INTRODUCTION TO DATA VISUALIZATION USING PYTHON
LESSON PLANS

Unit and lesson

Data visualization with matplotlib, lesson 1.

Related 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.

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.

Goal

Select data that can be visualized with a histogram; use matplotlib to generate the default histogram.

Key terms

Histogram, dataframe, matplotlib.

Opening activity

Students work in pairs to locate a data visualization on the internet or in a book. Together they sketch responses to questions for whole-class discussion: what did you learn about the underlying data set from the visualization? What elements of the visualization helped you see this? Why might we create a data visualization?

Direct instruction

Discussion based on the Introduction, Goal, and Histogram review sections of the student work Jupyter notebook. Give special attention to comparing the code and visualizations for the restaurant sales data.

Student work

Working independently or with a partner of the student’s choice, students complete classwork prompts.

Summary

Discussion based on Summary section of student work Jupyter notebook. Give special attention to number 11: “What questions do you have about this lesson, or what predictions can you make about what we might learn later in this unit?”

Strategies to improve accessibility

Students can scaffold their own code-writing by copying and pasting from earlier examples. Then they can edit specific sections of the code. Annotating example code with questions or descriptions of the function of each line is a useful way for students to work toward writing their own code. These annotations can be discussed with a classmate or teacher.

Students can also view examples of classmates' work posted on GitHub or some other public site. Analyzing, annotating, and evaluating another student's work could also help a student progress toward writing their own code.

Strategies to improve student engagement and collaboration

Students post their work on a public site, like a GitHub repository. Students could also collaborate in person with classmates or make use of an online or physical class message board to discuss the assignment.

LESSON PLAN

Unit and lesson

Data visualization with matplotlib, lesson 2.

Related 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.3. Interpret differences in shape, center, and spread in the context of the datasets, accounting for possible effects of extreme data points (outliers).

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.

Goal

Customize a basic histogram by adding labels and titles, and by setting the number of bins.

Key terms

Axes labels, bins, keyword argument

Opening activity

Give students a link to a folder containing the datasets available for this unit. Spend some time at the beginning of class looking at the columns, rows, and values of two data sets of students' choice. Part of students' work in this unit will be to choose what data to work with.

Direct instruction

Discussion based on the Introduction, Goal, and Labels and titles sections of the student work Jupyter notebook. Give special attention to the question, "What are the characteristics of good axes labels and titles?" One way to explore this would be for students to revise the labels and titles from the visualizations they found at the beginning of lesson 1. They can either improve or weaken the original labels/titles, and explain why their new choices are better/worse.

Student work

Working independently or with a partner of the student's choice, students complete classwork prompts. One optional strategy is for students to work independently for 10 minutes and then conference on their work so far with a classmate. After the conference, return to independent work.

Summary

How did the shape of the data help students decide what titles and labels might be reasonable? How did the values along the axes help them decide? How did students decide on the number of bins appropriate for their histogram? Did anyone make a big change?

Strategies to improve accessibility

Point out to students where they can find function names for setting titles and labels on the "cheat sheet" reference materials. What other sections of the references does the student think will be useful? Students could highlight methods that are discussed in class or methods they have questions about.

Strategies to improve student engagement and collaboration

This lesson lends itself well to students sharing and critiquing each others' work. Students who are shy about discussing their work could offer comments on histograms found online, although this would not be as useful as a discussion that could directly lead to improvements in the student's own histograms.

LESSON PLAN

Unit and lesson
Data visualization with matplotlib, lesson 3.

Related 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.

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.

Goal
Choose a measure of center appropriate for the data. Represent the center of the data set graphically.

Key terms
Mean, median, xlim.

Opening activity
Look at the two histograms in the student work Jupyter notebook. Talk with a neighbor about what you think typical values for each of these distributions might be. Do the titles and axes labels help you decide? How could we agree on a computation for a typical value?

Class discussion
Discussion based on the Introduction, Goal, and Measures of center sections of the student work Jupyter notebook. Give special attention to prompts 1-3, emphasizing the effects of skewness in a data set.

Student work
Working independently or with a partner of the student's choice, students complete classwork prompts.

Summary
Discussion based on Summary section of student work Jupyter notebook. Give special attention to how students used xlims to zoom in on the regions containing the means and medians. Did anyone use xlims to zoom out?

Strategies to improve accessibility

Students may benefit from a brief review of the arithmetic behind calculating mean and median. Experiment with different small sets of values. How is the median affected by a few very large values? How is the mean affected?

Strategies to improve student engagement and collaboration

Students post their work on a public site, like a GitHub repository. Students could also collaborate in person with classmates or make use of an online or physical class message board to discuss the assignment.

LESSON PLAN

Unit and lesson

Data visualization with matplotlib, lesson 4.

