STEM Literacy through Infographics

Creating Graphs of Distribution

Part of a Series of Tutorials on using Google Sheets to work with data for making charts in Venngage

TERC

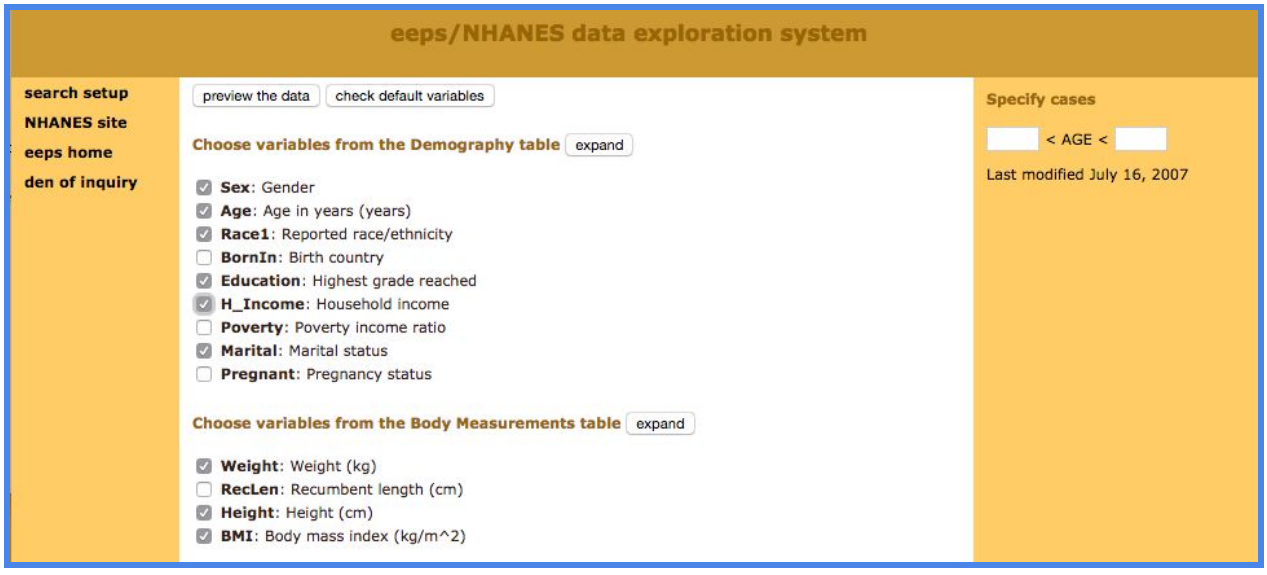
University of Colorado Boulder

These materials are based upon work supported by the National Science Foundation under Grant Nos. IIS-1441561, IIS-1441471, & IIS-1441481. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

Many data sets on the Web include a large number of cases, but you may be interested in making a graph that shows the distribution of those cases into categories - that is, what proportion or percentage of the cases fall into each of a small number of categories, such as gender, income groups or marital status. This document will show you how to use Google Sheets to figure out the distribution you are interested in, then use the results to create one of several kinds of graphs in Venngage.

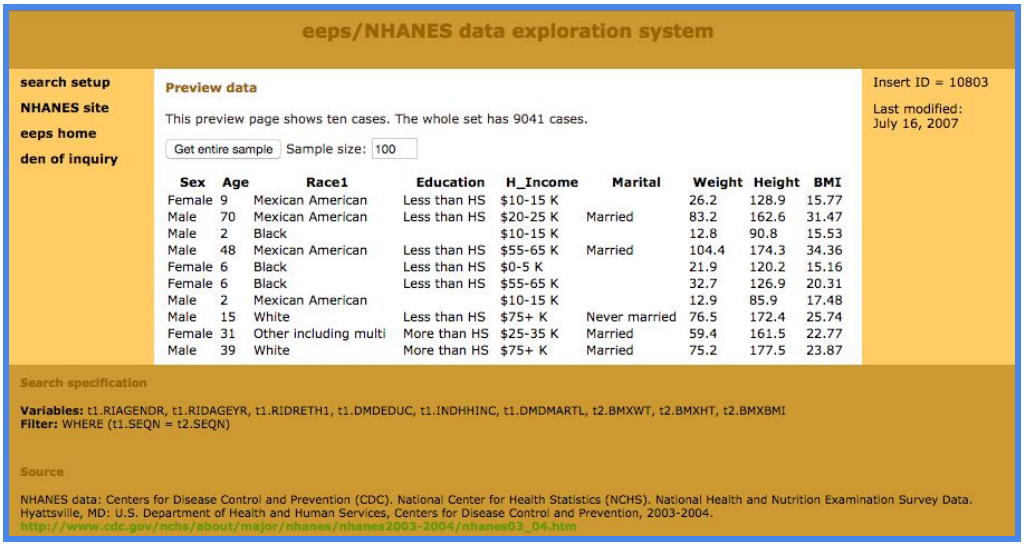
We're going to use as our example the National Health and Nutritional Examination Survey (NHANES) dataset. In the NHANES dataset, there are several thousand cases, each of which equates to a single person. NHANES data are highly encoded with numbers or abbreviations that stand in for readable words, and also come with a lot of metadata (data about data). These two factors make their datasets confusing for the general public. However, there is a useful data exploration interface for NHANES, which has been developed by Tim Erickson, a freelance science and math educator. The data have already been downloaded from NHANES and reformatted to be easily accessible.

