Applied Data Analysis

Module 5 Lab: Social Good Data Walkthrough

Learning Objectives

- Rearrange and sort data in Excel.
- Use Excel's COUNTIF function to turn categorical data into numerical counts.
- Use Excel's AVERAGEIFS function to compare averages across categories.
- Create and interpret histograms and scatterplots in Excel.
- Ensure that visualizations are properly labeled and readable.

Data Set

Mod5Lab.csv

Note: We're using a large data set for this lab, so make sure you copy and paste **all 3001 rows** of data into a new Excel Online worksheet.

What You'll Need

To complete the lab, you will need the online version of Microsoft Excel.

Overview

Imagine you're analyzing data for a nonprofit health organization. You've been tasked with showing whether soil contamination might be correlated with worsening health in three different regions. In this lab, you'll create several different graphs and tables to answer this question.

Exercise 1: Health Scores by Region

In our first exercise, we'll compare the general health scores of each region.

1. Open the data set in Excel Online. It contains data from 3000 randomly selected residents across 3 regions, testing each resident's home for soil contamination from industrial waste that might impact clean water (and thus health). The data also show generalized "health scores" for each resident, as well as their ages and the amount of time they spend outdoors. Here's a quick snapshot of the first few rows of the data set:

	A	B	C	D	E	F
1	id	region	contam	age	outdoors	health
2	1	Cleanville	0.8	58.9	3.1	8.1
3	2	Cleanville	3.9	81.9	1.5	5
4	3	Cleanville	1.2	63.3	5.1	6.9
5	4	Cleanville	1.2	29.1	5.7	8.7
6	5	Cleanville	1.9	58.3	6	7.2
7	6	Cleanville	0.3	49.8	5.8	8.4
8	7	Cleanville	3.3	30	4.3	7.7
9	8	Cleanville	0.7	55.4	3.8	7.9
10	9	Cleanville	0.6	76.6	3	7.1

Each row represents one person (there should be 3000 different people, which translates to 3001 different rows of data, counting the column titles). Here's what each column/variable represents:

id = a number representing the individual for ID purposes (i.e. Person 1, Person 2, etc.)
region = the town where that individual resides (either Cleanville, Monotowne, or Grimesburg)
contam = the "contamination score" for the soil at that individual's home, on a scale from 0 to 10 (where 0 = no contamination, and 10 = extremely contaminated)
age = the exact age of that individual, in years (as a decimal)
outdoors = the relative amount of time each individual spends outdoors in a given week, on a scale from 0 to 10 (where 0 = no time spent outdoors at all, and 10 = multiple hours spent outdoors every day)
health = the "health score" for that individual, on a scale from 0 to 10 (where 0 = extremely unhealthy, and 10 = extremely healthy)

2. Start out by creating a new mini-table off to the side of the data set, in columns H, I, J, K, and L. Include a row for each of the three towns (Cleanville, Monotowne, and Grimesburg) and columns for each town's number of residents, mean health score, minimum health score, and maximum health score. It should look like this:

	A	B	C	D	E	F	G	H	I	J	K	L
1	id	region	contam	age	outdoors	health			residents	mean health	min. health	max. health
2	1	Cleanville	0.8	58.9	3.1	8.1		**Cleanville**				
3	2	Cleanville	3.9	81.9	1.5	5		**Monotowne**				
4	3	Cleanville	1.2	63.3	5.1	6.9		**Grimesburg**				
5	4	Cleanville	1.2	29.1	5.7	8.7						
6	5	Cleanville	1.9	58.3	6	7.2						
7	6	Cleanville	0.3	49.8	5.8	8.4						

Here's a closer look at the mini-table itself:

	H	I	J	K	L
		residents	mean health	min. health	max. health
	Cleanville				
	Monotowne				
	Grimesburg				

You'll use this table to see how the relative health scores of each town compare to one another.

3. First, you want to see how evenly the data are distributed between the three towns. To find the number of residents from each town in this data, use Excel's COUNTIF function (which is way easier than counting through 3000 individual data points by hand, by the way).

   The syntax is **=COUNTIF(range, criteria)**. The range takes the form **firstcell:lastcell** (with a colon in between) and the criteria is whatever value you're looking for. **Note:** The criteria MUST be in quotation marks if you're looking for a text value (instead of a number).

   For example, to count the number of Cleanville residents in this data set, you want to look at column B, which shows the region names. The range runs from B2 (that's Person 1) all the way down to B3001 (that's Person 3000). (Remember, the entire data set has 3001 rows if you count the column titles.) Your criteria is "Cleanville" (in quotes because it's a text value, not a number).

   Here's what you should type into cell I2:

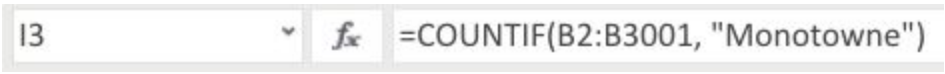
   Press Enter to get the count.

	H	I	J	K	L
		residents	mean health	min. health	max. health
	Cleanville	1000			
	Monotowne				
	Grimesburg				

   Our data set contains 1000 residents from Cleanville.

4. Repeat Step 3 for the number of residents in Monotowne and Grimesburg. The range is still B2:B3001 for both towns, but the criteria will change to either "Monotowne" or "Grimesburg" (again, don't forget the quotation marks).

   Here's the formula for Monotowne:



And here's the formula for Grimesburg:

Now your mini-table should look like this:

	H	I	J	K	L
		residents	mean health	min. health	max. health
2	Cleanville	1000			
3	Monotowne	1000			
4	Grimesburg	1000			

Excellent—the data set is evenly distributed between all three towns, with 1000 residents each.

5. Now find the mean (or average) health score for the people in each town. Use Excel's AVERAGEIFS function for this, which lets you find the average of cells that meet specific criteria (like finding the average health score in column F, but *only* for those people who also have "Cleanville" in column B).

