CS100: Studio 1

Introduction to Spreadsheets

September 11, 2019

Instructions

During today’s studio, you will begin your foray into processing data, using operations that are sometimes collectively referred to as data wrangling. You will be working in Google Sheets. As you complete this studio, write your answers on a sheet of paper, or type them into a text file. If you have any questions along the way, raise your hand, and a TA will come over and help you.

Upon completion of all tasks, a TA will give you credit for today’s studio. If you do not manage to complete all the assigned work during the studio period, do not worry. You can continue to work on this assignment until Wednesday before 6 PM. Come by TA hours any time before then to show us your completed work and get credit for today’s studio.

Objectives

By the end of this studio, you will know how to process data in spreadsheets using Google Sheets. Specifically, you will practice: * Loading csv (comma-separated values) files * Writing formulas * Sorting and filtering data * Charting data

Background

Standard data wrangling operations (i.e., ways in which to process data) include: * aggregate (e.g., sum, max, min, etc.) * group by (i.e., aggregate by group) * arrange (e.g., sort) * filter (i.e., delete rows) * mutate (e.g., add or delete columns) * merge (i.e., join two data sets on an attribute value)

Lectures 5a, 5b, and 5c explain these operations, and describe how to carry them out in a spreadsheet. We suggest you skim the slides now, and refer back to them as necessary while working through this studio.

Introduction to Spreadsheets

During this week’s studio, you will be learning to use Google Sheets. Google Sheets is a free alternative to other spreadsheets, such as Microsoft Excel.

Data

During today’s studio, you will be investigating the sugar content of different cereals. This exploration was inspired by the article ‘How Cereal Became the Quintessential American Breakfast’, which discusses the marketing of cereals as a nutritious breakfast. Many breakfast cereals are considered “healthful” due to their added vitamins and low fat content, but their sugar content may tell
a different story. Today, you will use the power of spreadsheets to explore the nutritional content of common breakfast cereals.

Click on this link to download cerealData.csv, a table of 77 brands of popular cereals in the US, along with nutritional, and other information about them.

Go to Google Drive, and click New → Google Sheets to create an empty spreadsheet. Click on File → Import → Upload → Select a file from your device and select cerealData.csv.

You will see some options appear. Import Location should be Replace spreadsheet, as the empty spreadsheet is no longer needed. Separator type should be Detect automatically, as this dataset has already been cleaned. You also want Convert text to numbers, dates, and formulas to be set to Yes. If you have trouble loading the dataset, call on a TA for help!

Spend a few minutes perusing the spreadsheet to get a handle on the types of measurements in the table. There are some cells with the value “?1”. This code indicates a missing data value.

Now that you have the data loaded into a spreadsheet, let’s get acquainted with Google Sheets.

**Getting Started**

To highlight a cell, you can simply click on it. To highlight consecutive cells, you can click on a cell and drag (left, right, up, down) to capture neighboring cells as well.

To highlight an entire column, click on the column’s letter at the top. To highlight an entire row, click on a row’s number at the left.

When you scroll down a spreadsheet, the first row of data disappears. In this case, this row—also called the header row—names the columns, so if it disappears, it will be difficult to remember the variable each column represents. To make your life easier, you can freeze the header row at the top of the screen. To do this, click on any cell in the spreadsheet, and then choose View → Freeze → 1 row. This will freeze the first row of the spreadsheet. To make your life even easier, let’s freeze the first column of data as well. Can you figure out how to do so? (Hint: do the same thing you did to freeze the header row, but for columns!) Later, to deactivate frozen rows, you can select View → Freeze → No Rows. Likewise, for columns.

**Color Scales**

Color scales are a nifty way for us to easily visualize a range of values, and identify extreme points: i.e., maximum and minimum values. You will use them to visualize which cereals are highest and lowest in sugar content (per ounce).
We would like you to color code the cereals based on their sugar content. To do so, highlight the “grams of sugars” column. Then on the main options bar, go to Format → Conditional Formatting → Color Scale. Adjust the settings of Minpoint and Maxpoint, so that the lowest value is white and highest value is a color of your choosing, with those in the middle along the scale. Click Done when you are done.

Notice that the color of a cell now reflects its relative value. Scroll through your newly colored data to identify which cereal is highest in sugar content.

**A.** What two cereals are highest in sugar content, as determined by their color?

**B.** List all the cereals that do not contain sugar.

**C.** How did the color scale assist you in answering these queries? Could you have detected this information without the aid of color scaling?

**D.** What is something you could have done to the data in one step that would make identifying both the lowest and highest values in the column very simple?

**Sorting**

You could have also determined the cereals that are highest and lowest in sugar content by sorting the data. In particular, if you sort data from smallest to largest, then the first value(s) will be the minima, and last value(s) will be the maxima.

**Remark:** Sorting is a fundamental concept in computer science. Indeed, sorting algorithms abound. Computer scientists choose among them depending on whether they want to prioritize time or space efficiency. If you continue your study of computer science beyond CS100, you will likely encounter many sorting techniques.

Google Sheets makes sorting very simple. Just select the column by which you would like to sort and then click Data → Sort sheet by. Try sorting with the A → Z option, and then try again with the Z → A option.

