Lesson 10: Structuring Data

Overview

Question of the Day: how can we make it easier for computers to process data?

In this lesson, students go further into the collection and interpretation of data, including cleaning and visualizing data. Students first look at how presenting data in different ways can help people to understand it better, and they then create visualizations of their own data. Using the results of a preferred pizza topping survey, students must decide what to do with data that does not easily fit into the visualization scheme that they have chosen. Finally, students discuss which parts of this process can be automated by a computer and which need a human to make decisions.

Purpose

This lesson demonstrates that raw data must be interpreted in some way to help people use it to make decisions. Students engage in both visualization and cleaning of data, and they see how data can be misinterpreted if it is not cleaned properly. Students also see how data must be structured in particular ways to be used by a computer.

Assessment Opportunities

1. **Identify and remove irrelevant data from a data set.**

   As students clean their data in the digital activity, circulate and ask them about the choices that they are making. You may also use the discussion afterwards as a time for them to explain what data they identified as irrelevant.

2. **Create a bar chart based on a set of data.**

   Activity Guide: The bar chart should be filled in. Answers may vary slightly, but should overall be approximately the same as in the exemplar.

3. **Explain why a set of data must be cleaned before a computer can use it.**

   Activity Guide: Students should identify data that needs to be cleaned and explain why it is problematic in its raw form. You can also use the discussion afterwards to prompt students to give a more explicit explanation of why it is necessary.

Objectives

Students will be able to:

- Create a bar chart based on a set of data.
- Explain why a set of data must be cleaned before a computer can use it.
- Identify and remove irrelevant data from a data set.

Links

Heads Up! Please make a copy of any documents you plan to share with students.

For the teachers

- **CSD Unit 5 - Data & Society** - Slides

For the students

- **Structuring Data 2021** - Activity Guide

Standards

Full Course Alignment

CSTA K-12 Computer Science Standards (2017)

- ▶ **DA** - Data & Analysis

Agenda

Warm Up (5 minutes)
Journal

Activity (35 minutes)

Wrap Up (5 minutes)
Journal

Teaching Guide

Warm Up (5 minutes)

As students enter, have the warm-up slide projected on the board which contains 3 different representations from a "Best Class Pet" survey.

Journal

Prompt Which one of these makes it easiest for a human to make a decision about which pet is the most popular?Which one makes it easiest for a computer to make a decision?

Discuss: Have students journal individually, then share with a neighbor, and finally discuss as a whole class. You can record their ideas on the board to refer back to later in the class.

Discussion Goal

Students should understand that different forms of data make it easier for people to make decisions. They should also see that people often do best with visuals, such as the bar chart, while computer do better with numbers, such as the table.

Remarks

> Sometimes the "raw" data, the way the information is first collected, needs to be put in a different form so that humans and computers can more easily understand what it means.

Question of the Day: how can we make it easy for computers to process data?

Activity (35 minutes)

Group: Put students into pairs and give each pair a copy of **Structuring Data 2021**.

Structuring Data Activity Guide

Read the instructions together as a class, ensuring that students understand the problem that they are trying to solve (choosing a pizza topping for the pizza party).

Do This: Have students create the bar chart for the set of raw data given. Some of the answers will intentionally not easily fall into the given choices.

Circulate: Encourage students to use their best judgment on the answers that are difficult to put into the chart, and that these challenges are a normal part of the data problem solving process. Listen for students discussing & collaborating with their partners as they make these decisions.

Reflect: Have students answer the reflection questions at the bottom of the activity guide.

Discuss: After students finish making the chart and filling out the reflection questions, have students share their answers with the class.

Assessment Opportunity

Students should see that there are several ways that answers might be difficult to categorize, whether they are completely irrelevant, not specific enough, or not a given choice. Ignore spelling for now if students don't bring it up.

Remarks

> We've made this chart by hand, but it's also possible for the computer to make it for us. This is especially useful when you have lots of data. What would happen if the computer tried to make a chart with the same data you started with? Do you think it will look the same as yours?

Discuss: Ask students to discuss this prompt with a neighbor. As they do this, pull up the Pizza Party Data App so all students can see it.

Pizza Party Data App: Demonstrate the app to the class, explaining that this app has the same data you all had plus a little more. You can scroll through the first screen to see all of the data available, pointing out some new answers that may also be hard to categorize.

Emphasize the "Show Chart" button at the top of the app and ask students to predict what the chart will look like on the next screen. You can ask students to share their ideas, but don't interpret them as "right" or "wrong" - let students brainstorm a few guesses before clicking the button and revealing the next screen.

1 **Pizza Party Data**

Prompt Ask students to discuss in pairs why the chart looks the way it does and why wasn't the computer able to put everything into the correct category. Have students share their answers with the class.

Discussion Goal

Students should notice that the computer used all the answers in the chart, even the ones that were irrelevant. Students should also notice that it didn't correct any spelling or fix any differences in capitalization - for example, "Cheese" and "cheese" appear as two separate rows on the chart.

Remarks

> When we created our charts, we knew that we needed to leave off some of the answers that didn't make sense, and that some answers, such as "peppers" and "green peppers", actually meant the same thing. We also put everything that had been misspelled into the correct category. Computers

> don't know how to do this, because they don't actually understand what a "pepper" is, or that a misspelled word is the same as a correctly spelled word. That means that we have to clean the data before the computer is able to use it.

Have students turn to the next page in their Activity Guide. We will use this page to help "clean" the data for a computer so it will be easier to interpret.

Do This: Ask students to finish in pairs, cleaning the data and narrowing in on at most 7 choices from the answers. Then decide which pizza topping is the best choice.

Prompt: What changes did you need to make to the data? Was there any that you just needed to throw away completely? Why?

Assessment Opportunity

Students should notice that some data, such as minor misspellings, could be easily "cleaned" into an appropriate category, but that other data, such as "I will be absent", needed to be removed from the set entirely.

Prompt: This was a lot of work, and it was only about twenty votes. How much time do you think it would take to clean the data for a nationwide survey? Can you think of any ways to make sure that we got clean data from the beginning, to save us all of this work?

Discussion Goal

In the end, students should realize that constraining a user's choices can help with this - for example, by using multiple choice rather than a write in answer. They may also bring up ignoring capitalization so “cheese” and “Cheese” aren’t two different results.

Allow students to discuss in pairs, then share out with the class.

Remarks

> When we work with large amounts of data, we want to automate as much of the problem solving process as we can. Because computers can't make the same connections that people can, that means that people have to help organize data in a way that computers can understand it. That means either cleaning the data, or collecting data in a way that makes sure it's clean when we get it.

Wrap Up (5 minutes)

Journal

Prompt: Can you think of a time in the past when you had data collected about you, maybe by filling out a form? What do you think were some strategies this form used to help make sure it collected clean data?