Lesson 3: Filtering and Cleaning Data

Overview

In this lesson, student explore the challenges of working with a messy dataset. First students learn how to identify issues using the Data Visualizer, and then manually clean the data. Following this, students learn about the filtering tools in the Data Visualizer, and use a guided activity to answer data questions that require filtering a dataset.

Purpose

The goal of this lesson is to introduce two crucial concepts when working with data: cleaning and filtering. The datasets in App Lab are generally clean, so a "messy" dataset is imported to a level for students to explore. After discovering why datsets need to be cleaned, students manually clean the dataset.

When working with a dataset to answer a question, the user may want to focus on a subset of the data. In this lesson, students are introduced to the filtering tools in the Data Visualizer so they can accurately filter for the information they are looking for.

Standards

Full Course Alignment

CSP Conceptual Framework

- **DAT-2** - Programs can be used to process data, which allows users to discover information and create new knowledge.

CSTA K-12 Computer Science Standards (2017)

- **DA** - Data & Analysis

Agenda

Warm Up (2 minutes)

Activity (33 minutes)
Filtering Data

Wrap Up (10 minutes)

Objectives

Students will be able to:

- Create filtered charts that answer specific questions
- Explain why data needs to be cleaned
- Use the Data Visualizer to filter data

Preparation

- Preview the filter tool in the Data Visualizer
- Prepare for the demo in the Activity

Links

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

For the teachers

- **CSP Unit 9 - Data** - Slides

For the students

- **Filtering Data Unit 9 Lesson 3** - Activity Guide

Teaching Guide

Warm Up (2 minutes)

Remarks

> We've started to explore how to use charts to process data stored in a table, but there are challenges with doing this. The ability to process that data depends on the users and available tools. Today, we are going to explore ways to refine the data the we can answer even more questions using the Data Visualizer.

Display: Set the scene for the first activity.

Remarks

> Imagine you have used a survey to collect information from students. This is aligned with the first step in the Data Analysis Process.
>
> All of that data is now stored in a table. You are excited to dig into the data and see what you can learn. Let's go!

Activity (33 minutes)

Do This: Instruct students to navigate to Level 1 on Code Studio.

- Open the data tab
- Familiarize yourself with the imported table
- Open the Data Visualizer
- Make Charts:
  - Average Hours of Sleep
  - Favorite Subject

Discuss: *With partners, students discuss the following prompts:*

- What problems came up when trying to create these charts?
- What problems do you see in the data?

Discussion Goal: Students should point out some issues like `4` and `"four"` being charted as different values.

1 Cleaning Data

Remarks

> Datasets can bring about challenges, no matter what their size. There can be incomplete data and invalid data. You might want to combine two tables, with inconsistent data. All of this requires data to be cleaned.
>
> When does data need to be cleaned?
>
> - Data is incomplete
> - Data is invalid
> - Multiple tables combined into one
>
> What leads to "messy data"?
>
> - Users enter in different types of data ("two", 2)
> - Users use different abbreviations to represent the same information ("February", "Feb", "Febr")
> - Data may have different spellings ("color", "colour") or inconsistent capitalization ("spring",

"Spring")

When we clean data, the goal is to make it uniform without changing meaning.

For example: "two" can be changed to 2

Do This: With a partner:

- Clean the Student Info table
- Look for:
  - Different types of data ("two", 2)
  - Different abbreviations to represent the same information ("February", "Feb", "Febr")
  - Different spellings ("color", "colour")
  - Inconsistent capitalization ("spring", "Spring")
- Manually update cells with messy data so they are consistent with other cells, while not changing the meaning of the data.

Do This: Once students have finished cleaning their data, they should remake the original charts and compare with others.

Remarks

Your charts should look similar to others, but it depends how you cleaned your data.

Now you are able to accurately visualize the data in your table!

For this activity, we cleaned a dataset that is similar to one you might create yourself or have uploaded from another location. However, the datasets in the dataset library have already been cleaned for you!

Discuss: *What if I only wanted to look at a subset of my data? How could I do this? For example: I only want to investigate dogs with long lifespans.*

Discussion Goal: We are building towards the idea of filtering data. Students filtered programmatically in previous units, but don't yet know how to use the Data Visualizer tool to accomplish the same goal.

In the discussion, students may suggest deleting unwanted rows or creating new tables. Challenge students by asking what they would do if they wanted to look at several different subsets in the same data? Is it reasonable to create a new table for each subset?

Filtering Data

Remarks

&fa-desktop; The best way to look at a subset of data is to use a filter. In Unit 5, we filtered data programmatically using traversals in order to gain insight into knowledge from that data.

Software programs with built-in tools (like the Data Visualizer) can also be used to filter data. These tools help us find specific information in the data and look for patterns.

Do This: Demonstrate how to filter data for the class.

1. Open up Level 2.
2. Discuss that you want to find out about the peak level of female representation in your state's legislature. What will you filter by? Percentage of Females in Legislature or State?
3. We will filter by state. This means the data for female state legislators who are from your home state are the only data that will be shown.

4. Select "Bar Chart" and "Percentage of Females in Legislators" from the dropdowns.
5. Discuss what the chart displays: the bar farthest to the right represents the year(s) when female representation in the legislature was at its highest in your state's history.

2 Filtering Data

Remarks

> When filtering, the most challenging part is deciding what value you will filter by. Think to yourself: what's the limiting factor? What do I want to make sure all these percentages have in common to be included in this subset?
>
> It's time for you to play around in the tool yourself!

Distribute: Share the Activity Guide with students. Alternatively, they can click on the link in Level 1. This Activity Guide is designed to be used digitally - do not print.

Do This: For the rest of the activity time, students work through the Activity Guide filtering two datasets and answering questions. Students need to copy/paste their charts into the Activity Guide.

Discuss: *Have students share their answers from the Activity Guide and reactions to those findings.*

Discussion Goal: Use this discussion to help students practice the skill covered in lesson 1, distinguishing between:

- What the data shows
- Why that might be the case

Wrap Up (10 minutes)

Discuss: *Why is "Clean and/or Filter" an important part of the Data Analysis Process? What are situations when you would filter vs. clean your data?*

Discussion Goal: Messy data can lead to confusing charts that could be misinterpreted. Data should be filtered when you want to focus on a subset of the data. Filtering allows you to return to the full dataset if you have unanswered questions.

Remarks

> Great job filtering and cleaning data today! This is a useful skill that will help you create more powerful visualizations.
>
> There are other things that programs can to do to better prepare data for visualizations including combining datasets, clustering data, and classifying data. The Data Visualizer isn't the best tool for these more advanced concepts, so if you are interested in exploring more data analysis we suggest researching tools like spreadsheets.

Assessment: Check For Understanding

Check For Understanding Question(s) and solutions can be found in each lesson on Code Studio. These questions can be used for an exit ticket.

Question: What makes manually cleaning data challenging?

3 Check For Understanding