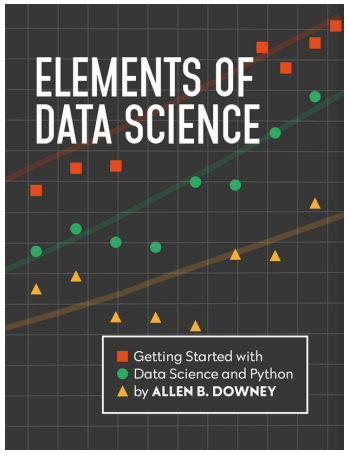
Elements of Data Science

Print to PDF

Contents



Printed copies of *Elements of Data Science* are available now, with a **full color interior** from Lulu.com.

Elements of Data Science is an introduction to data science for people with no programming experience. My goal is to present a small, powerful subset of Python that lets you do real work with data as quickly as possible.

I don't assume that the reader knows anything about programming, statistics, or data science. When I use a term, I try to define it immediately, and when I use a programming feature, I try to explain it.

This book is in the form of Jupyter notebooks. Jupyter is a software development tool you can run in a web browser, so you don't have to install any software. A Jupyter notebook is a document that contains text, Python code, and results. So you can read it like a book, but you can also modify the code, run it, develop new programs, and test them.

The notebooks contains exercises where you can practice what you learn. Most of the exercises are meant to be quick, but a few are more substantial.

This material is a work in progress, so suggestions are welcome. The best way to provide feedback is to click here and create an issue in this GitHub repository.

Case Studies

In addition to the notebooks below, the *Elements of Data Science* curriculum includes these case studies:

- Political Alignment Case Study: Using data from the General Social Survey, this case study explore changing opinions on a variety of topics among survey respondents in the United States. Readers choose one of about 120 survey questions and see how responses have changed over time and how these changes relate to political alignment (conservative, moderate, or liberal).
- Recidivism Case Study: This case study is based on a well known paper, "Machine Bias", which was published by Politico in 2016. It relates to COMPAS, a statistical tool used in the criminal justice system to assess the risk that a defendant will commit another crime if released. The ProPublica article concludes that COMPAS is unfair to Black defendants because they are more likely to be misclassified as high risk. A response article in the Washington Post suggests that "It's actually not that clear." Using the data from the original article, this case study explains the (many) metrics used to evaluate binary classifiers, shows the challenges of defining algorithmic fairness, and starts a discussion of the context, ethics, and social impact of data science.
- Bite Size Bayes: An introduction to probability with a focus on Bayes's Theorem.
- Astronomical Data in Python: An introduction to SQL using data from the Gaia space telescope as an example.

The notebooks

For each of the notebooks below, you have three options:

- If you view the notebook on NBViewer, you can read it, but you can't run the code.
- If you run the notebook on Colab, you'll be able to run the code, do the exercises, and save your modified version of the notebook in a Google Drive (if you have one).
- Or, if you download the notebook, you can run it in your own environment. But in that case it is up to you to make sure you have the libraries you need.

Notebook 1

Variables and values: The first notebook explains how to use Jupyter and introduces variables, values, and numerical computation.

Click here to run this notebook on Colab

or click here to download it

Notebook 2

Times and places: This notebook shows how to represent times, dates, and locations in Python, and uses the GeoPandas library to plot points on a map.

Click here to run this notebook on Colab

or click here to download it

Notebook 3

Lists and Arrays: This notebook presents lists and NumPy arrays. It discusses absolute, relative, and percent errors, and ways to summarize them.

Click here to run this notebook on Colab

or click here to download it

Notebook 4

Loops and Files: This notebook presents the `for` loop and the `if` statement; then it uses them to speed-read *War and Peace* and count the words.

Click here to run this notebook on Colab

or click here to download it

Notebook 5

Dictionaries: This notebook presents one of the most powerful features of Python, dictionaries, and uses them to count the unique words in a text and their frequencies.

Click here to run this notebook on Colab

or click here to download it

Notebook 6

Plotting: This notebook introduces a plotting library, Matplotlib, and uses it to generate a few common data visualizations and one less common one, a Zipf plot.

Click here to run this notebook on Colab

or click here to download it

Notebook 7

DataFrames: This notebook presents DataFrames, which are used to represent tables of data. As an example, it uses data from the National Survey of Family Growth to find the average weight of babies in the U.S.

Click here to run this notebook on Colab

or click here to download it

Notebook 8

Distributions: This notebook explains what a distribution is and presents 3 ways to represent one: a PMF, CDF, or PDF. It also shows how to compare a distribution to another distribution or a mathematical model.

Click here to run this notebook on Colab

or click here to download it

Notebook 9

Relationships: This notebook explores relationships between variables using scatter plots, violin plots, and box plots. It quantifies the strength of a relationship using the correlation coefficient and uses simple regression to estimate the slope of a line.

Click here to run this notebook on Colab

or click here to download it

Notebook 10

Regression: This notebook presents multiple regression and uses it to explore the relationship between age, education, and income. It uses visualization to interpret multivariate models. It also presents binary variables and logistic regression.

Click here to run this notebook on Colab

or click here to download it

Notebook 11

Resampling: This notebook presents computational methods we can use to quantify variation due to random sampling, which is one of several sources of error in statistical estimation.

Click here to run this notebook on Colab

or click here to download it

Notebook 12

Bootstrapping: Bootstrapping is a kind of resampling that is well suited to the kind of survey data we've been working with.

Click here to run this notebook on Colab

or click here to download it

Notebook 13

Hypothesis Testing: Hypothesis testing is the bugbear of classical statistics. This notebook presents a computational approach to the topic that makes it clear that there is only one test.

Click here to run this notebook on Colab

or click here to download it