API CAN CODE

COLLEGE OF EDUCATION

Unit 2: Computational Foundations of Data Science

Unit 2 Project

In this lesson, students practice using many of the concepts they've learned over the course of Unit 2, including drawing data in from a JSON, filtering data, and examining the format of data.

Duration: 1 class period

Objective: By the end of this project, students will know how to filter a dataset down to a desired set of variables and observations.

CSTA Standards in this Lesson

Concepts	Standard Identifiers
Data & Analysis	3A-DA-09, 3A-DA-10, 3A-DA-11, 3A-DA-12, 3B-DA-06
Impacts of Computing	3A-IC-24, 3A-IC-27, 3B-IC-25
Algorithms & Programming	3A-AP-13, 3A-AP-14, 3A-AP-15, 3A-AP-16, 3A-AP-17, 3A-AP-18, 3A-AP-19, 3A-AP-23
Computing Systems	3A-CS-03

Lesson activities

1. Students identify a dataset. The instructions guide them to use Kaggle, although you could use another database that provides data in CSVs and follow the same instructions if you prefer. Prompts students respond to in the worksheet will vary depending on the particular data they are interested in and what they find:
    a. **Submit a link to your chosen dataset. What topic does your dataset cover?**
    b. Evaluate the dataset using the 5 V's.
        i. **Is it up to date?**
        ii. **Does it have enough data?**
        iii. **Does it have a variety of different data types?**
        iv. **Does the data appear trustworthy?**
        v. **What columns of data interest you? What questions could these help you answer?**

This work was made possible through generous support from the National Science Foundation (Award # 2141655).

   c. Download the CSV of this dataset.
2. Students use the API Can Code CSV converter to convert this downloaded CSV into a JSON file. We recommend selecting all the headers and converting the entire CSV to a JSON file, then doing the filtering over in EduBlocks. If a student has a particularly huge dataset, it might be appropriate to drop some of the variables at this step by selecting a subset of the headers. **Note: there is a tutorial on the converter here.** Students follow the instructions below:
   a. Select ALL the columns to convert to JSON.
   b. How does this affect the format of the data?
   c. How does the structure change?
   d. What is our experience working with CSV and JSON files in the past?
   e. Generally speaking, explain how rows and columns should be organized in a dataset. What does each row represent? What columns are most useful? What information can be gained from them? Are there any columns you don't plan on using? Why not?
3. Students insert their JSON text into this EduBlocks program. Make sure to remind them to clone and rename this program so it's saved in their library! Students follow the instructions below:
   a. Identify one variable you'd like to focus on. Modify the code so that only this variable prints out!
   b. Modify your print statement with this block: **(students have found this a little confusing in the past. It may help to clarify with them that the 'var,' 'var2,' etc. should be the actual names of their variables within the JSON)**

   (Found in the "Text" tab of blocks.)
   Use this block to print out a series of variables you are interested in for each observation. What do you notice? What do you wonder?
4. Students filter their data to help answer a question, or questions, of interest. They follow the steps below:
   a. Using an "if" statement from the Logic tab of blocks, filter your observations down based on some criteria of interest. (In the previous lesson, we filtered NBA players by team; in another lesson, we filtered the top 100 movies by genre. What might you use for your dataset?)
   b. Submit a link to your EduBlocks program.
   c. What columns did you select? What trends did you notice in these columns? What do you wonder/what additional questions do you have about the data?
   d. **Explain 2 different ways you can filter columns of your data using certain conditional statements. (lots of answers might be appropriate here,**

This work was made possible through generous support from the National Science Foundation (Award # 2141655).

including a direct match to a particular value or a > or < comparison to a value to get a range)

5. Revision & Reflection. Students reflect on the process of this project, responding to the prompts below:
    a. **Was the data formatted in a way that was useful? (missing values, numbers as strings, columns containing multiple values, etc.)**
    b. **What additional columns of data would have been helpful in answering your question? How could these columns improve your analysis?**

Assessment:

Assess student understanding through participation in class discussions and class activities.

This work was made possible through generous support from the National Science Foundation (Award # 2141655).