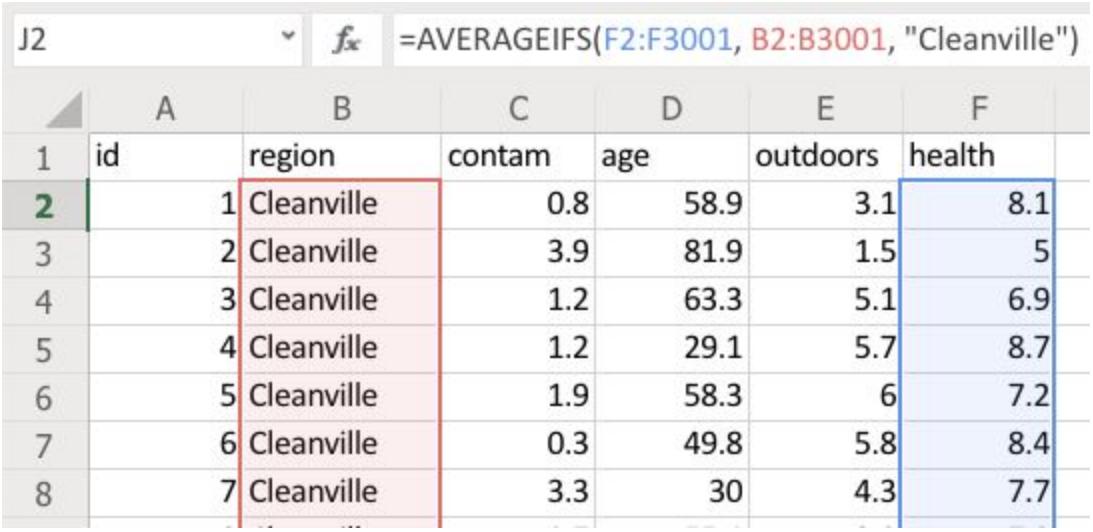
   The syntax is **=AVERAGEIFS(average_range, criteria_range, criteria)**. The average_range is the range of cells you want to pull the average from (that's column F in this case). The criteria_range is the range of cells that dictate which rows to include in the average (column B). The criteria is what needs to show up in that criteria range for Excel to count the row.

   Both ranges take the form **firstcell:lastcell** (with a colon in between). This time around, the average_range will be F2:F3001, because that covers everything in the "health" column. The criteria_range will be B2:B3001, because you want to narrow things down and only look at those residents with "Cleanville" in the "region" column. The criteria is "Cleanville" (and once again, it needs to be in quotation marks because it's text, not a number).

   So here's what you should type into cell J2 of your mini-table:

	H	I	J	K	L	M
		residents	mean health	min. health	max. health	
2	Cleanville	1000	=AVERAGEIFS(F2:F3001, B2:B3001, "Cleanville")			
3	Monotowne	1000				
4	Grimesburg	1000				

It pulls from the original data set like this (with fancy color-coding so you can see what's happening):



	A	B	C	D	E	F
1	id	region	contam	age	outdoors	health
2	1	Cleanville	0.8	58.9	3.1	8.1
3	2	Cleanville	3.9	81.9	1.5	5
4	3	Cleanville	1.2	63.3	5.1	6.9
5	4	Cleanville	1.2	29.1	5.7	8.7
6	5	Cleanville	1.9	58.3	6	7.2
7	6	Cleanville	0.3	49.8	5.8	8.4
8	7	Cleanville	3.3	30	4.3	7.7

Press Enter, and Excel will calculate the average health score from column F, but *only* for those folks with "Cleanville" in column B.

	H	I	J	K	L
		residents	mean health	min. health	max. health
	Cleanville	1000	6.9198		
	Monotowne	1000			
	Grimesburg	1000			

6. Repeat Step 5 to find the average health scores for Monotowne and Grimesburg. You can use AVERAGEIFS with the exact same syntax as before, but change the criteria to "Monotowne" or "Grimesburg" (both in quotes).

   Here's what you'll type for Monotowne (in cell J3):

   And here's what you'll type for Grimesburg (in cell J4):

   With all three mean health scores calculated, your table should look like this:

	H	I	J	K	L
		residents	mean health	min. health	max. health
	Cleanville	1000	6.9198		
	Monotowne	1000	7.7761		
	Grimesburg	1000	6.1113		

Now you can see how the three regions measure up. Monotowne had the highest average health score at about 7.7761, while Grimesburg did a bit worse with an average of 6.1113. Cleanville was right in the middle.

Note: In real-life data analysis, we would typically report on other summary statistics as well besides just the mean (like standard deviation, skewness, etc.). But we'll save those for another class.

7. It can also be useful to see what each region's lowest and highest health scores were. In column K, use Excel's MINIFS function to find the lowest (minimum) score for each region.

   The syntax is very similar to what you used for the AVERAGEIFS function: **=MINIFS(minimum_range, criteria_range, criteria)**. The minimum_range is the range of cells you want to pull the minimum from (that's column F again). The criteria_range is the range of cells that dictate which rows to include (column B again). The criteria is what needs to show up in that criteria range for Excel to count the row.

   So the minimum_range will be F2:F3001 from the "health" column. The criteria_range will be B2:B3001, because you still want to only look at those residents with "Cleanville" in the "region" column. The criteria is "Cleanville" (in quotes).

   Here's what you should type into cell K2 of your mini-table:

   Hit Enter.

	H	I	J	K	L
		residents	mean health	min. health	max. health
	Cleanville	1000	6.9198	1.3	
	Monotowne	1000	7.7761		
	Grimesburg	1000	6.1113		

   So the lowest health score for anyone in Cleanville was 1.3.

