Table of Contents

Introduction

I had two main reasons for writing this book. When I first started learning data science, I could not find a centralized overview of all the important topics on this subject. A practitioner of data science needs to be proficient in at least one programming language, learn the various aspects of data preparation and visualization, and also be conversant with various aspects of statistics. The goal of this book is to provide a consolidated resource that ties these interconnected disciplines together and introduces these topics to the learner in a graded manner. Secondly, I wanted to provide material to help readers appreciate the practical aspects of the seemingly abstract concepts in data science, and also help them to be able to retain what they have learned. There is a section on case studies to demonstrate how data analysis skills can be applied to make informed decisions to solve real-world challenges. One of the highlights of this book is the inclusion of practice questions and multiple-choice questions to help readers practice and apply whatever they have learned. Most readers read a book and then forget what they have read or learned, and the addition of these exercises will help readers avoid this pitfall.

The book helps readers learn three important topics from scratch - the Python programming language, data analysis, and statistics. It is a self-contained introduction for anybody looking to start their journey with data analysis using Python, as it focuses not just on theory and concepts but on practical applications and retention of concepts. This book is meant for anybody interested in learning Python and Python-based libraries like Pandas, Numpy, Scipy, and Matplotlib for descriptive data analysis, visualization, and statistics. The broad categories of skills that readers learn from this book include programming skills, analytical skills, and problem-solving skills.

The book is broadly divided into three parts - programming with Python, data analysis and visualization, and statistics. The first part of the book comprises three chapters. It starts with an introduction to Python - the syntax, functions, conditional statements, data types, and different types of containers. Subsequently, we deal with advanced concepts like regular expressions, handling of files, and solving mathematical problems

with Python. Python is covered in detail before moving on to data analysis to ensure that the readers are comfortable with the programming language before they learn how to use it for purposes of data analysis.

The second part of the book, comprising five chapters, covers the various aspects of descriptive data analysis, data wrangling and visualization, and the respective Python libraries used for each of these. There is an introductory chapter covering basic concepts and terminology in data analysis, and one chapter each on NumPy (the scientific computation library), Pandas (the data wrangling library), and the visualization libraries (Matplotlib and Seaborn). A separate chapter is devoted to case studies to help readers understand some real-world applications of data analysis. Among these case studies is one on air pollution, using data drawn from an air quality monitoring station in New Delhi, which has seen alarming levels of pollution in recent years. This case study examines the trends and patterns of major air pollutants like sulfur dioxide, nitrogen dioxide, and particulate matter for five years, and comes up with insights and recommendations that would help with designing mitigation strategies.

The third section of this book focuses on statistics, elucidating important principles in statistics that are relevant to data science. The topics covered include probability, Bayes theorem, permutations and combinations, hypothesis testing (ANOVA, chi-squared test, z-test, and t-test), and the use of various functions in the Scipy library to enable simplification of tedious calculations involved in statistics.

By the end of this book, the reader will be able to confidently write code in Python, use various Python libraries and functions for analyzing any dataset, and understand basic statistical concepts and tests. The code is presented in the form of Jupyter notebooks that can further be adapted and extended. Readers get the opportunity to test their understanding with a combination of multiple-choice and coding questions. They also get an idea about how to use the skills and knowledge they have learned to make evidence-based decisions for solving real-world problems with the help of case studies.