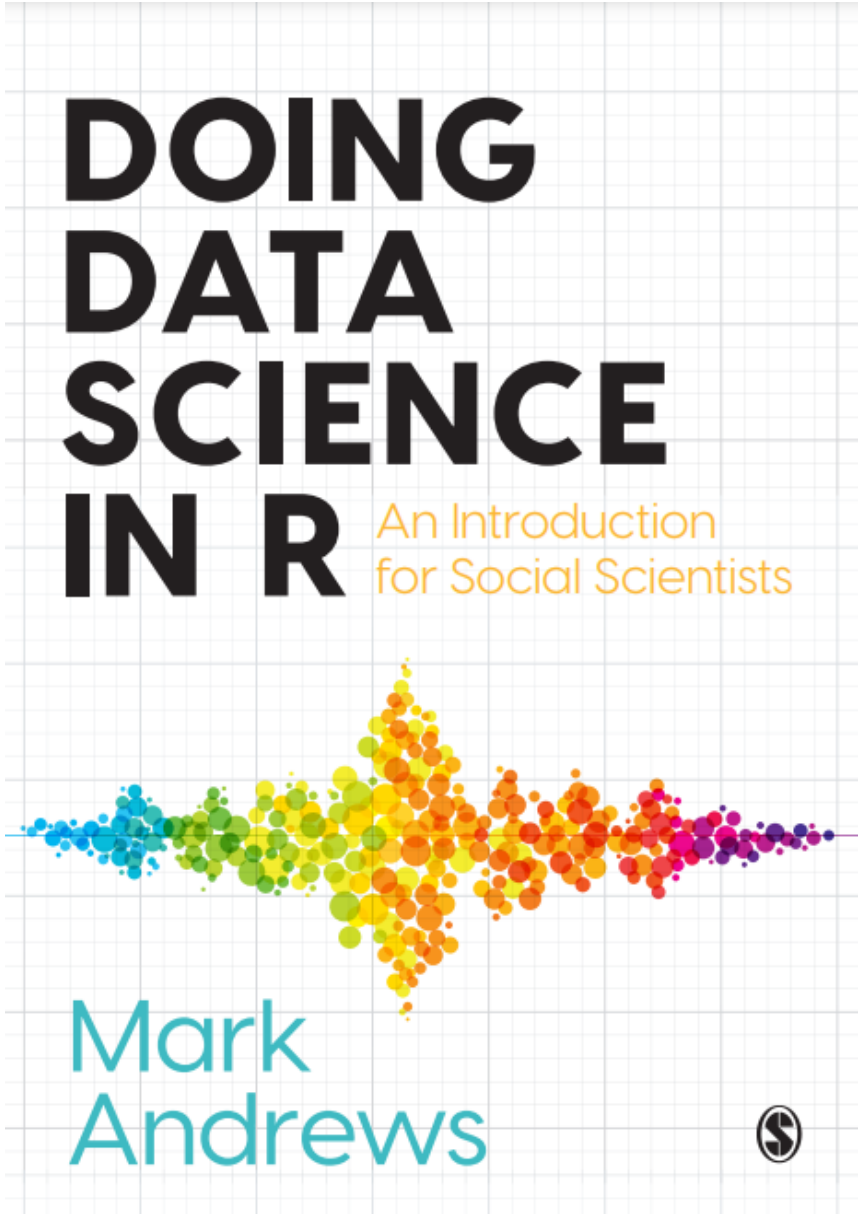

Doing Data Science in R: An Introduction for Social Scientists

Chapter 1

Data Analysis And Data Science

Introduction

This book is about statistical data analysis of real-world data using modern tools.

It is aimed at those who are engaged in analysis of statistical data of the kind that might arise at or beyond PhD level scientific research

Analysing this data almost always requires data wrangling, exploration, and visualization. It involves modelling the data using flexible probabilistic models. This book aims to address all of these topics.