Go to the EEPS NHANES data exploration system web page: http://www.eeps.com/zoo/nhanes/source/choose.php. There you will find a webform where you can select the variables you want to examine, including demographics, body measurements, and biochemistry bloodwork information. Next to each variable, it shows you what units the variable is in. For example, "Age" is in years, "Weight" is in kilograms, and "Height" is in centimeters.



In addition to the default variables selected, we are also going to look at Household income, so also check the box for H_Income (Household income), which is in the Demography section. In this interface you are required to look at a preview of your data before you get a

large sample, so click on the button “Preview the Data” at the top. The data exploration system will load a table in the web page.



eeps/NHANES data exploration system

search setup
NHANES site
eeps home
den of inquiry

Preview data

This preview page shows ten cases. The whole set has 9041 cases.

Get entire sample Sample size: 100

Insert ID = 10803
Last modified: July 16, 2007

Sex	Age	Race1	Education	H_Income	Marital	Weight	Height	BMI
Female	9	Mexican American	Less than HS	$10-15 K		26.2	128.9	15.77
Male	70	Mexican American	Less than HS	$20-25 K	Married	83.2	162.6	31.47
Male	2	Black		$10-15 K		12.8	90.8	15.53
Male	48	Mexican American	Less than HS	$55-65 K	Married	104.4	174.3	34.36
Female	6	Black	Less than HS	$0-5 K		21.9	120.2	15.16
Female	6	Black	Less than HS	$55-65 K		32.7	126.9	20.31
Male	2	Mexican American		$10-15 K		12.9	85.9	17.48
Male	15	White	Less than HS	$75+ K	Never married	76.5	172.4	25.74
Female	31	Other including multi	More than HS	$25-35 K	Married	59.4	161.5	22.77
Male	39	White	More than HS	$75+ K	Married	75.2	177.5	23.87

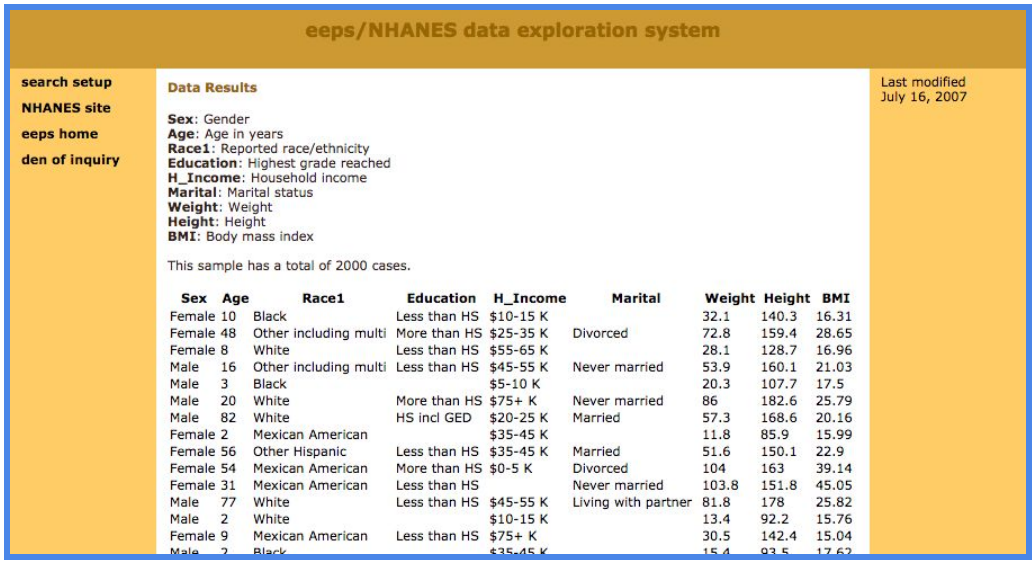
Search specification

Variables: t1.RIAGENDR, t1.RIDAGEYR, t1.RIDRETH1, t1.DMDEDUC, t1.INDHHINC, t1.DMDMARTL, t2.BMXWT, t2.BMXHT, t2.BMXBMI
Filter: WHERE (t1.SEQN = t2.SEQN)