8. Repeat Step 7 and use the MINIFS function again to find the lowest score for the other two regions. Use the same ranges as the previous step, but change the criteria to “Monotowne” or “Grimesburg” (both in quotes).

   Here’s Monotowne’s minimum health score (in cell K3):

   And here’s Grimesburg’s minimum score (in cell K4):

   With those minimum health scores calculated, your table is looking good:

H	I	J	K	L
	residents	**mean health**	**min. health**	**max. health**
Cleanville	1000	6.9198	1.3	
Monotowne	1000	7.7761	2.8	
Grimesburg	1000	6.1113	1.3	

   Looks like Monotowne had a significantly higher minimum score of 2.8.

9. As you may have guessed, there’s also a MAXIFS function that you can use to find the *highest* (maximum) health score in each region.

   The syntax is pretty similar to the MINIFS function: **=MAXIFS(maximum_range, criteria_range, criteria)**. The maximum_range is the range of cells you want to pull the maximum from (column F again). The criteria_range is the range of cells that dictate which rows to include (column B again). The criteria is what needs to show up in that criteria range for Excel to count the row.

   So the maximum_range is F2:F3001 again from the “health” column. The criteria_range is B2:B3001 again, because you’re still looking at those residents with “Cleanville” in the “region” column. The criteria is “Cleanville” (in quotes).

   Type this into cell L2 of your mini-table to find the highest health score in Cleanville:

   Press Enter to calculate it.

H	I	J	K	L
	residents	mean health	min. health	max. health
Cleanville	1000	6.9198	1.3	10
Monotowne	1000	7.7761	2.8	
Grimesburg	1000	6.1113	1.3	

Unsurprisingly, the highest health score for this region was a 10, which is the highest possible score. It's probably a good guess that the other two regions will have at least one resident with a perfect 10 as well, but it's best to double check.

10. Repeat Step 9 for the other two regions, using MAXIFS to find the highest health score in each region. The data ranges won't change at all (still F2:F3001 and B2:B3001), but change the criteria to "Monotowne" or "Grimesburg" (both in quotes).

    Here's Monotowne's maximum health score (in cell L3):

    And here's Grimesburg's maximum health score (in cell L4):

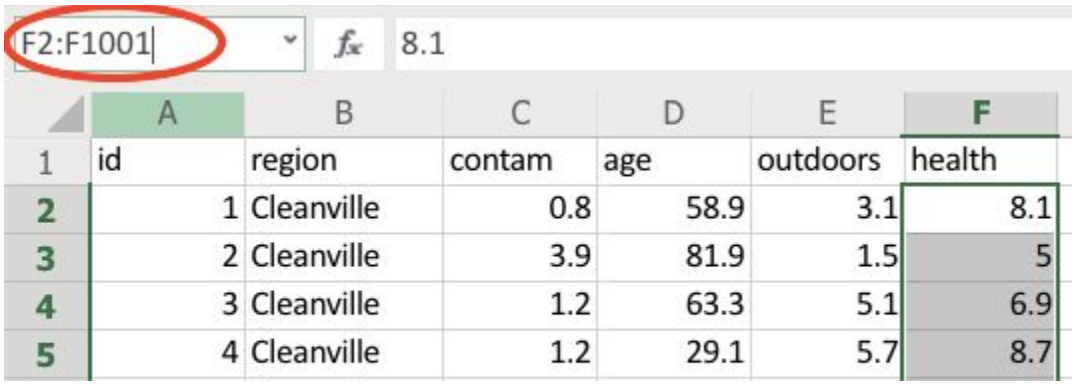
    Sure enough, all three towns had a max health score of 10.

H	I	J	K	L
	residents	mean health	min. health	max. health
Cleanville	1000	6.9198	1.3	10
Monotowne	1000	7.7761	2.8	10
Grimesburg	1000	6.1113	1.3	10

11. Based on this table, it's probably safe to say that Grimesburg is slightly less healthy than the other two regions. But let's see how those health scores play out when we create visualizations for them.

    First off, create a histogram to show the health score distribution in Cleanville only. Thankfully, the data are already sorted by town—in other words, all of the Cleanville residents are listed first (Person 1 through Person 1000 in the spreadsheet), followed by the Monotowne residents (Persons 1001–2000), and then the Grimesburg residents (Persons 2001–3000). (Note: If they weren't sorted already, you could click **Data > Filter** to bring up dropdown menus at the top of each column, which would let you sort all of the data by any particular column.)