Related 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.

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.

Goal

Choose a measure of spread appropriate for the data. Represent the spread of the data graphically.

Key terms

Standard deviation, interquartile range, variability.

Opening activity

Discuss the question, Why might we want to predict a random data point from a distribution? What sorts of predictions could we make with these distributions / data sets? This is a motivation for why we might be interested in the variability of a data set.

Students can discuss with a partner a data set that might have a large variability, and one that might have a very small variability (but not zero). Is there some way you could alter the systems mentioned to increase or decrease their variability?

Direct instruction

Discussion based on the Introduction, Standard deviation, and Interquartile range sections of the student work Jupyter notebook. Emphasize that standard deviation and IQR are not an attempt to measure the exact same interval. For normal distributions, the fraction of the data lying within 1 standard deviation of the mean is about 68%.

Student work

Working independently or with a partner of the student's choice, students complete classwork prompts.

Summary

Discussion based on Summary section of student work Jupyter notebook. What if we were to move the bars representing interquartile range to different regions of the dataset? Would they still enclose 50% of the data? Why or why not?

Strategies to improve accessibility

In this lesson, students are starting to bring together a significant amount of statistical and coding knowledge. Encourage students to refer to their previous work in this unit for examples and for reviews of key concepts. Some students may benefit from representing variability in a way that doesn't involve numbers. Questions like, "What is greater: the variability of the temperature in the first week of January, or the variability of the temperature across the entire year?" can help students connect with the idea of dispersion in a collection of measurements.

Strategies to improve student engagement and collaboration

Students post their work on a public site, like a GitHub repository. Students could also collaborate in person with classmates or make use of an online or physical class message board to discuss the assignment.

LESSON PLAN

Unit and lesson
Data visualization with matplotlib, lesson 5.

Related 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).

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.

Goal
Drop missing values from a data set. Remove extreme values. Consider when these manipulations may or may not be appropriate.

Key terms
Descriptive statistics, NaN, filter, subset.

Opening activity
Full class discussion: what are some situations that might result in “messy” data? What might messy data look like?

Direct instruction
Discussion based on the Introduction, Caution!, Goal, and Examine the data sections. Pay special attention to the difference between counting distinct names in this data set, and counting children with a given name. Each row represents a name, not a child.

Student work
Working independently or with a partner of the student’s choice, students complete classwork prompts. Because this data is so extremely skewed, it might be hard for some students to understand the meanings of the histograms, and what the data themselves are showing. One question that could yield understanding is, “Why are the bars representing the most common names, the shortest? Where are they located on the histogram?”

Summary
Discussion based on Summary section of student work Jupyter notebook. If it was difficult for students to understand how this data was represented visually, what other suggestions for visualizing this data set do they have? What are some of the important patterns / trends that they see in this data?

Strategies to improve accessibility

A simpler data set could still illustrate the main ideas in this lesson about tidying data. Students may choose to populate a CSV with information from a class list, and intentionally omit some values. This could provide some conceptual understanding as the student works toward manipulating the large data set for this lesson.

Students may also want to take some time discussing different ways to filter the data set before proceeding with the lesson. What conditions might some rows satisfy, and other rows fail? How can you state these conditions in Python?

Strategies to improve student engagement and collaboration

Students post their work on a public site, like a GitHub repository. Students could also collaborate in person with classmates or make use of an online or physical class message board to discuss the assignment.

LESSON PLAN

Unit and lesson
Data visualization with matplotlib, lesson 6.

Related 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.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.

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.

Goal
Use subplots to compare groups in the baby names data set.

Key terms
Groups, filtering, bar plot, subplots.

Opening activity
Class discussion: What are some different ways that we could group the baby names data? What predictions can you make about how the baby names data might vary by group? How many different ways can we think of to break this data into groups?

Direct instruction
Discussion based on the Introduction, Goal, and Grouping data sections of the student work Jupyter notebook. Give special attention to whether the summary values in the table with grouped data are meaningful. Why is the column heading `first_letter` lower than the others? Because it contains the indices of each row. The index plays a special role in a data frame, and we can refer to the index column with `df.index`.

Student work
Working independently or with a partner of the student's choice, students complete classwork prompts.

Strategies to improve accessibility
Grouping using `groupby()` followed by a summary function can be a difficult leap to make for some students. Students can use the Pandas reference sheet to help them experiment with using different summary functions on different columns of the data. Each time, they should try to describe the meaning

of each resulting column and indicate whether this would be a useful way to summarize the information or not.

Strategies to improve student engagement and collaboration

Students post their work on a public site, like a GitHub repository. Students could also collaborate in person with classmates or make use of an online or physical class message board to discuss the assignment. For this assignment in particular, students may be interested to explore each others' findings about patterns in the baby names data.