Source

NHANES data: Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2003-2004.
http://www.cdc.gov/nchs/about/major/nhanes/nhanes2003-2004/nhanes03_04.htm

If you are happy with the preview of your data, choose your sample size, enter it in the “Sample size” text box and click on “Get entire sample”. We will choose a relatively large sample - 2000 - to demonstrate how Google Sheets can help you deal with a large amount of data easily. It will now give you 2000 samples.



eeps/NHANES data exploration system

search setup
NHANES site
eeps home
den of inquiry

Data Results

Sex: Gender
Age: Age in years
Race1: Reported race/ethnicity
Education: Highest grade reached
H_Income: Household income
Marital: Marital status
Weight: Weight
Height: Height
BMI: Body mass index

This sample has a total of 2000 cases.

Last modified
July 16, 2007

Sex	Age	Race1	Education	H_Income	Marital	Weight	Height	BMI
Female	10	Black	Less than HS	$10-15 K		32.1	140.3	16.31
Female	48	Other including multi	More than HS	$25-35 K	Divorced	72.8	159.4	28.65
Female	8	White	Less than HS	$55-65 K		28.1	128.7	16.96
Male	16	Other including multi	Less than HS	$45-55 K	Never married	53.9	160.1	21.03
Male	3	Black		$5-10 K		20.3	107.7	17.5
Male	20	White	More than HS	$75+ K	Never married	86	182.6	25.79
Male	82	White	HS incl GED	$20-25 K	Married	57.3	168.6	20.16
Female	2	Mexican American		$35-45 K		11.8	85.9	15.99
Female	56	Other Hispanic	Less than HS	$35-45 K	Married	51.6	150.1	22.9
Female	54	Mexican American	More than HS	$0-5 K	Divorced	104	163	39.14
Female	31	Mexican American	Less than HS		Never married	103.8	151.8	45.05
Male	77	White	Less than HS	$45-55 K	Living with partner	81.8	178	25.82
Male	2	White		$10-15 K		13.4	92.2	15.76
Female	9	Mexican American	Less than HS	$75+ K		30.5	142.4	15.04
Male	2	Black		$35-45 K		15.4	93.5	17.62

Select all the contents of the table and copy. Then paste it into a new Google Sheet file. It should paste correctly into each spreadsheet cell. If it doesn’t, make sure you have only selected the contents of the table. If you select any text before the table, after the table or in the side panel, everything will get pasted into a single cell.

Now that our data are in Google Sheets, we are ready for analysis.

Starting off, it's a good idea to look at what's going on here. In this dataset sample, we have Sex, Age, Race, Education, Household Income, Marital Status, Weight (kg), Height (cm), and BMI. (The kg and cm labels come from the EEPS NHANES web page where we requested the data.) Age, Weight, Height, and BMI have numerical values. Sex, Race, Education and Marital Status have text values that have limited options for what the text is.

Column	Text Values
Sex	Male, Female
Race	Black, White, Mexican American, Other Hispanic, and Other including multi
Education	More than HS, Less than HS, HS incl GED, blank
H_Income	$0-5 K, $5-10 K, $10-15 K, $15-20 K, $20-25 K, $25-$35 K, $35-45 K, $45-55 K, $55-65 K, $65-75 K, $75+ K
Marital	Never Married, Married, Divorced, Widowed, Separated, Living with Partner, blank

The values for Household Income look like numerical values, but they are actually categories, each of which indicates a range of incomes.

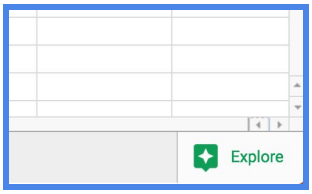
Education is generally blank for young children who are not old enough to have entered school yet.

Marital Status is blank for people 13 years of age and under.

We'll now look at two different ways to explore the distribution of Race in our sample. We'll use Google Sheets to generate the data, then create both a pie chart and a column graph in Venngage.