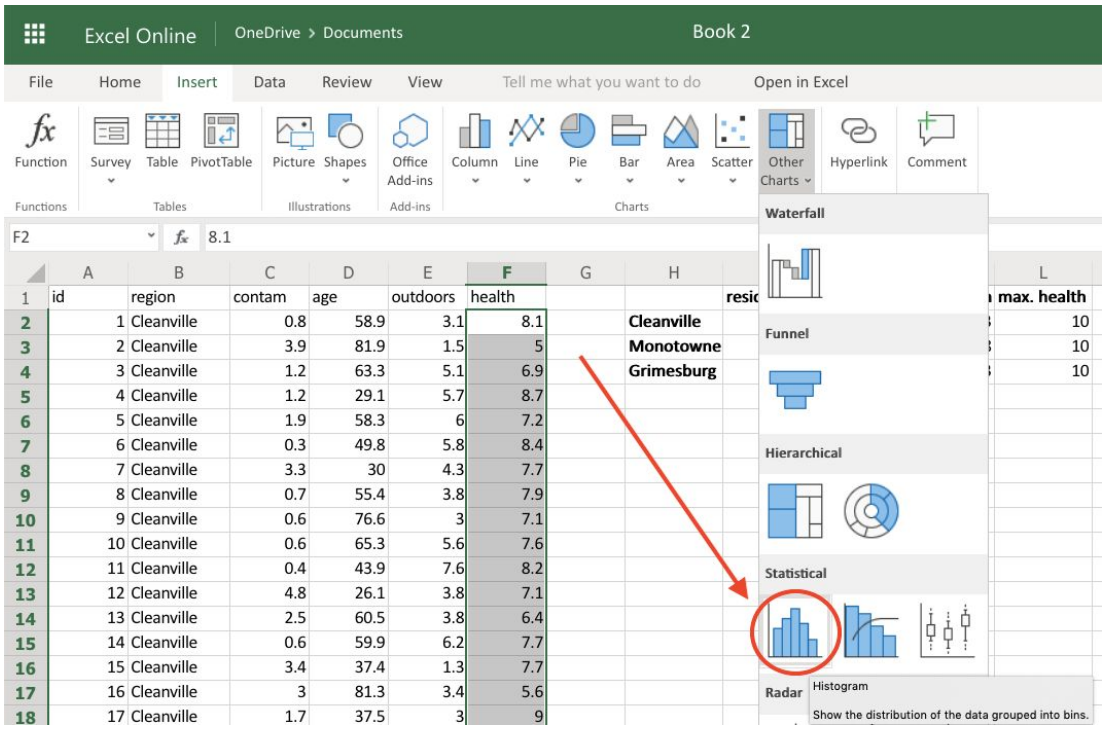
To select *only* the health scores for Person 1 to Person 1000 (the people who live in Cleanville), click into the name box in the upper-left corner and type in **F2:F1001**, like this:



	A	B	C	D	E	F
1	id	region	contam	age	outdoors	health
2	1	Cleanville	0.8	58.9	3.1	8.1
3	2	Cleanville	3.9	81.9	1.5	5
4	3	Cleanville	1.2	63.3	5.1	6.9
5	4	Cleanville	1.2	29.1	5.7	8.7

Hit Enter to highlight that range of values.

12. With cells F2:F1001 highlighted, click Insert > Other Charts > Histogram (it's typically the first icon under Statistical).



	A	B	C	D	E	F
1	id	region	contam	age	outdoors	health
2	1	Cleanville	0.8	58.9	3.1	8.1
3	2	Cleanville	3.9	81.9	1.5	5
4	3	Cleanville	1.2	63.3	5.1	6.9
5	4	Cleanville	1.2	29.1	5.7	8.7
6	5	Cleanville	1.9	58.3	6	7.2
7	6	Cleanville	0.3	49.8	5.8	8.4
8	7	Cleanville	3.3	30	4.3	7.7
9	8	Cleanville	0.7	55.4	3.8	7.9
10	9	Cleanville	0.6	76.6	3	7.1
11	10	Cleanville	0.6	65.3	5.6	7.6
12	11	Cleanville	0.4	43.9	7.6	8.2
13	12	Cleanville	4.8	26.1	3.8	7.1
14	13	Cleanville	2.5	60.5	3.8	6.4
15	14	Cleanville	0.6	59.9	6.2	7.7
16	15	Cleanville	3.4	37.4	1.3	7.7
17	16	Cleanville	3	81.3	3.4	5.6
18	17	Cleanville	1.7	37.5	3	9

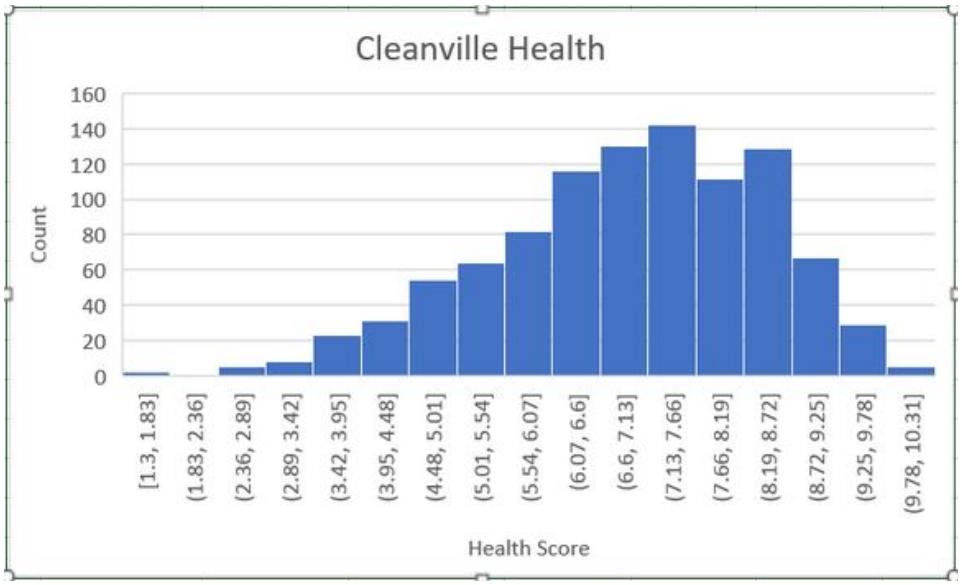
13. A histogram should pop up. Just like we did in the other labs, click directly on the chart and add some helpful labels by doing the following:

Horizontal axis title: **Chart Tools > Axis Titles > Primary Horizontal Axis Title > Edit Horizontal Axis Title**, then type "Health Score"

Vertical axis title: **Chart Tools > Axis Titles > Primary Vertical Axis Title > Edit Vertical Axis Title**, then type "Count"

Chart title: **Chart Tools > Chart Title > Edit Chart Title**, then type "Cleanville Health"

