UNIT PLAN OVERVIEW

Title
Introduction to Data Visualization using Python

Author
Daniel Moscoe
dmoscoe@gmail.com

Contents

Related courses
New York Algebra I
New York High School Computer Science
New York High School Data Science
AP Computer Science Principles

Essential questions and enduring understandings
EQ: How can we summarize large datasets?
EU: We can summarize large datasets with descriptive statistics. Descriptive statistics are useful because they summarize a large amount of information with just a few numbers.

EQ: How can we compare large datasets using visualizations?
EU: We can compare some aspects of large datasets using histograms. Histograms help us describe and compare the center, shape, and spread of univariate datasets. In general, data visualizations are images that help us see patterns in datasets, and they help us compare datasets with each other. Data visualizations are a powerful, widely used tool for telling stories with data.

EQ: How can programming help us find meaning in data?
EU: Programming helps us work with data by streamlining repetitive computations and procedures. We can create our own programs, and we can modify programs created by others to develop powerful and specialized tools that help us understand our data.

Introduction
Getting started with data visualization in Python is an exciting and challenging task for high school students. They're called upon to invoke technical skills from math class and earlier computer science classes. But they may also need to draw on ideas from social studies or information literacy in order to understand and communicate new ideas about the data they're working with.

This unit offers students a step-by-step guide to begin to meet the challenge of data visualization. The unit is divided into six lessons and three summative assessments evenly spaced throughout the unit.

Each lesson is designed to require about 1 hour of class time, although there are some ideas that may merit deeper exploration. For example, it may be worthwhile to spend one class period just sharing and discussing the different ways that students grouped and analyzed the baby names data in lessons 5 and 6.

Data scientists usually work with reference material close at hand, and students should be encouraged to do likewise. Many popular Python modules, including Pandas and matplotlib, are partially described in "cheat sheets" that give very brief summaries of the main tools of the module. In this unit these materials are not cheating, so we refer to them as "reference sheets," even though this is not what they're usually called. In addition to these references, students should practice making use of their previous work as sources of example code, and even collaborate with each other. The goal is always to be able to explain in as much detail as possible how the code functions, and to make progress toward being able to produce working code oneself.

This unit focuses on histograms and simple descriptive statistics about center and dispersion. When students complete the unit, they should be well-positioned to explore bivariate data visualizations. They should also be able to exercise judgment about how different kinds of data (quantitative, factors) lend themselves to different visualization strategies. During the unit, students can always go further by experimenting with or researching neighboring functions on the reference materials.

Standards

NY State Computer Science and Digital Fluency Learning Standards

- 9-12.CT.1. Create a simple digital model that makes predictions of outcomes.
- 9-12.CT.2. Collect and evaluate data from multiple sources for use in a computational artifact.
- 9-12.CT.3. Refine and visualize complex data sets to tell different stories with the same data set.

NY State Next Generation Mathematics Learning Standards

- AI-S.ID.1. Represent data with plots on the real number line (dot plots, histograms, and box plots).
- AI-S.ID.2. Use statistics appropriate to the shape of the data distribution to compare center (median, mean) and spread (interquartile range, sample standard deviation) of two or more different datasets.
- AI-S.ID.3. Interpret differences in shape, center, and spread in the context of the datasets, accounting for possible effects of extreme data points (outliers).
- AII-S.IC.2. Determine if a value for a sample proportion or sample mean is likely to occur based on a given simulation.
- AII-S.IC.4. Given a simulation model based on a sample proportion or mean, construct the 95% interval centered on the statistic (+/- two standard deviations) and determine if a suggested parameter is plausible.

AP Computer Science Principles Learning Objectives

- DAT-2.A. Describe what information can be extracted from data.
- DAT-2.C. Identify the challenges associated with processing data.
- DAT-2.D. Extract information from data using a program.
- DAT-2.E. Explain how programs can be used to gain insight and knowledge from data.

Summary of student portfolio pieces

- *Designing Histograms*. Given a dataset as CSV, students create histograms that reveal the shape of their datasets. Students choose axis labels, colors, bin widths, and titles that make their plots informative summaries of the data they describe. Students explain their choices in the context of the data. Code and explanations accompany the plots.
- *Histograms and Descriptive Statistics*. Students compute descriptive statistics for their chosen datasets. Students add elements to their histograms that show appropriate measures of center and spread. Code and explanations accompany the updated plots.
- *Using Histograms to Compare Groups*. Students analyze how the properties of their dataset may vary by group. Students visually compare the shape, center, and spread of groups within their dataset by creating histograms in stacked subplots ("small multiples"). Students interpret their findings in the context of the data. Code and explanations accompany the plots.