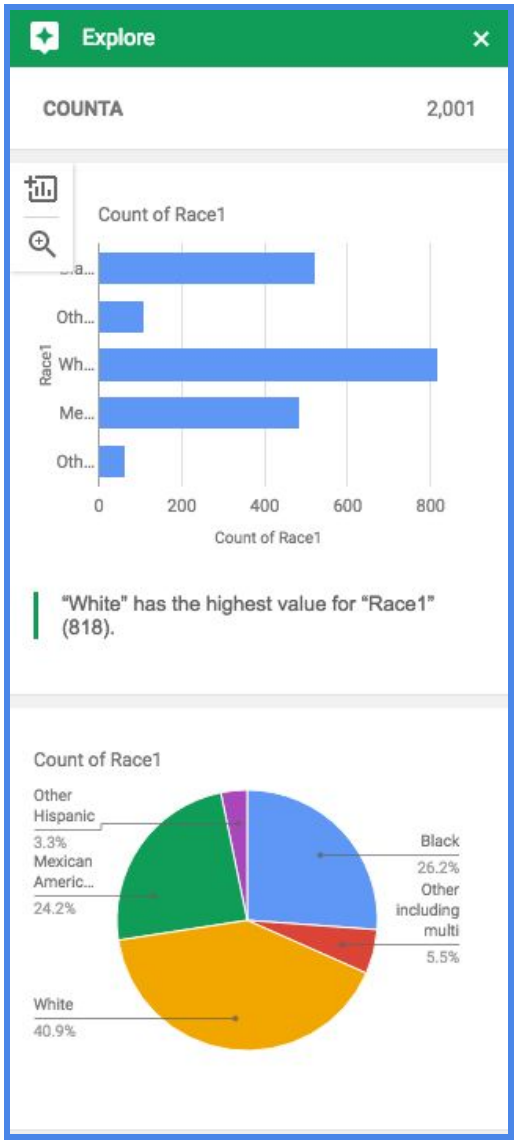
What is the distribution of Race in our sample?

For this question, we will only look at the Race column. First, select the Race column by clicking on the letter above the column. On the bottom right-hand corner of the screen, you'll see a green Explore button. Click on that.



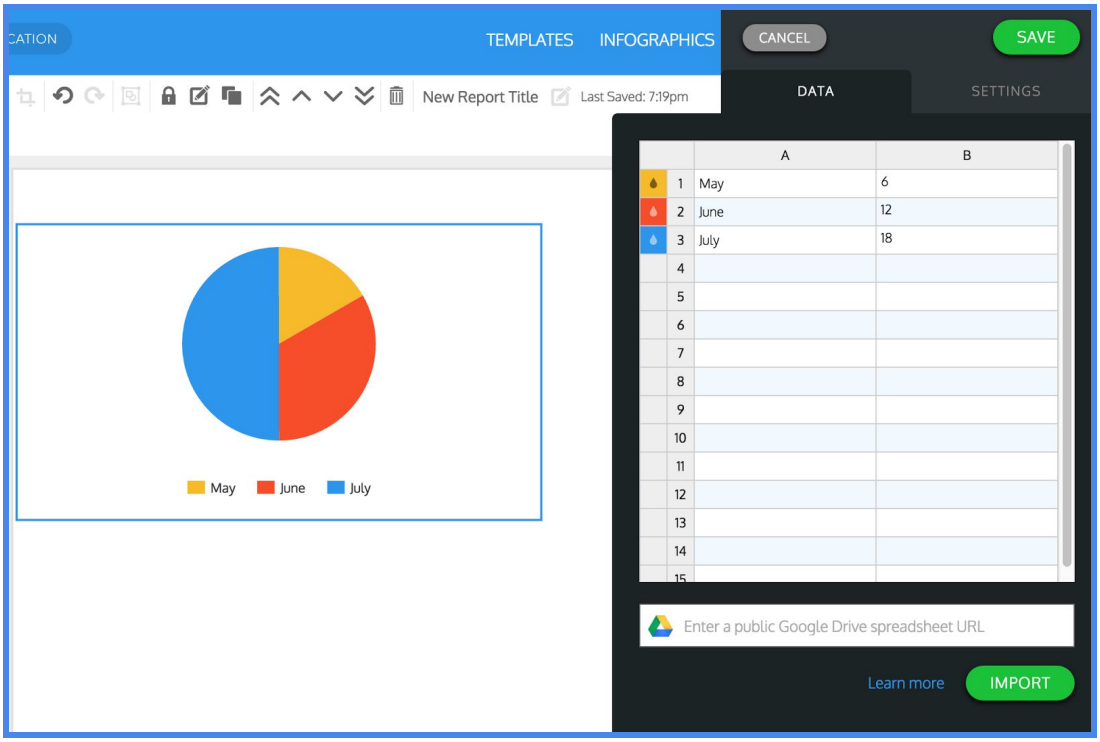
For a single variable, the Explore tool will display statistics about that column. For a column with numerical values selected, the Explore tool will give you the SUM, AVERAGE, MIN, MAX, and COUNTA of the column. COUNTA is the number of cells in the column that contain data. In this dataset, you will see a value of 2,001 for COUNTA. The extra “1” is for the row of column labels at the top. When you select a column with text values, it will only show COUNTA, since SUM, AVERAGE etc. don’t make sense for text values.

