STEM Literacy through Infographics

Visualizing Relationships with Scatter Plots

Part of a Series of Tutorials on using Google Sheets

Last updated: October 12, 2017

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.

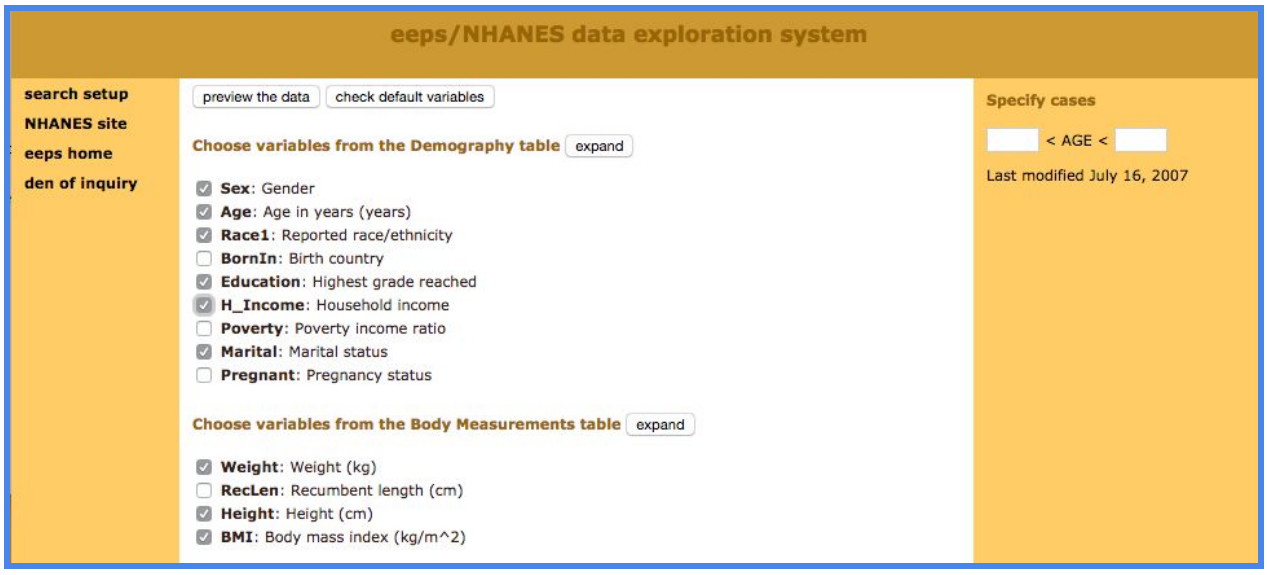
Scatter Plots

One way to investigate whether two variables are related to one another is to create a scatterplot, which plots one variable on the X-axis and one on the Y-axis, and shows you how they may be related.

Getting the Sample Data

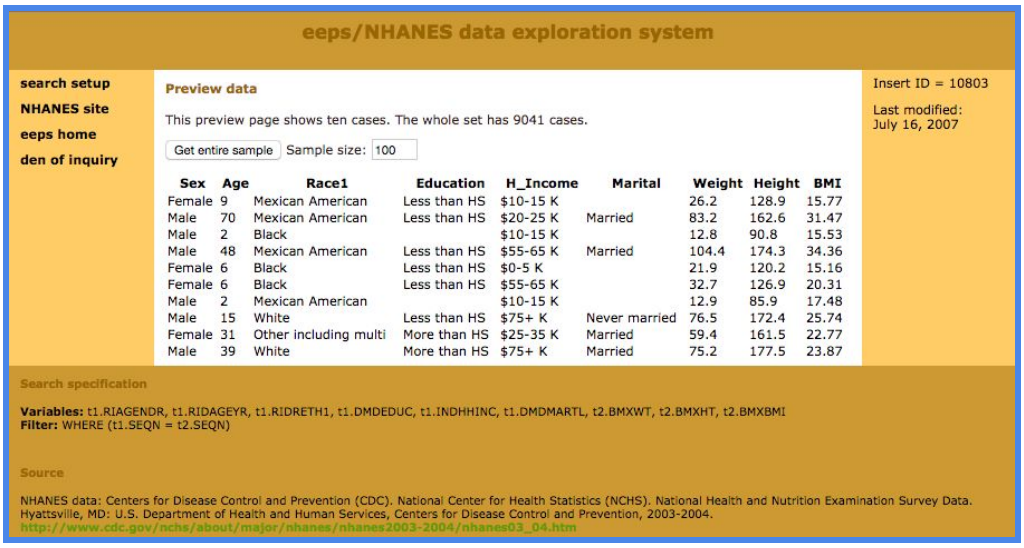
We're going to use, as our example, the National Health and Nutritional Examination Survey (NHANES) dataset. In the NHANES dataset, there are several thousand cases, each of which has data about a single person. NHANES data are highly encoded with numbers or abbreviations that stand in for readable words, 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. Where applicable, the units are next to each variable. For example, "Age" is in years, "Weight" is in kilograms, and "Height" is in centimeters.



