
INTRODUCTION
TO
DATA SCIENCE
FOR SOCIAL
AND POLICY
RESEARCH
Collecting and Organizing Data
with R and Python
JOSE MANUEL MAGALLANES REYES

Data science has now firmly moved from computer science and engineering to the disciplines of the social sciences, where scholars are harnessing the insightful power of ever larger and more complex data sets. This volume provides a clear introduction for social scientists and policy researchers into the use of R and Python, including best practice of working with data files, command files, and outputs. The step-by-step approach with real-world examples will be of great value to students, scholars, and practitioners engaged in data analytic approaches to social problems."

– Todd Landman, Professor of Political Science and Pro Vice Chancellor, Faculty of Social Sciences, University of Nottingham

"The irruption of big data and the need to comply with high standards of research reproducibility require social scientists and policy analysts to be conversant in data collection and management techniques. Unfortunately, even those with sophisticated methodological training often lack the necessary tools to take on these requirements. Magallanes Reyes's book at long last collects and organizes a large amount of information and useful advice on how to curate data for scientific analysis. Through agile narrative and compelling examples, he walks the reader through the use of open-source tools of data science such as R, Python, and Github. The book is an invaluable resource for students and scholars at different levels of proficiency, from neophytes to advanced users."

– Guillermo Rosas, Washington University in St. Louis

"This new, practical, reader-friendly how-to manual on computational social data analysis is both long overdue and a must-have for analysts and researchers. The range of problem-solving strategies and demonstrations is impressive. While eminently practical, Magallanes Reyes's contribution is also rigorous and true to its scientific aims, which will please both basic and applied scientists and practitioners."

– Claudio Cioffi-Revilla, Professor of Computational Social Science and Director, Center for Social Complexity, George Mason University, and founding President, Computational Social Science Society of the Americas

"Magallanes Reyes's excellent book on data science for researchers and policy analysts is an accessible yet thorough introduction to data management and analyses in R and Python. It has a broad coverage of the techniques required to capture, clean, and process complex information. It is the perfect companion for sophisticated policy analysts and researchers that are ready to take advantage of the wealth of data that is available to skilled computer scientists."

– Ernesto Calvo, University of Maryland

CAMBRIDGE
UNIVERSITY PRESS
www.cambridge.org

ISBN 978-1-107-54025-5

Cover image courtesy of © Westhoff / E+ / Getty Images

Cover design by Holly Johnson

MAGALLANES REYES

INTRODUCTION TO DATA SCIENCE FOR SOCIAL AND POLICY RESEARCH

CAMBRIDGE

Introduction to Data Science for Social and Policy Research

Collecting and Organizing Data with R and Python

JOSÉ MANUEL MAGALLANES REYES
University of Washington and Pontificia Universidad Católica del Perú

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

4843/24, 2nd Floor, Ansari Road, Daryaganj, Delhi - 110002, India

79 Anson Road, #06-04/06, Singapore 079906

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781107117419
DOI: 10.1017/9781316338599

First published 2017

Printed in the United Kingdom by Clays, St Ives plc

A catalogue record for this publication is available from the British Library.

Library of Congress Cataloging-in-Publication Data
Names: Magallanes Reyes, Jose Manuel, author.
Title: Introducing data science to social and policy analysts : from collecting to organizing data with R and Python / Jose Manuel Magallanes Reyes, Pontificia Universidad Católica del Perú.
Description: Cambridge, United Kingdom ; New York, NY : Cambridge University Press, 2017. | Includes bibliographical references and index.
Identifiers: LCCN 2017020965 | ISBN 9781107117419 (alk. paper)
Subjects: LCSH: Policy sciences – Statistical methods. | Policy sciences – Data processing. | Python (Computer program language) | R (Computer program language)
Classification: LCC H97.M336 2017 | DDC 300.72/7 – dc23
LC record available at https://lccn.loc.gov/2017020965

ISBN 978-1-107-11741-9 Hardback
ISBN 978-1-107-54025-5 Paperback

Introduction to Data Science for Social and Policy Research

Real-world data sets are messy and complicated. Written for students in social science and public management, this authoritative but approachable guide describes all the tools needed to collect data and prepare it for analysis. Offering detailed, step-by-step instructions, it covers collection of many different types of data including web files, APIs, and maps; data cleaning; data formatting; the integration of different sources into a comprehensive data set; and storage using third-party tools to facilitate access and shareability (from Google Docs to GitHub).

Assuming no prior knowledge of R and Python, the author introduces programming concepts gradually, using real data sets that provide the reader with practical, functional experience.

JOSE MANUEL MAGALLANES REYES is an associate professor of political science and public policy at Pontificia Universidad Católica del Perú, senior data science fellow at the eScience Institute, and visiting professor at the Evans School of Public Policy and Governance at the University of Washington, Seattle. His research focuses on social complexity, applying computational thinking to governance issues to inform public policy. Over the past 15 years he has served in government and has been involved in several initiatives in Peru's public sector to make better use of data for policy, political research, and decision making.

"It is rare indeed to pick up a new manuscript and immediately think how much you wish it had been written five years earlier, but I suspect many people will have that reaction to this book. This timely, thorough, and remarkably clear tutorial to both R and Python serves as a much needed on-ramp to the data part of data science, and will undoubtedly soon grace the bookshelves of many social scientists – both students and their instructors. If you are intrigued by the possibilities of data science but concerned about the start-up costs, look no further: help has arrived."

Joshua Tucker, New York University

"If you need to develop new skills in R and Python but you don't know where to start, this is the book for you. With simple language, Magallanes Reyes shows you how to install the programs, retrieve data using APIs and scrape Internet sources, and how to get the data ready for modeling. This book is a gem."

Aníbal Pérez-Liñán, University of Pittsburgh

Contents

Illustrations

Tables

PART ONE

GETTING STARTED