With the column for Race selected, the Explore tool will show you a bar graph and a pie chart. (It may take a few seconds, as we’re dealing with a lot of data.) These two charts are two representations of the same distribution. The difference is that the first one shows us how many cases are in each category, while the second one shows us what percentage of cases are in each category. First, we will use the percentages from the pie chart to create a pie chart in Venngage.

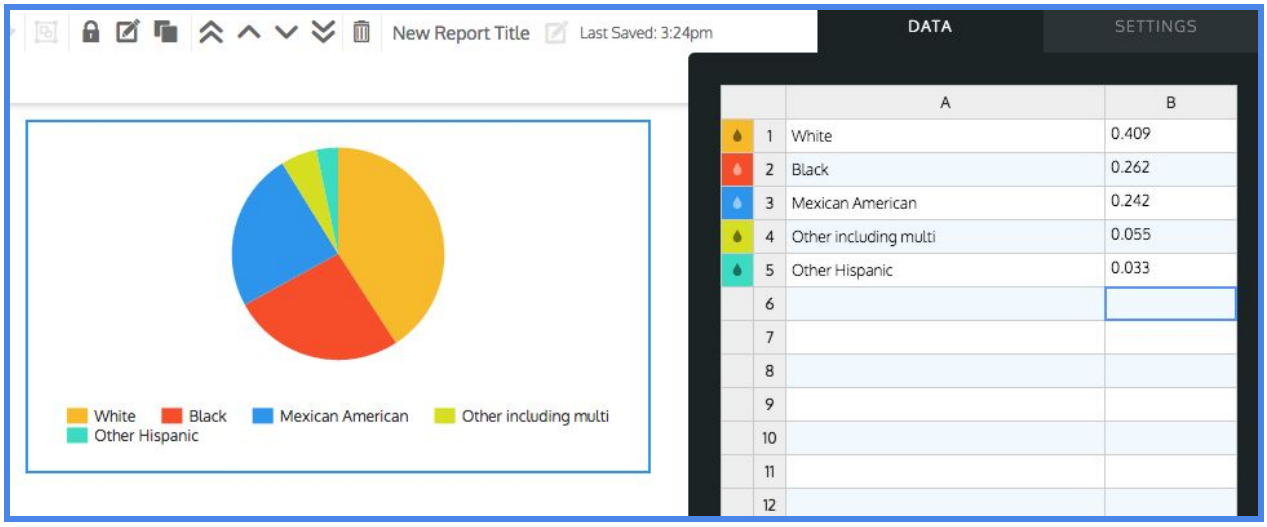


Creating the Pie Chart in Venngage

In Venngage, click on the Charts tab on the left side panel and drag the pie chart to the canvas. Double-click on the pie chart to edit it. You'll see that the sample data is formatted with the labels in the first column and the values in the second column.



Go back and forth between the pie chart in the Google Sheets Explore tool and Venngage to fill in the data table for the pie chart. For each row, add the title of the category in Column A and the percent in Column B. If you are entering something that is not a number in the Venngage data table, you have to double-click on the cell. You may want to order the numbers in increasing or decreasing order, rather than copy directly from Google Sheets. This makes the information easier to grasp.