Introduction

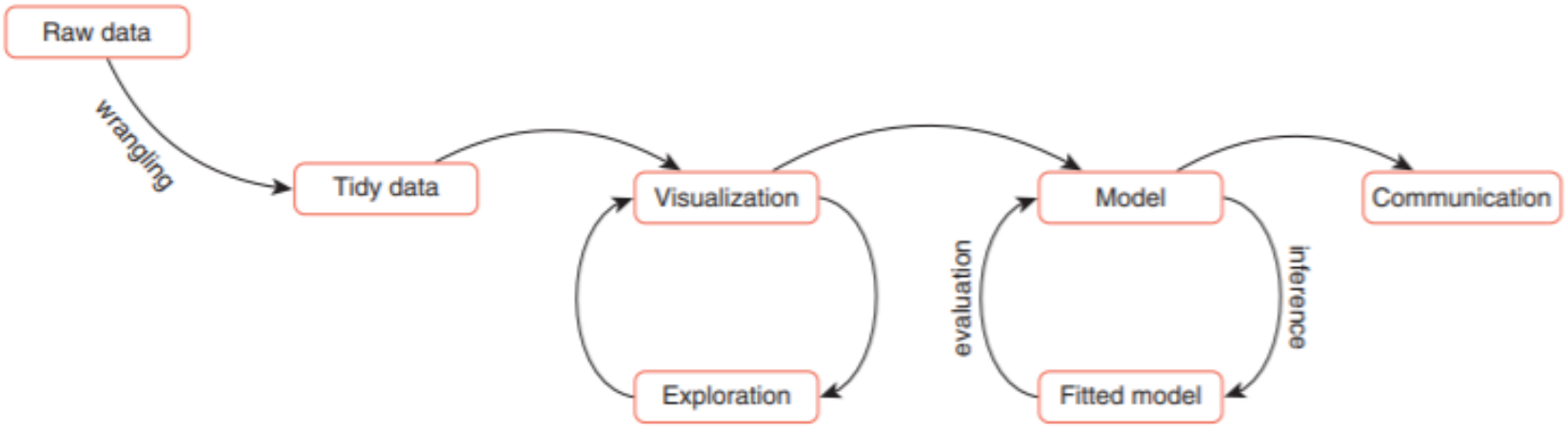

Figure 1.1 The data science workflow

Introduction

Data science is a set of interrelated computational or mathematical methods and tools that are used in the general data analysis workflow

The process of transforming the data so that it amenable to further analysis is data wrangling, and the resulting data sets are said to be tidy.

The exploratory analysis stage then leads us to posit a tentative probabilistic model of the data.

Introduction

Each of the stages of this data science workflow involves computational and mathematical concepts and methods

In addition to computing tools, many of the stages of the data science workflow involve mathematical and statistical concepts and methods. T

Being aware of statistical modelling as a flexible and systematic framework that is based on pragmatic and theoretical principles allows us to more competently and confidently perform statistical analysis

What is data science?

This general point about real-world data analysis being more than just the traditional focus of mathematical statistics was actually made decades ago by Tukey (1962).

Modern data analysis has a character that goes beyond Tukey's vision, however broad and comprehensive it was. This is due to the computing revolution.

What is data science?

In summary, in this book, we use the term ‘data science’ as the general term for modern data analysis, which is something that always involves a tight integration of computational and statistical methods and tools.

In this, we are hopefully faithfully following the broad and general understanding of what real-world data analysis entails as described by Tukey (1962), albeit with the additional vital feature of intensive use of computational tools

Why R, Not Python?

Given our conception of the data science workflow that we outlined in Figure 1.1, R is an inevitable choice.

Everything we cover in this book could be done using another programming language, or possibly using some set of different languages. Chief among these alternatives is Python

Who is this book for?

Those engaged in data analysis in scientific research, specifically research at or beyond PhD level

This book is heavily focused on using data to build and interpret statistical or probabilistic models of the scientific phenomenon being studied