Exploratory Data Analysis in the 21st Century

Di Cook and Emi Tanaka

2021-06-14

Contents

Chapter 1

Preface

Exploratory data analysis is about "playing in the sand with your data" to allow you to find the unexpected or at least get an understanding about the data that you have in hand. You can think about it as like travelling to a new place. You might have a purpose for visiting, perhaps to attend a conference, or family and friends. Some of your movements will be pre-determined, or guided by the advice of others, but hopefully you will spend some of the time you wandering around without guidance, perhaps even aimlessly. It is in these times you might find something special, a cafe in a garden with great carrot cake, a cuddling pair of rainbow lorikeets, a little library full of Jane Austen books, or even a cheap gas station. Walking around without an agenda helps you get to know the new neighbourhood.

The first book on exploratory data analysis was published by **?**. It has been the gold standard for learning about data analysis for many decades, although it has been quickly dated because all the techniques described can be accomplished with pencil and paper. Today, exploratory data analysis has come of age, and is a fundamental part of data science. Most of what we do in data analysis is conducted using the computer, not pencil and paper. Data sets are often quite large, too.

There are many books, and courses on exploratory data analysis. Virtually all of these are missing key ingredients of Tukey's spirit. Exploratory data analysis has become synonymous with descriptive statistics, and this is sad. The exploratory part of exploratory data analysis has been subsumed by humdrum data summary. The purpose in writing this book is to communicate the enjoyment of working with data, to reclaim the original intent, to "forces us to notice what we never expected to see" (**?**), with modern computational techniques.

Talk about teaching

All of the examples in this book are produced using the open source software R.

If you are new to R, a good place to start before reading this book is **?**.

Do they need to know some statistics?

About the exercises.

Chapter 2

What is exploratory data analysis?

Data analysis is a process of cleaning, transforming, inspecting and modelling data with the aim of extracting information.

Data analysis includes:

- exploratory data analysis,
- confirmatory data analysis, and
- initial data analysis.

In a confirmatory data analysis, the focus is on statistical inference and includes processes such as testing hypothesis, model selection, or predictive modelling.