For this tutorial, check the box for H_Income (Household Income) in the Demography section and keep the default variables already marked. 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 with a handful of sample cases 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

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

Insert ID = 10803

Last modified:
July 16, 2007

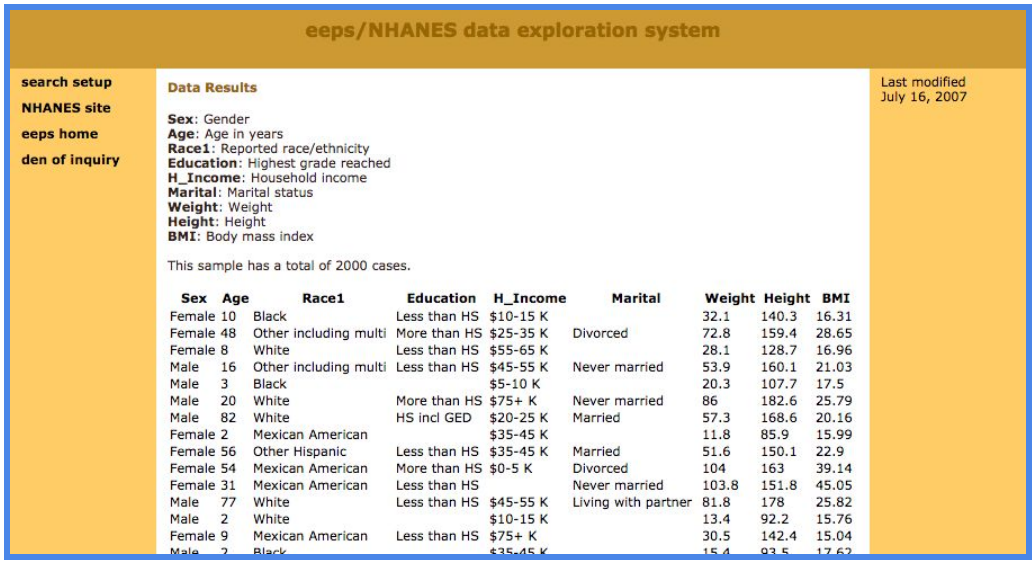
Search specification

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 the data are what you expected, enter 2000 in the “Sample size” text box and click on “Get entire sample”. We chose a sample of 2000 to demonstrate how Google Sheets can help you deal with a large amount of data easily. The interface will give you 2000 cases, as requested.



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.

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

Last modified
July 16, 2007

Select all the contents of the table, then copy and paste the data into a new Google Sheet file. It should paste correctly into each spreadsheet cell. If it doesn’t, make sure you have selected only 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; this is a common error. Now that our data are in Google Sheets, we are ready for analysis.

Reading the Data

Starting off, take a 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 (body mass index, a measure of body fat based on height and weight). (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, Household Income, and Marital Status have text values with limited options, as listed below.