14. Your fancy new histogram should look like this:

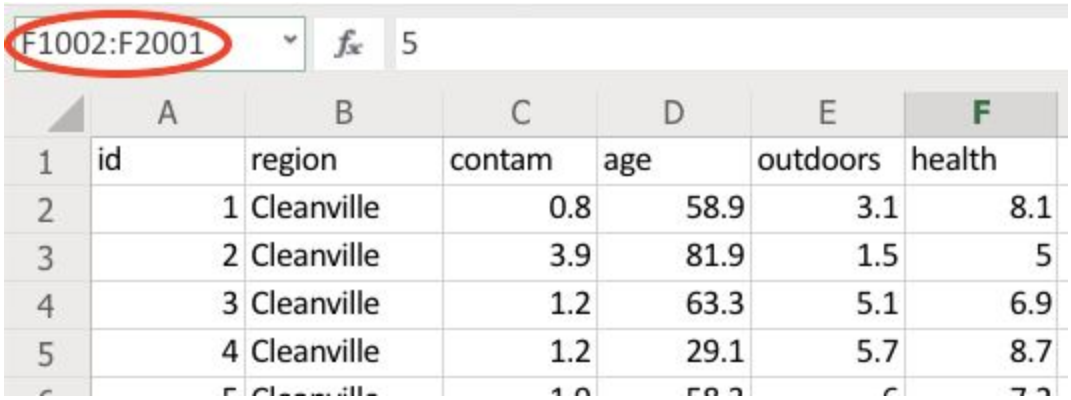


Here's how to read this thing: The horizontal axis along the bottom shows different ranges, or **bins**, of the health scores. For example, the bin for the tallest bar (the one labeled (7.13, 7.66]) represents all the people in Cleanville who had a health score between 7.13 and 7.66 (including 7.66 itself).

The vertical axis on the left shows the *number* of people who fall into each bin category. For example, the tallest bar is slightly above the marker for 140, which means that just over 140 people had a health score between 7.13 and 7.66.

Sadly, there's no way to change the bin size in Excel Online—the program automatically chooses the bins for you based on an even distribution of the data. (But if you have the desktop version of Excel, you can click Open in Excel in the ribbon and then double-click the graph to specify the bin width).

15. Now repeat Steps 11, 12, and 13 to create a histogram for Monotowne. To select *only* the health scores for Person 1001 to Person 2000 (the people who live in Monotowne), click into the name box in the upper-left corner and type in **F1002:F2001**, like this:



Press Enter. Note: The program will probably jump around a bit as it tries to select those cells. It will also probably move your view down to cell F1002 before it allows you to actually select that range of values. If necessary, **type F1002:F2001 into the name box a second time**, and press Enter again. (You can also just highlight cells F1002 to F2001 manually by clicking and dragging, if you prefer.) You should see this:

16. With cells F1002:F2001 highlighted, click Insert > Other Charts > Histogram (the first icon under Statistical), just like you did in Step 12.

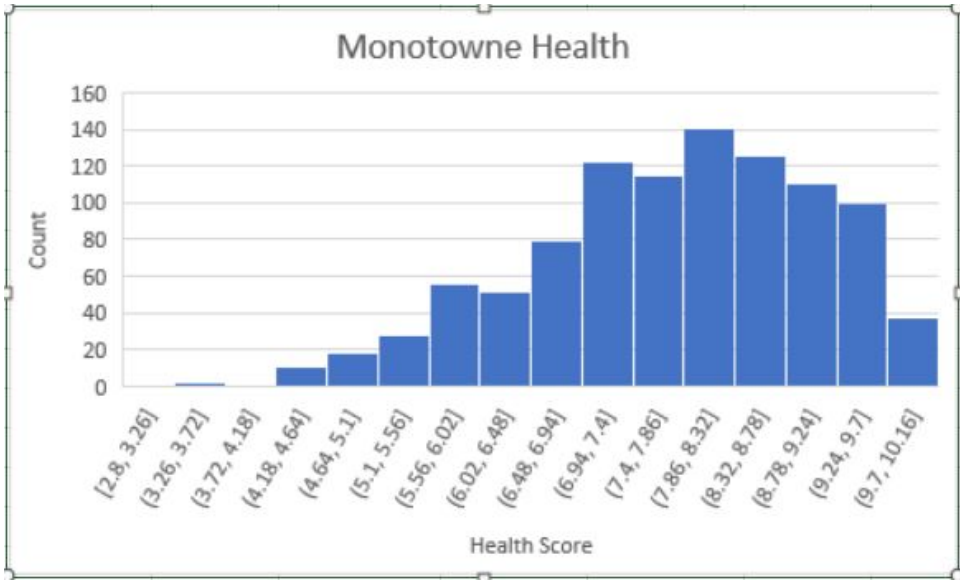
17. Click directly on the new histogram and add labels to it by doing the following:

    Horizontal axis title: **Chart Tools > Axis Titles > Primary Horizontal Axis Title > Edit Horizontal Axis Title**, then type “Health Score”

    Vertical axis title: **Chart Tools > Axis Titles > Primary Vertical Axis Title > Edit Vertical Axis Title**, then type “Count”

    Chart title: **Chart Tools > Chart Title > Edit Chart Title**, then type “Monotowne Health”

18. Your histogram should look like this now:



19. One more time: repeat Steps 11, 12, and 13 again to create a histogram for Grimesburg. To select *only* the health scores for Person 2001 to Person 3000 (the people who live in Grimesburg), click into the name box in the upper-left corner and type in **F2002:F3001**, like this:

	A	B	C	D	E	F
1	id	region	contam	age	outdoors	health
2	1	Cleanville	0.8	58.9	3.1	8.1
3	2	Cleanville	3.9	81.9	1.5	5
4	3	Cleanville	1.2	63.3	5.1	6.9
5	4	Cleanville	1.2	29.1	5.7	8.7

Press Enter. Again, the program will probably jump around a bit as it tries to select those cells. It will also probably move your view down to cell F2002 before it allows you to actually select that range of values, so **type F2002:F3001 into the name box a second time**, and press Enter again. (You can also just highlight cells F2002 to F3001 manually by clicking and dragging, if you prefer.) You should see this:

	A	B	C	D	E	F
1998	1997	Monotowne	5.3	60.9	4.8	5.6
1999	1998	Monotowne	2.1	44.5	4	8.6
2000	1999	Monotowne	1.9	20.3	6	9.3
2001	2000	Monotowne	1.5	45	5	9
2002	2001	Grimesburg	2.7	41.6	3.5	7.8
2003	2002	Grimesburg	0.8	33.5	7.6	8.6
2004	2003	Grimesburg	7.4	82	1.9	3.2
2005	2004	Grimesburg	9.5	45.1	5	3.4
2006	2005	Grimesburg	3.3	60.2	5.4	6.3

20. With cells F2002:F3001 highlighted, click Insert > Other Charts > Histogram (the first icon under Statistical), just like you did in Step 12.

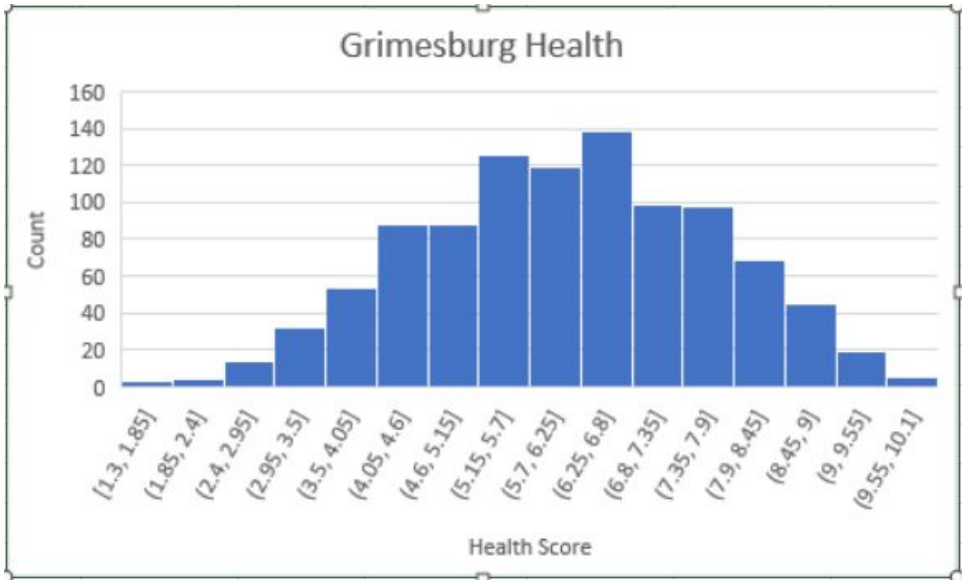
21. Click on the new histogram and add labels again:

    Horizontal axis title: **Chart Tools > Axis Titles > Primary Horizontal Axis Title > Edit Horizontal Axis Title**, then type “Health Score”

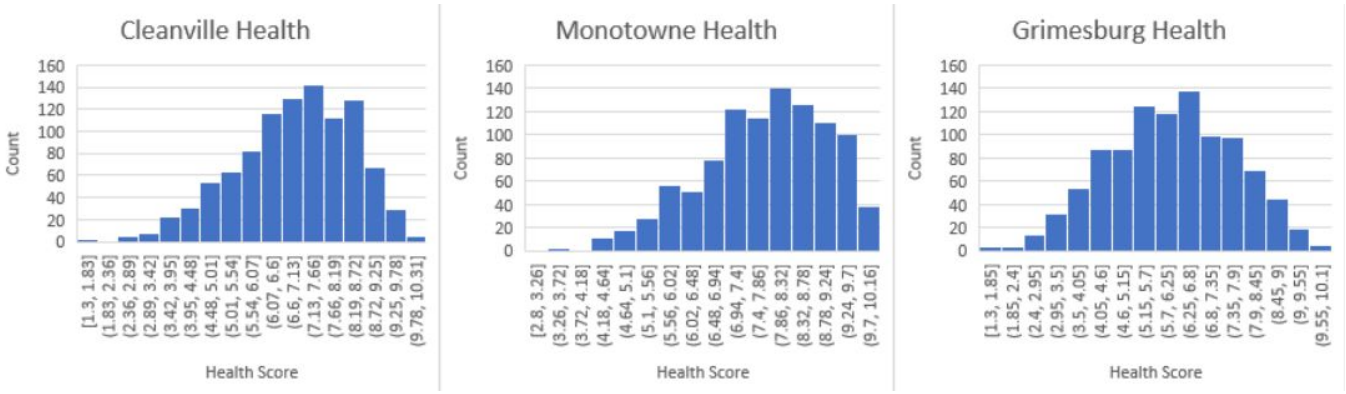
    Vertical axis title: **Chart Tools > Axis Titles > Primary Vertical Axis Title > Edit Vertical Axis Title**, then type “Count”

    Chart title: **Chart Tools > Chart Title > Edit Chart Title**, then type “Grimesburg Health”

22. And now you’ve got your third histogram:



23. Drag and resize the three histograms in Excel so that they're side by side, which makes comparing them easier:



The taller "bumps" show where most health scores for that region fall. Now you can see that Grimesburg does indeed seem to be the least healthy of the three regions, because its "bump" is closer to the 5–6 range (remember, lower numbers = lower health). This supports the mean/average data you saw earlier, where Grimesburg had a lower average health score than the other regions. Monotowne seems to be the healthiest, since its "bump" is higher up the scale than the other regions.

Of course, because we're presumably dealing with only a *sample* of each region's population, we would need more sophisticated statistical tests to know if if these differences are "real" in the broader population. But this is a good place to start!