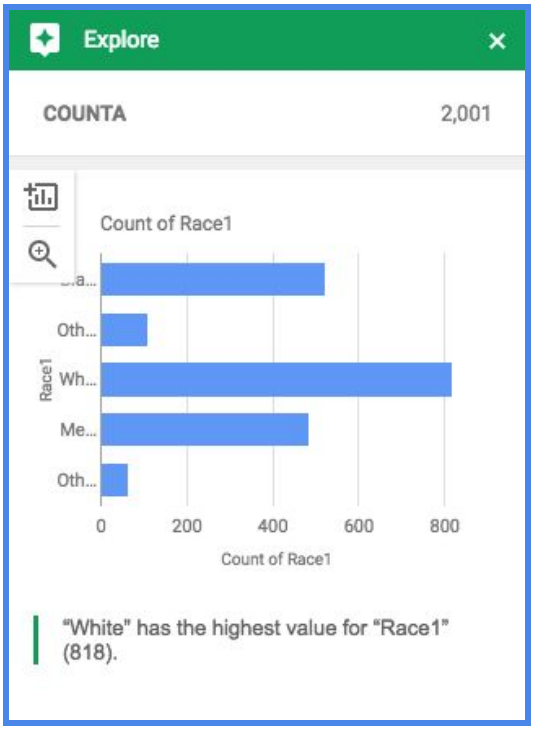


When you are entering percents, don't use the percent sign (%) in Venngage; just enter a number. You can use the number without the percent sign, or if you'd like, the decimal form of the number. For 40.9% you can enter either 40.9 or 0.409. We now have a pie chart in

Venngage showing the distribution of Race in our sample and you can use other Venngage editing tools to customize the pie chart.

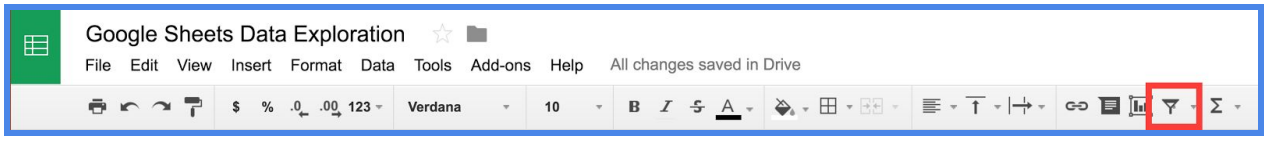
Creating a Column Graph of the Distribution of Race in Venngage

The Explore tool also gave us a bar graph of the distribution of Race. We can also recreate this kind of graph in Venngage, but it's not going to be as easy as pie. Notice the text on the bottom of the graph. It says, "'White' has the highest value for 'Race1' (818)." This gives us the number of people in the White category, but there is no easy way to get the number of people in the other categories directly from the graph. You could estimate what the numbers are, but to recreate the graph in Venngage we need exact values.

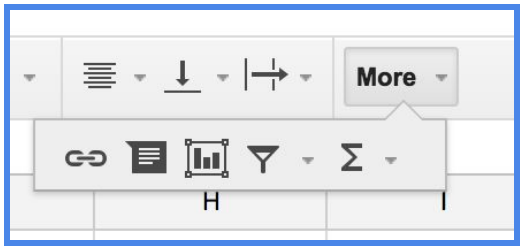


How are we going to get exact values without having to go through and count everything by hand? Turns out Google Sheets has a nifty Filter tool. We will use that in addition to the Explore tool to get our numbers.

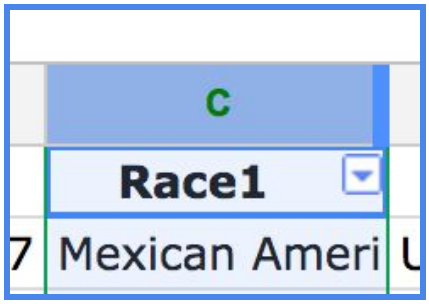
The icon for the Filter tool looks like a funnel.



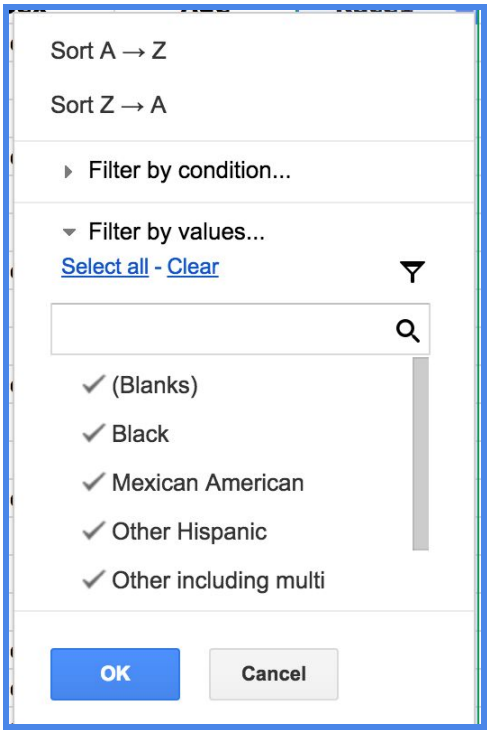
If you don't see it, it means your browser width is too small and the button is hidden under "More".



