Data Science from Scratch

First Principles with Python

Stephen CUI

March 20, 2023

Contents

Chapter 1

Working with Data

1.1 Exploring Your Data

After you've identified the questions you're trying to answer and have gotten your hands on some data, you might be tempted to dive in and immediately start building models and getting answers. But you should resist this urge. Your first step should be to explore your data.

1.1.1 Exploring One-Dimensional Data

An obvious first step is to compute a few summary statistics. You'd like to know how many data points you have, the smallest, the largest, the mean, and the standard deviation.

But even these don't necessarily give you a great understanding. A good next step is to create a histogram, in which you group your data into discrete buckets and count how many points fall into each bucket.

1.1.2 Two Dimensions

1.1.3 Many Dimensions

With many dimensions, you'd like to know how all the dimensions relate to one another. A simple approach is to look at the *correlation matrix*, in which the entry in row i and column j is the correlation between the ith dimension and the jth dimension of the data.

A more visual approach (if you don' t have too many dimensions) is to make a scatterplot matrix Figure 1.1 showing all the pairwise scatterplots.

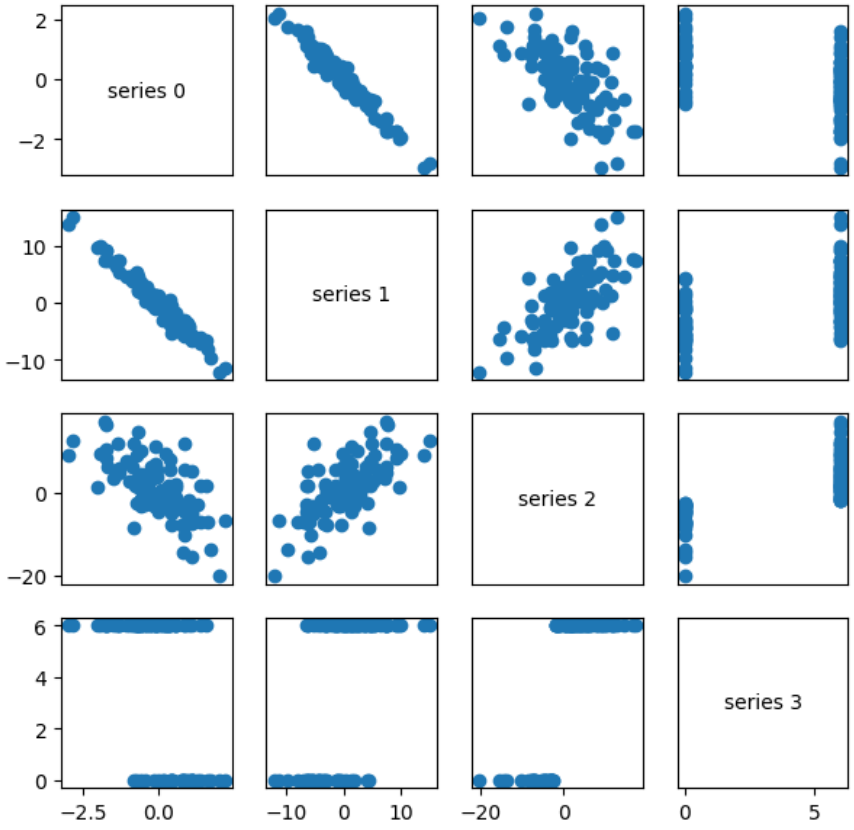

Figure 1.1: Scatterplot matrix