**E.** What is the difference between the A → Z and Z → A options when sorting numerical data?

**F.** Use the sorted data to verify your answers to questions A and B. Name one advantage and one disadvantage of using color scaling versus sorting for identifying minimum and maximum values in a dataset.

**Formulas**

In the interest of your health, we’d like for you to compare the sugar content of various cereals to that of an Oreo cookie. How can we make this comparison meaningful? Should you compare one cookie to one fruit loop? In the last exercise, you did a cursory comparison of “grams of sugar” across cereals. But this was not actually an apples-to-apples comparison. Why not?
Well, because serving sizes differ! An apples-to-apples comparison would require comparing measurements of a standardized unit size. Spreadsheets allow us to mutate data through formulas. You will use them to standardize these measurements. Specifically, you will mutate “grams of sugar” to “grams of sugar per ounce”.

Enter the title “grams of sugar per ounce” at the top of column Q.

To start, you will calculate the grams of sugar per ounce for the cereal in row 2 (‘100%_Bran’). You’ll want to divide the grams of sugar per serving (J2) by the weight in ounces (N2). Thus, you should enter =J2/N2 in cell Q2. (Note: the equal sign is the symbol that denotes the beginning of a formula.) When you press enter, Google Sheets will display the numerical value that results from executing the formula. However, when you select cell Q2, you can see =J2/N2 in the formula bar up above.

Now, select cell Q2, drag the handle in the lower right hand corner down until you have highlighted the whole column, and release. (Note: alternatively, you can simply double click on the handle.) The equation in the first cell should have been copied all the way down the column.

Click on a cell between Q3 and Q78. Does it say =J2/N2? It shouldn’t. The formula should have been changed to reflect the cereal’s row each step of the way. The values that were copied down the column are specific to each row, and hence, each cereal.

You should see that the “Quaker_Oatmeal” row displays #VALUE! in column Q. This symbol indicates there was an error with the formula. The error was caused by the missing numerical value for grams of sugar (represented by “?1”). We will deal with missing values later in the course. For now, we want to manually ignore these missing values, since they will interfere with our subsequent analyses. Delete #VALUE! for Quaker_Oatmeal, leaving Q59 empty.

Color scale column Q (“grams of sugar per ounce”) like you did column J (“grams of sugar”) above. *H. Then sort the data by “grams of sugar per ounce”. Observe the resulting color scale in column Q.

G. What two cereals are highest in sugar content per ounce? Are they the same two cereals that are highest in (straight-up) sugar content?

We are now ready to compare the sugar content of these cereals to that of Oreo cookies. Oreos have approximately 11.7 grams of sugar per ounce (2.5 cookies, if you’re wondering). We are interested in the number of cereals whose sugar content is above this benchmark of 11.7 grams per ounce. How can we find this out?

One possible way would be to use the COUNTIF function. COUNTIF takes in two arguments, a range of data, and a predicate. A range is a contiguous block of data, such as A1: F1 or F1: F8 or A1:F8. A predicate is a function, such as <4.2 or =Bran_Flakes, that evaluates to a Boolean value. There are two
Boolean values, ‘true’ and ‘false’. The function COUNTIF evaluates the predicate on every observation in the given range. As ‘true’ is traditionally represented by a 1, and ‘false’, by a 0, COUNTIF works by tallying a 1 if the predicate is true, and a 0, if not, for every observation.

In cell R78, write a formula in the format of  
\[
\text{=COUNTIF(}<\text{insert range here}>,<\text{insert predicate here}>)
\]
to figure out the number of cells in the “grams of sugar per ounce” column whose value is greater than 11.7. We have left the range and the predicate blank for you to figure out. Note that your predicate must be entered within quotation marks.

**H. How many cereals have more sugar per ounce than Oreos?**

**Charting**

Another informative, and potentially beautiful, way to visualize data is to make charts.

That is precisely what we are going to do next. Begin by making a copy of your spreadsheet. To do so, right click on the ‘cerealData’ sheet tab at the bottom of the page and choose Duplicate. A copy of ‘cerealData’ should now appear as a second sheet. Rename this sheet ‘cerealData2’. You will make your charts on the ‘cerealData2’ sheet.

Choose Insert -> Chart. A chart editor window will appear. Choose column chart. The range of your dataset is A1:Q78. However, for your chart, we only want you to graph the data of two columns: “cereal name” (column A) and “grams of sugar per ounce” (column Q). You can specify these columns by changing the range to A1:A78, Q1:Q78. Make sure that the checkbox for Use row 1 as headers is selected and for Aggregate is not selected. A vertical bar graph should appear on your sheet. If column Q is not already sorted, sort it now!

**I. Looking at the chart, write down your best estimate for the average number of grams of sugar per ounce across all cereals.**

Is this chart easier or harder to comprehend than a color-scaled column? Discuss the pros and cons of both visualizations with your partner.

While the chart allows us to visually estimate the average number of grams of sugar per ounce, we can compute the exact value using the AVERAGE formula.