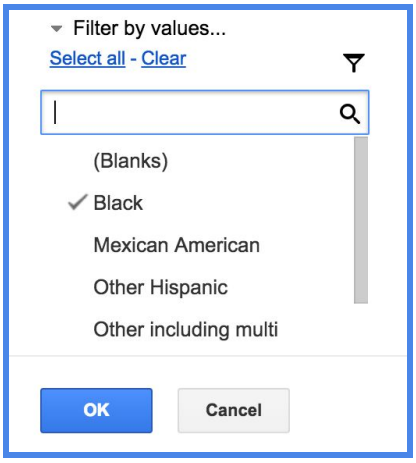
Select the Race column, then click on the Filter tool. A downwards arrow will appear next to the column label.



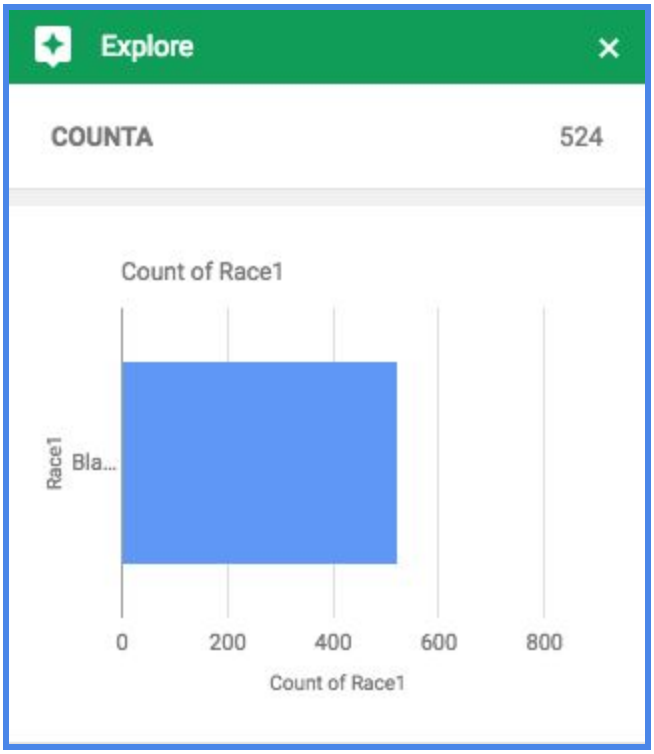
Click on the arrow button. You'll see a pop-up box that looks like the image below. All the different text values for the Race column are listed under "Filter by values". All of them have a checkmark next to it, meaning that all the categories are being included in the graph. Any rows that are NOT checked will not be included in an analysis or graph. Clicking on a checkmark will uncheck that category. Let's try it out.



Uncheck all of the values except for one. You can also click on the word "Clear" to uncheck everything, then check a single value. Below we have unchecked all of the categories except for Black.

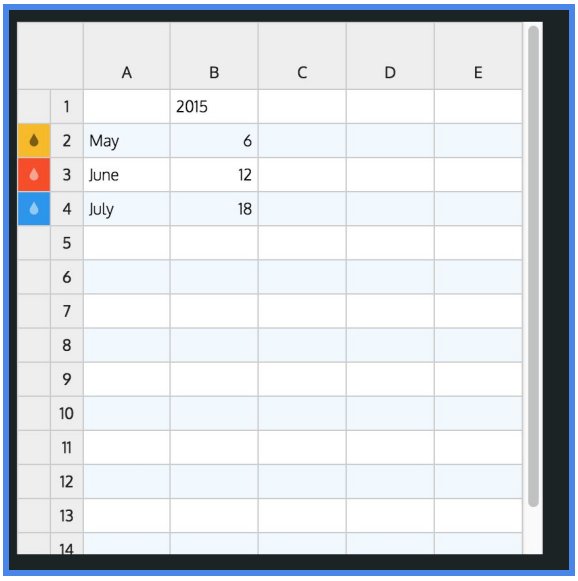


The Explore panel on the right will now show statistics for the filtered data, i.e. just for the category that is checked.



Here, remember COUNTA gives us the number of rows with our selected value for Race plus one for the column label. So the number of cases that have the value “Black” for Race is 523. This number will be different for you because you get a random sample from the EEPS NHANES data explorer.

