Data Science from Scratch

First Principles with Python

Stephen CUI

March 20, 2023

Chapter 1

Visualizing Data

There are two primary uses for data visualization:

- To explore data
- To communicate data

1.1 Bar Charts

A bar chart is a good choice when you want to show how some quantity varies among some discrete set of items.

A bar chart can also be a good choice for plotting histograms of bucketed numeric values in order to visually explore how the values are distributed.

Be judicious when using plt.axis. When creating bar charts it is considered especially bad form for your y-axis not to start at 0, since this is an easy way to mislead people (Figure 1.1)

1.2 Line Charts

We can make line charts using plt.plot. These are a good choice for showing trends.

1.3 Scatterplots

A scatterplot is the right choice for visualizing the relationship between two paired sets of data.

If you're scattering comparable variables, you might get a misleading picture if you let matplotlib choose the scale. If we include a call to `plt.axis("equal")`, the plot (Figure 1.2) more accurately shows that most of the variation occurs on test 2.

1.4 For Further Exploration

- The matplotlib Gallery will give you a good idea of the sorts of things you can do with matplotlib (and how to do them).
- seaborn is built on top of matplotlib and allows you to easily produce prettier (and more complex) visualizations.
- Altair is a newer Python library for creating declarative visualizations.
- D3.js is a JavaScript library for producing sophisticated interactive visualizations for the web. Although it is not in Python, it is widely used, and it is well worth your while to be familiar with it.

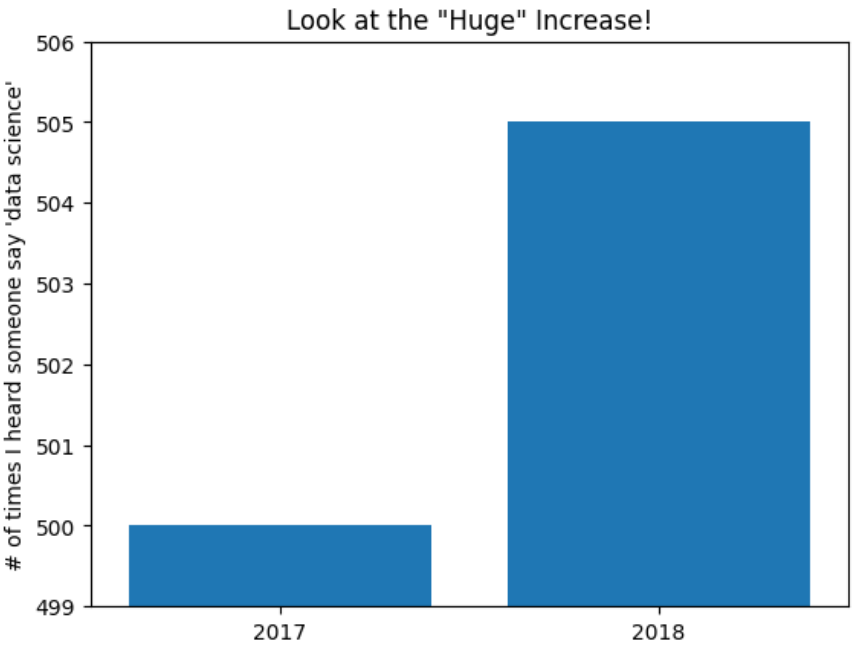

Figure 1.1: A chart with a misleading y-axis

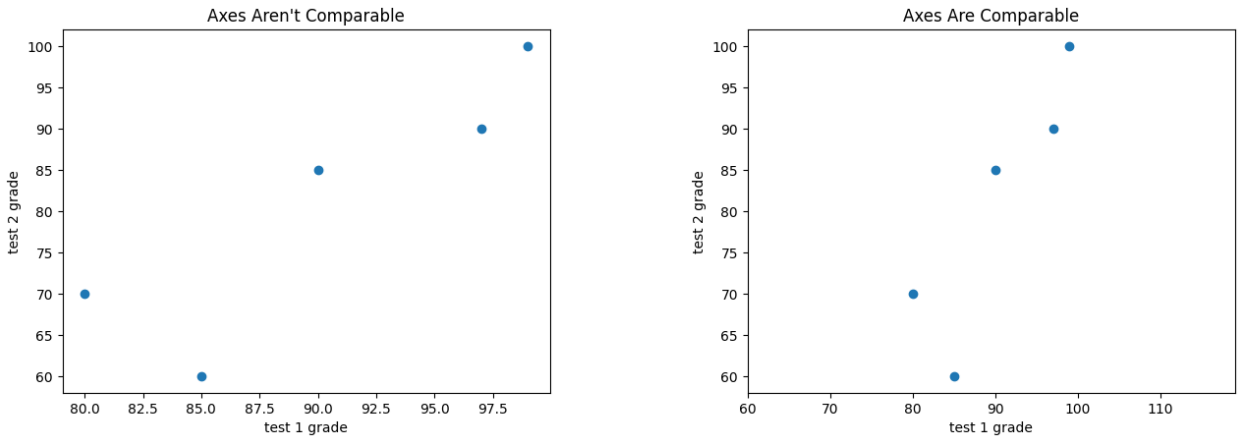

Figure 1.2: User comparable scale in scatterplot

- Bokeh is a library that brings D3-style visualizations into Python.