Exercise 2: Soil Contamination by Region

Now we'll look at the soil contamination data in each region, using basically the same steps we did in Exercise 1 to compare averages, minimums, maximums, and histograms.

1. Directly below the table you created in Exercise 1, create a new mini-table off to the side of the data set, in columns H, I, J, and K. Include a row for each of the three regions (Cleanville, Monotowne, and Grimesburg) and columns for each region's mean contamination score, minimum contamination score, and maximum contamination score. Like this:

	A	B	C	D	E	F	G	H	I	J	K	L
1	id	region	contam	age	outdoors	health			**residents**	**mean health**	**min. health**	**max. health**
2	1	Cleanville	0.8	58.9	3.1	8.1		**Cleanville**	1000	6.9198	1.3	10
3	2	Cleanville	3.9	81.9	1.5	5		**Monotowne**	1000	7.7761	2.8	10
4	3	Cleanville	1.2	63.3	5.1	6.9		**Grimesburg**	1000	6.1113	1.3	10
5	4	Cleanville	1.2	29.1	5.7	8.7						
6	5	Cleanville	1.9	58.3	6	7.2			**mean contam.**	**min. contam.**	**max. contam.**	
7	6	Cleanville	0.3	49.8	5.8	8.4		**Cleanville**				
8	7	Cleanville	3.3	30	4.3	7.7		**Monotowne**				
9	8	Cleanville	0.7	55.4	3.8	7.9		**Grimesburg**				

Close-up view:

	mean contam.	min. contam.	max. contam.
Cleanville			
Monotowne			
Grimesburg			

2. Like you did in Exercise 1, use Excel's AVERAGEIFS function to find the mean/average contamination score in the three regions. The syntax is **=AVERAGEIFS(average_range, criteria_range, criteria)**. (See Step 5 of Exercise 1 if you need a more detailed refresher on how the syntax works for this function.)

   This time, the average_range will be C2:C3001, because column C contains the "contam" data. The criteria_range will be B2:B3001, because you want to narrow things down and only look at the residents of one particular region at a time. The criteria is either "Cleanville" or "Monotowne" or "Grimesburg" (in quotes).

   To get the mean contamination score in Cleanville, here's what you'll type in cell I7 of your new mini-table (and then press Enter):

   For the mean contamination score in Monotowne, type this into cell I8 and press Enter:

   For the mean contamination score in Grimesburg, type this into cell I9 and press Enter:

   And with your mean/average contamination scores filled in, your table should look like this:

	mean contam.	min. contam.	max. contam.
Cleanville	2.7683		
Monotowne	2.7855		
Grimesburg	4.3205		

   It appears that Grimesburg (true to its name) has a higher average amount of soil contamination than the other towns.

3. In the next column of your mini-table, use Excel's MINIFS function to find the lowest (minimum) score for each region, exactly like you did in Exercise 1. (See Step 7 of Exercise 1 for a refresher on how this function works.)

   Basically, you'll type in **=MINIFS(C2:C3001, B2:B3001, "Region")** every time, but swap out **Region** for the name of the actual town.

   Type this into cell J7 to get the minimum contamination score in Cleanville:

   Type this into cell J8 to get the minimum contamination score in Monotowne:

   Type this into cell J9 to get the minimum contamination score in Grimesburg:

   Now your table should look like this:

	mean contam.	min. contam.	max. contam.
Cleanville	2.7683	0	
Monotowne	2.7855	0	
Grimesburg	4.3205	0	

   So all three towns had at least one resident with a contamination score of 0.

4. In the last column of your mini-table, use Excel's MAXIFS function to find the highest (maximum) score for each region, exactly like you did in Exercise 1. (If you need a refresher on how the syntax for this function works, go back and look at Step 9 of Exercise 1.)

   Long story short: you'll type in **=MAXIFS(C2:C3001, B2:B3001, "Region")** every time, but swap out **Region** for the name of the actual town.

   Type this into cell K7 to get the max contamination score in Cleanville:

   Type this into cell K8 to get the max contamination score in Monotowne:

And type this into cell K9 to get the max contamination score in Grimesburg:

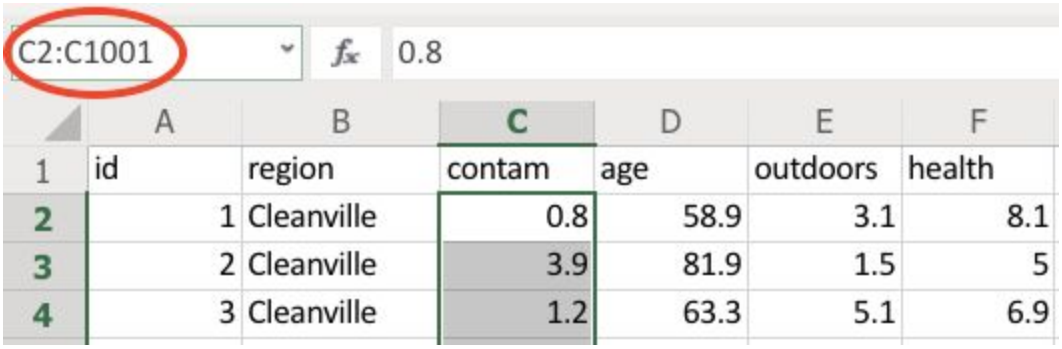
Your finished table should now look like so:

	mean contam.	min. contam.	max. contam.
Cleanville	2.7683	0	9.4
Monotowne	2.7855	0	8.9
Grimesburg	4.3205	0	10

Things aren't looking great for Grimesburg, are they?

5. Once again, creating a few histograms will make it easier to visualize the contamination in each region. Start with the contamination score distribution in Cleanville. Like we mentioned before, the data are already sorted by town—all of the Cleanville residents are listed first (Person 1 through Person 1000 in the spreadsheet), followed by the Monotowne residents (Persons 1001–2000), and then the Grimesburg residents (Persons 2001–3000).