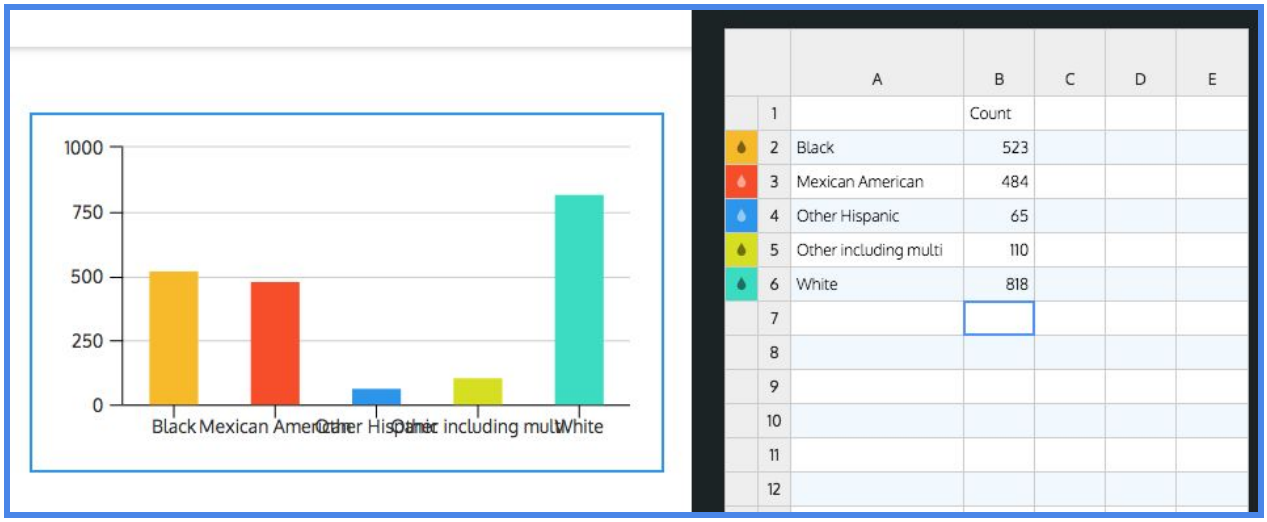
To enter this value into Venngage, drag a column chart from the Charts list in Venngage and double click on it to edit the chart.



	A	B	C	D	E
1		2015			
2	May	6			
3	June	12			
4	July	18			
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					

Column A here contains the label for the column on the graph and Column B contains the numeric values. Cell B1 contains the label for the series. (You can have multiple series in a column graph.) Change cell B1 to say “Count”. In cell A2, enter “Black”, and then “523” in B2.

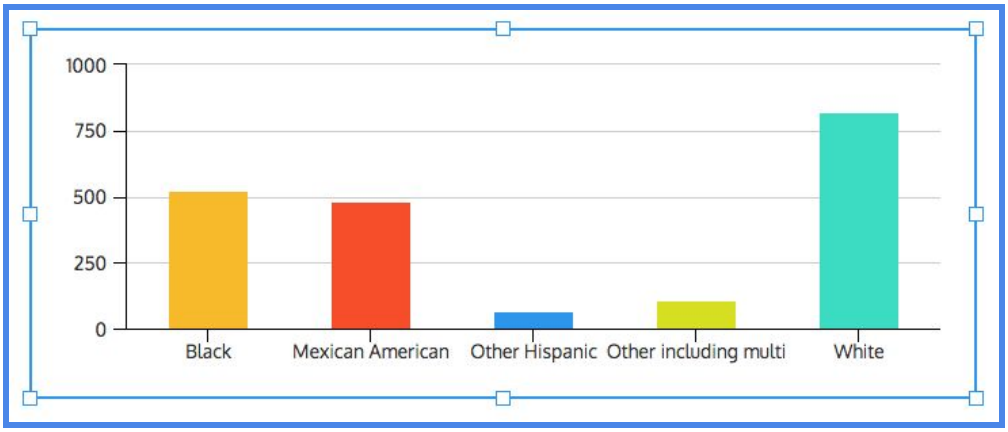
Go back to Google Sheets and use the Filter tool figure out the count for each category in the column Race, then transfer that number to Venngage. (Don’t forget to subtract 1 from COUNTA for the first row that contains labels!)



	A	B	C	D	E
1		Count			
2	Black	523			
3	Mexican American	484			
4	Other Hispanic	65			
5	Other including multi	110			
6	White	818			
7					
8					
9					
10					
11					
12					

Once you have all your numbers, verify that they add up to 2000, the total number of cases in our data.

In Venngage, save your changes to the chart. Since the labels are crowded, you can resize the chart by clicking and dragging any of the small boxes on the edges and corners of the chart.



You now have a column graph showing the distribution of race in the NHANES dataset sample.