Column	Possible 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 Status	Never Married, Married, Divorced, Widowed, Separated, Living with Partner, blank

Some of the things we might notice about the data are:

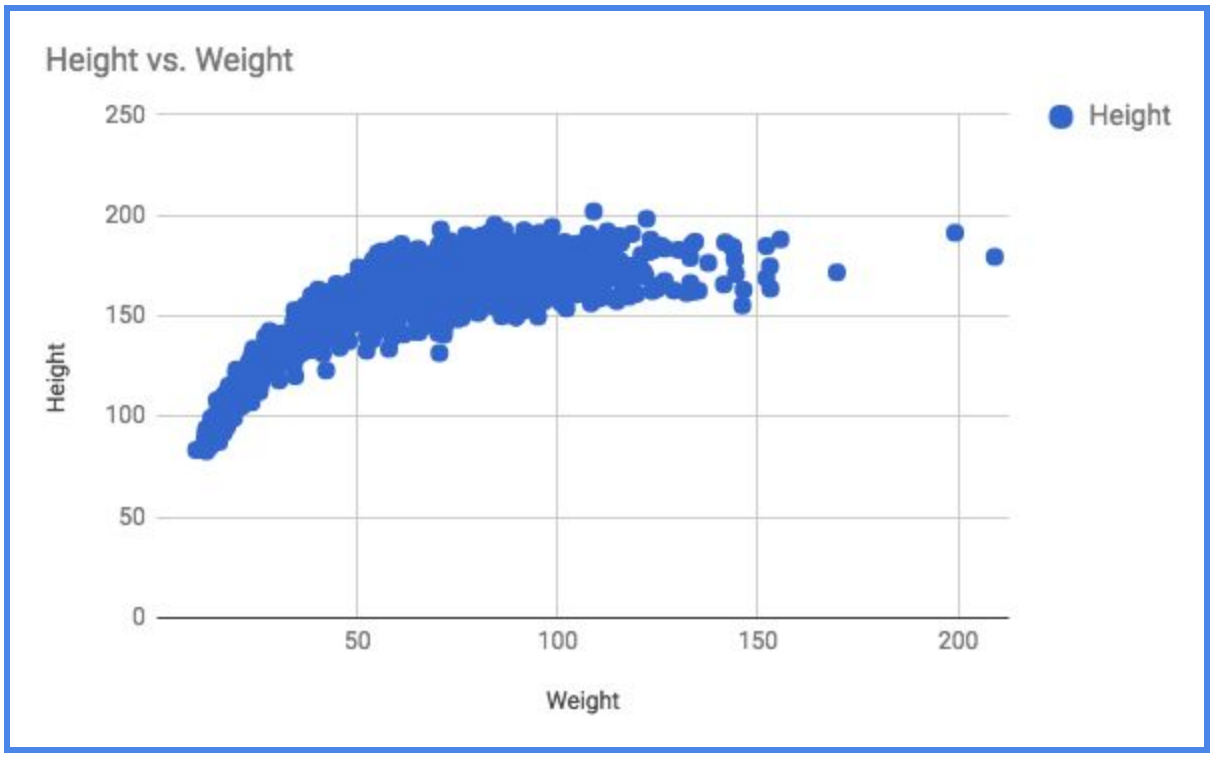
- The values for Household Income look like numerical values, but they are actually categories, each of which indicates a range of incomes.
- Education is generally blank for young children who are not old enough to have entered school yet.
- Marital Status is blank for people 13 years of age and under.
- Values for Height and Weight have decimals.

Two variables that we can easily imagine might have a relationship to one another are height and weight; we might expect that taller people are also heavier. We will create a scatterplot to see if this is the case.

