API CAN CODE

COLLEGE OF EDUCATION

Unit 2: Computational Foundations of Data Science

Lesson 2.2: Manually Processing CSV Data

In this lesson, students will do some basic data manipulation from a CSV, first in Google Sheets and then briefly in EduBlocks. The goal of the lesson is primarily to suggest the benefit of programming over manual manipulation.

Duration: 90 minutes

Objective: By the end of this lesson, students will know some basic functions and techniques to work with data in Google Sheets. They will also know the distinction between doing these functions by hand vs. programmatically.

Lesson Walkthrough: Unit 2 Lesson 2 - Teacher Walkthrough

CSTA Standards in this Lesson

Identifier	Concept	Subconcept	Standards
HS-DAT-DC-23	Data & Analysis	Data Processing	Use a digital tool to clean and organize text-based data.

CSTA Data Science Specialty Standards in this Lesson

Identifier	Concept	Subconcept	Standards
S1-DSC-CC-02	Data Science	Creation & Curation	Interpret metadata when using data collected by others.
S1-DSC-CC-03	Data Science	Creation & Curation	Develop programs to manipulate and transform data to prepare for analysis.

Lesson activities

Warmup - Explore the Streetlights Dataset (10 min)

(CSTA standards in this activity: 3A-DA-09, 3A-DA-10)

- **[Optional].** Students download the OpenDataDC Streetlights dataset. Do this if you want students to use the most up-to-date version of the Streetlights dataset!
    - Click the button that looks like a cloud with a downwards arrow pointing from it.
    - Under the “CSV” category, select “Download file previously generated on [last update date]”

This work was made possible through generous support from the National Science Foundation (Award # 2141655). API Can Code is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) License

    - Save the file locally.
- **[Optional].** They import the CSV into Google Sheets.
- Open the dataset in Google Sheets [this is an older copy of the Streetlights dataset – if you use this, some of the comparison numbers will be slightly off towards the end of the lesson]. Explore the dataset. What is the data about? What kinds of variables are stored within the dataset?
- ***Teacher Note:*** *This dataset is quite large, and students might encounter problems like frozen screens or slow responses. If this happens, they should close other browser tabs and exit other applications.*

Streetlights - Manual Manipulation (40 min)

(CSTA standards in this activity: 3A-DA-09, 3A-DA-10)

- Students do some basic data manipulation within Google Sheets.
- **Missing data.** Students use conditional formatting across the entire dataset to highlight missing data cells.
    - Press Ctrl+A to select the entire dataset.
    - Go to Format → Conditional Formatting.
    - Format cells if… "Is empty."
    - Color the cells any color you want!
    - Look through the data. Is there any other cell value that you might like to have highlighted?
    - *Teacher Note: Students might want to highlight N/A values as well.*
- **Countif.** Students use a "CountIf" function to count the number of observations/cell meeting a given condition/particular variable value. We first introduce the CountIf function and its components, then the following example:
    - Go to the bottom of the "Fixture Style Description" column, column O. If necessary, use the "Add 1000 more rows at the bottom button," but before you press "Add," change the 1000 to a 1.
    - In the row **below** the last data value in this column, enter the following: =COUNTIF("Posttop", O2:O72037). 72037 may be replaced by whatever is the last row of data in your sheet, since your downloaded data may differ slightly from our version!
    - Press Enter. What number appears? What does this number mean?
    - ***Teacher Note:*** *This number represents the number of streetlights meeting the "Posttop" fixture style.*
    - Go to Column Y, "LIGHTMANUFACTURER_DESC". In the row below the last data value, type a similar function: =COUNTIF("General Electric", Y2:Y72037) and press "Enter". How many street lights are made by General Electric?
    - How might missing data cells affect these counts?
    - ***Teacher Note:*** *If a street light is made by General Electric, but that manufacturer isn't recorded here, then the count will be too low.*