Navigate back to the original spreadsheet. Let’s create a new row in which to store average values. Type Average in cell A79. For now, we’ll just compute the average for column Q. In cell Q79, enter the formula =AVERAGE(<insert range here>), and enter the correct range.

**J. What is the average number of grams of sugar per ounce? How good was your estimate?**
Filtering

There is a psychology behind the layouts of supermarkets intended to manipulate consumer purchases. Decision variables include where certain departments are located in the store, where items are stacked on the shelves, etc. In our ‘cerealData’ spreadsheet, you should notice a column titled “display shelf.” There are 3 grocery shelf locations: 1, 2, and 3, which correspond to bottom, middle, and top shelves, respectively. Look at the data to see if you can detect any associations between cereals and the shelf that they are placed on.

One tool that might help you here is filtering. Go to Data → Create a filter. You will see a little triangle next to the display shelf header name. Click on that to manually choose which display shelf you want to see. Only cereals at that display shelf will be shown.

Do you notice anything interesting about the differences in values for each different shelf?

Turn off your filter by selecting Data → Turn off Filter.

Grouping Data

By now we have probably peaked your curiosity: is there a relationship between display shelf and sugar content? To answer this question, we need a way to compare cereals across display shelves. We can accomplish this by grouping the data by display shelf. There is a nifty and automatic way of doing just this in spreadsheet, called a pivot table. But you will perform these calculations manually today. That way, you will have a better understanding of what pivot tables do when we learn about them in section.

Let’s start by calculating the average “number of grams of sugar” for each “display shelf”. Discuss with your partner some strategies for manipulating the data so that you can compute these averages. Don’t read on until you’ve formulated a plan.

We will walk you through one way of manually manipulating the data to compute these averages. (If you prefer to compute these values some other way, go ahead. Just make sure you can easily compare the average sugar content of the various display shelves).

Click on the plus sign near the bottom left-hand side of the window, next to the current tab (“cerealData”), to create a new sheet. On the new sheet that appears, label cells A2, A3, and A4 as “Shelf 1”, “Shelf 2”, and “Shelf 3”, respectively. Label column B, “average grams of sugar” in cell B1.

Click on the original tab. Sort the data in this sheet by column M, “display shelf”, in ascending order. Observe that rows 2 through 21 correspond to shelf 1, rows 22 to 42 correspond to shelf 2, and rows 43 to 78 correspond to shelf 3.

Return to your new sheet. In cell B2, type =AVERAGE(cerealData!J2:J21).
This formula will average the values in the range J2:J21 in the “cerealData” sheet. (Use the “!” following a sheet name to indicate that a range is from a different sheet than the current one. Fill in cells B3 and B4 accordingly to find the average sugar content per ounce for shelves 2 and 3.

K. Which shelf has the highest average sugar content?

Use the “display shelf” filter to look at the cereals on shelves 1 and 3. You’ll notice that many of these cereals include bran and wheat. Contrast them with the cereals on shelf 2, where you’ll see cereals like Cinnamon Toast Crunch, Fruity Pebbles, and Cocoa Puffs.

L. Why do you think cereals on that shelf might have the highest average sugar content? Consider the audience they might be marketing to and their line of sight when walking around the store.

Scatterplots

A scatterplot is a graph on the Cartesian (x-y) plane that shows the relationship between two numeric variables. The ‘rating’ variable in the cereal data is a score (out of 100) assigned to the cereal by Consumer Reports based on its quality. You can make scatterplots to try to determine which factors impact a cereal’s rating: e.g., does sugar content impact a cereal’s rating?

Navigate to “cerealData2”. Let’s start by plotting “rating” vs. “calories per serving”. Create a chart, but this time choose Scatterplot. Limit the data range to the columns for calories per serving and rating. Fill in the X-Axis with the range for calories per serving. Make sure the series is ‘rating’. Look at the resulting graph, do you notice a relationship between rating and sugar content?

We can create a trendline to check any perceived relationship! In the chart window, go to Customize → series and select the trendline option. A line showing the linear relationship between calories and rating should appear! Finally use the customize window to title the graph and the x and y axes. The x axis is, “calories per serving”, and the y axis is “rating”.

M. What is the relationship between calories per serving and a cereal’s rating? How clear is the trend?

Independent Application: Exploring Spotify Data

Now that we have walked you through one example of data exploration in a spreadsheet, you are ready to try your own! At the end of each year, Spotify compiles a playlist of the tracks streamed most often that year. The most recent such playlist (Top Tracks of 2018) includes 100 songs and variables capturing features such as “valence” and “instrumentalness” (You can read about these cool ways to quantify and capture music data in depth here, looking at the figure in the column titled ‘Columns’!). Using the same procedure as you did for the cereal data csv file, download the spotifyData as a new Google Sheet
and perform an analysis of your choosing, with the tools introduced above. Consider sorting the data based on features that interest you, using a function like COUNTIF, writing a “formula” for a new metric based on the given features (i.e. Danceability * Energy = ‘Party Rating’), or even exploring which artists appear the most in the top songs. Record your exploration.

End of Studio

When you are done please call over a TA to review your visualizations, and check you off for the studio. If you do not finish within the 2-hour studio period, remember to come to TA office hours to get checked off.