Creating a Scatterplot in Google Sheets

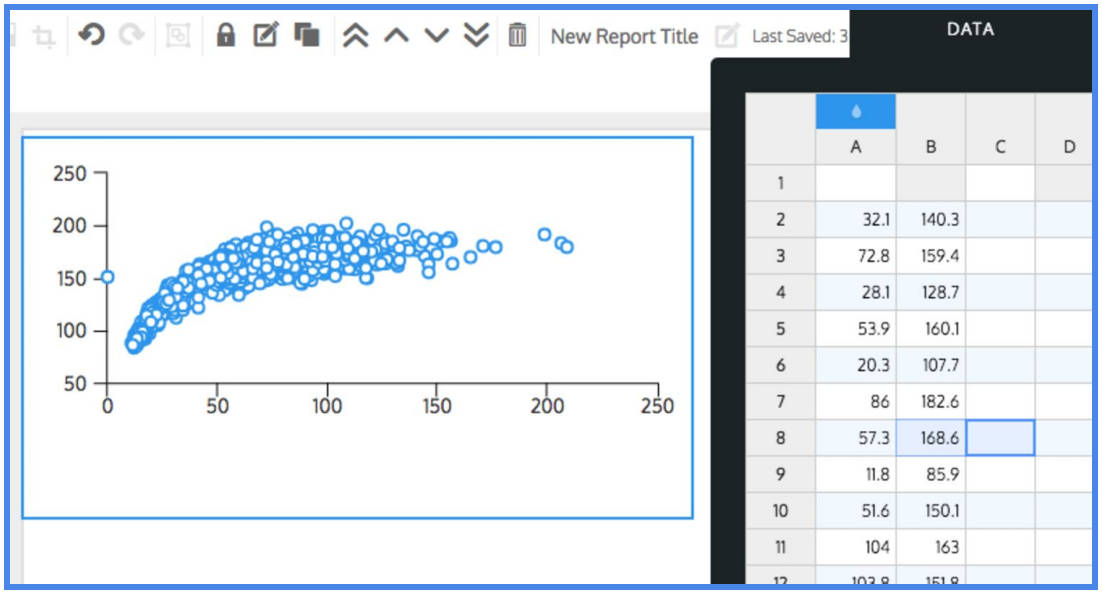
Click and drag over the column letters for Weight and Height to select them. (They should be next to each other because that's how we requested the data.) Then go to the the menu

Insert->Chart... and it will automatically recognize that the data in the columns are decimal numbers and create a scatter plot with an appropriate chart title and labeled axes.

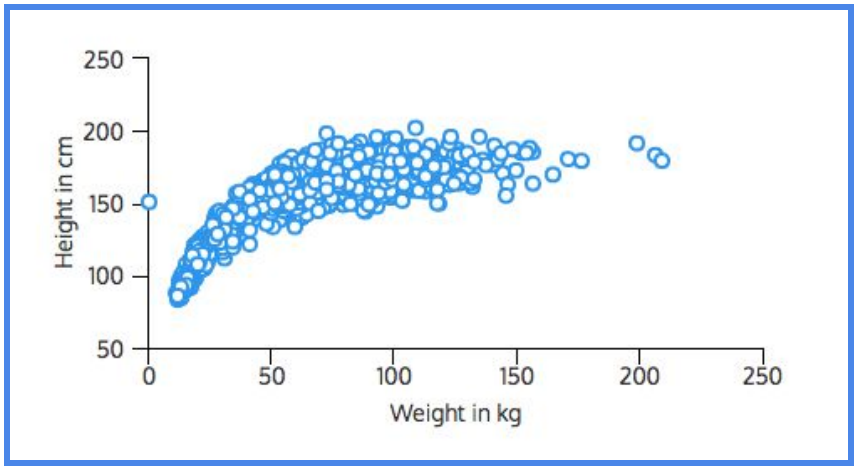


Creating a Scatter Plot in an Infographic Canvas Tool

Find the scatter plot option if it is available. In most tools, you will need to place the sample scatter plot on the canvas first. Go into the edit data view for the scatter plot. Getting to the edit data view will be different from one infographic canvas tool to the next. Our x-axis will be Weight, and our y-axis will be Height. In Google Sheets, select all the data in the column for Weight, not including the label on the first row. Paste this into the data table, starting at cell A2, leaving row 1 for the label if the canvas tool asks for it. Copy and paste the Height data into column B, starting at cell B2. In Venngage, our scatter plot will look something like this:



It's hard to read the graph without knowing what each axis represents, so make sure you label the axes. For the x-axis Title, type in "Weight in kg." Notice that it's not just "Weight" because that doesn't convey the units to the reader. Do the same with the y-axis and label it "Height in cm." If you don't remember the units, you can recheck on the data source. The kg and cm come from the EEPS NHANES data exploration tool web page.



Reflecting on the Data

Now that you've completed this walkthrough, here are some key points to take away:

- Weight and height increase together but there's a curve so it's not a linear relationship.
- There's a data point that has a height of about 150 cm but a weight of 0 kg. This means there's missing data for that person.