This work was made possible through generous support from the National Science Foundation (Award # 2141655). API Can Code is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) License

- **Average.** Students use a function to calculate a mean value for several of the variables.
    - Go to the bottom of column AV, “WATTAGE 1”, and in the row below the last data value, type =AVERAGE(AV2:AV72037). What is the average wattage of street lights in the dataset? How might missing data affect this average?
    - ***Teacher Note:*** *Missing data values won’t be treated as 0s for the purposes of the average function, so they won’t mathematically change the average. However, if these lights have wattages that are actually above or below the calculated average, because the values are missing, they won’t be counted – so the REAL average could be higher or lower, respectively!*
    - Go to the bottom of column AH, “POLEHEIGHT_DESC”, and type a similar function: =AVERAGE(AH2:AH72037) and press “Enter”.
    - This function doesn’t seem to work. Why not? What went wrong, and how could we fix it?
    - ***Teacher Note:*** *This column stores a string because each cell contains both a number and the units (ft). To calculate the average of these values, we’d need numerical values ONLY in these cells!*
- **Variable transformation.** Students create new columns using a formula.
    - Look at column AG, “POLECOMPOSITION_DESC.” What information does this column store?
    - Click on the header “AG,” right-click, and select “Insert 1 column right.” This will create a new, blank column. In the title row, enter “Wood.”
    - Go to cell AH2. Enter =IF(AG2=”Wood”, “Yes”, “No”) and press “Enter”.
    - Click in cell AH2, click on the small circle in the bottom right corner of the cell, and drag downwards all the way to the bottom of the sheet (whoa!). This should apply this formula to *all* the cells in this column.
    - Note: Does this keep comparing cells to AG2? Or does it compare each new cell to a different cell? (Click in AH72037 to check!)
    - How does missing data affect these “Yes” and “No” values?
    - ***Teacher Note:*** *Missing data will be recorded as “No” because the AG cell does not match “Wood.” A ‘CountIf’ of AH cells would probably under-count the number of wooden poles.*

This work was made possible through generous support from the National Science Foundation (Award # 2141655). API Can Code is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) License

Streetlights - Programmatic Manipulation (20 min)

(CSTA standards in this activity: 3A-DA-12, 3A-AP-16)

- Students look at this program, which imports the live-updated Streetlights data, prints an example streetlight data structure, and counts the number of streetlights with Wood poles.
    - Each "block" in this program represents a line of Python code (you can also see the Python text equivalent on the right-hand side). Which parts of this program do you recognize? What do you NOT recognize?
    - The data structure here is quite different from the tabular structure of the CSV. How is this data organized?
    - Look at the green "if" block. What criteria is required for a "match"? What happens within the "if" statement if these criteria are met?
    - Does the "Wood Pole Count" match the output of our countif statement on wood poles from earlier?
- There are clear differences between the block-based programming process and our Google Sheets work earlier.
    - Which do you think was easier? In what ways?
    - What is challenging about the Google Sheets work?
    - What is challenging about the programming work?
    - *Teacher Note - If students are struggling with answers here, ask them to consider things like needing to scroll through the dataset in Google Sheets, and challenges with understanding the blocks in EduBlocks. Likely, EduBlocks will be seen as a bit more complicated, while Google Sheets has some tedious manual work.*

Exit Ticket (10 min)

(CSTA standards in this activity: 3A-DA-09, 3A-DA-10)

- If I want to take the average of a set of values, which function will give me that?
    - **=AVERAGE(B2:B5)**
    - =COUNTIF(B2:B5, "Yes")

- If I want to count the number of "Blue" cars in a column, which function will do that?
    - =AVERAGE(AD2:AD57)
    - **=COUNTIF(AD2:AD57, "Blue")**

- I surveyed 100 randomly-selected students and calculated their average GPA, but 23 students did not provide a GPA value. I come up with an average GPA of 2.75. How might the missing data have affected this average?

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). API Can Code is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) License