunteer, data science

Abstract

We are excited to announce the translation of the second edition of "R for Data Science", by Hadley Wickham, Mine Çetinkaya-Rundel, and Garrett Golemund into Brazilian Portuguese. This project was a community-driven effort, carried out by volunteers from the R community in Brazil and Latin America. The translated book aims to facilitate access to data science for Portuguese-speaking practitioners, students, and enthusiasts, fostering inclusion and outreach of this valuable resource.

Introduction

The R community is rich in resources, providing numerous free learning materials related to R and data science in English, including blogs, tutorials, books, and videos. Brazil is the largest country in Latin America with a population of over 215 million (2022) and yet faces a significant language barrier. The "EF English Proficiency Index EPI 2023 - Ranking of Countries and Regions"¹, rated Brazil as a country with low proficiency in English, with most of the population not speaking English at an intermediate level or higher. Translating these materials into Brazilian Portuguese can substantially expand access to technical content, democratize knowledge dissemination, enhance collaboration among diverse groups, and provide opportunities to leverage advancements in data science.

R for Data Science (1e)

"R for Data Science" (<https://r4ds.had.co.nz>), by Wickham, Çetinkaya-Rundel, and Golemund, is an excellent resource on data science with R and the tidyverse. It is freely

¹ Available at:

<https://www.ef.com/assetscdn/WIBIwq6RdJvcD9bc8RMd/cefcom-epi-site/reports/2023/ef-epi-2023-english.pdf>

DEMOCRATIZING DATA SCIENCE IN PORTUGUESE: A COMMUNITY-LED TRANSLATION OF "R FOR DATA SCIENCE" (2ND EDITION)

available online and can also be purchased as a physical copy. The first edition of the book was translated into Spanish (available at: <https://es.r4ds.hadley.nz/>) by volunteers from the Latin American community, making it freely available online in Spanish. However, the first edition was translated into Portuguese by another publisher, and it was not available for free online. This created a financial barrier to accessing content that was freely available in English.

Package {dados}

In 2020 during the COVID-19 pandemic, the Latin American community began drafting the first translation project into Portuguese: the package `{dados}`. This is a sibling package of `{datos}`, which translates datasets into Spanish. With the efforts of 10 volunteers, several datasets were translated into Brazilian Portuguese, including those used in the book "R for Data Science". This package has been available on CRAN since 2022.

R for Data Science (2e)

In 2021, the R-Ladies São Paulo started a book club to discuss the book "R for Data Science." While attending the activities, one of the organizers noticed that a second edition of the book was being developed. This led to a conversation with the authors in May 2022 about the possibility of translating the second edition into Portuguese, but this time making it freely available online.

Translating the book

In September 2023, the volunteer team gained permission from the book's authors to start the translation. To translate the book into Portuguese, the team used the initiative to translate the book into Spanish as a starting point. Each chapter had one translator and two reviewers. More than 20 volunteers were involved in the translation and review process.

The main platform used in the translation was GitHub. Essential GitHub features such as Issues, Pull Requests, and Projects were crucial for managing the translation process. We created an Issue to discuss each chapter to be translated, and these Issues were linked to a Project Board to facilitate progress tracking. The use of GitHub also added value to the contributors while applying industry-standard software development concepts (i.e. versioning control and Kanban). The GitHub Project for the translation is available at: <https://github.com/orgs/cienciadedatos/projects/2/views/1>.

Translation guide

The team developed a [translation guide](#) that contains basic style and framework information, as well as tutorials on how to contribute using GitHub, further democratizing the process since not all the volunteers had previous experience with it. This guide is continually being improved, and we hope it can be used in future translation projects.

Translation status

As of the end of July 2024, all the chapters of the book have been translated, but a few are still under review. The chapters that have already been translated and reviewed are available at: <https://cienciadedatos.github.io/pt-r4ds/>. We aim to conclude the reviews before the conference in November 2024.