   To select *only* the contamination scores for Person 1 to Person 1000 (the people who live in Cleanville), click into the name box in the upper-left corner and type in **C2:C1001**.



   Press Enter to highlight that entire range.

6. With cells C2:C1001 highlighted, click Insert > Other Charts > Histogram (the first icon under Statistical—to see what the icon looks like, scroll back up to Step 12 of Exercise 1).

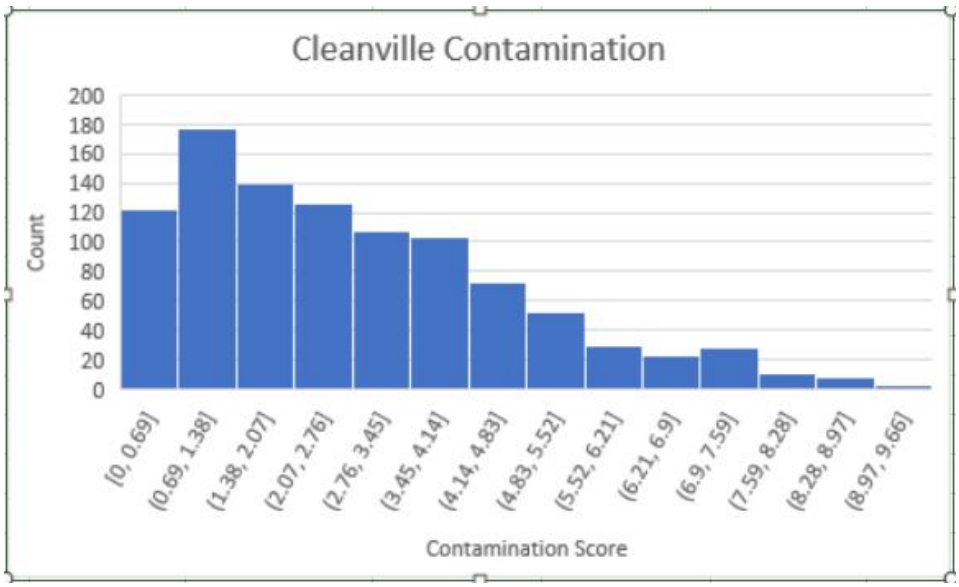
7. When the new histogram pops up, click directly on the chart and add labels:

Horizontal axis title: **Chart Tools > Axis Titles > Primary Horizontal Axis Title > Edit Horizontal Axis Title**, then type “Contamination Score”

Vertical axis title: **Chart Tools > Axis Titles > Primary Vertical Axis Title > Edit Vertical Axis Title**, then type “Count”

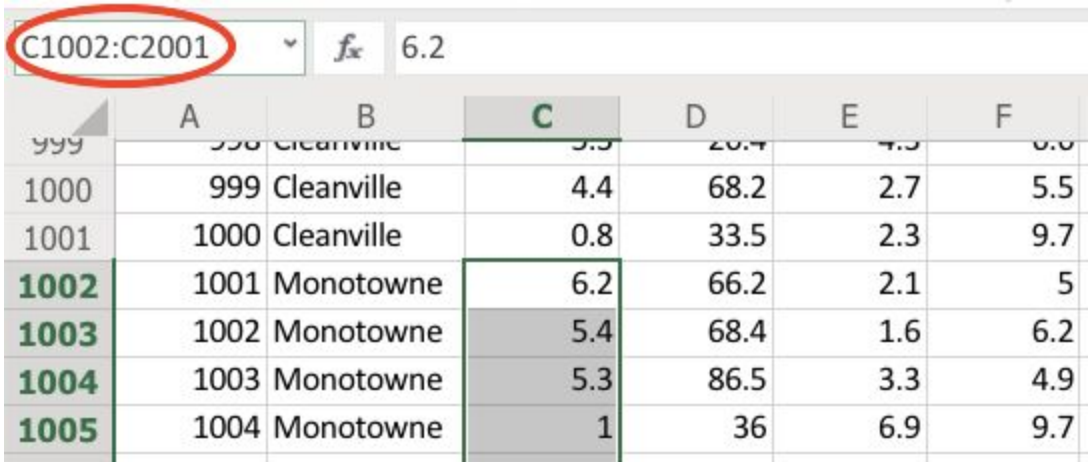
Chart title: **Chart Tools > Chart Title > Edit Chart Title**, then type “Cleanville Contamination”

8. Check out your new histogram:



Thankfully for Cleanville, the contamination scores are mostly clustered around the low end of the scale.

9. Now create a similar histogram for Monotowne. Select *only* the health scores for Person 1001 to Person 2000 (the people who live in Monotowne) by clicking into the name box in the upper-left corner and typing in **C1002:C2001**. Then press Enter. Note: The program will probably jump around and move your view down to cell C1002 before it allows you to actually select that range of values. If necessary, **type C1002:C2001 into the name box again and press Enter a second time**. (You can also just highlight cells C1002 to C2001 manually by clicking and dragging, if you prefer.)



	A	B	C	D	E	F
999	998	Cleanville	5.5	20.4	4.5	6.6
1000	999	Cleanville	4.4	68.2	2.7	5.5
1001	1000	Cleanville	0.8	33.5	2.3	9.7
1002	1001	Monotowne	6.2	66.2	2.1	5
1003	1002	Monotowne	5.4	68.4	1.6	6.2
1004	1003	Monotowne	5.3	86.5	3.3	4.9
1005	1004	Monotowne	1	36	6.9	9.7

10. With cells C1002:C2001 highlighted, click Insert > Other Charts > Histogram (the first icon under Statistical—to see what the icon looks like, scroll back up to Step 12 of Exercise 1).

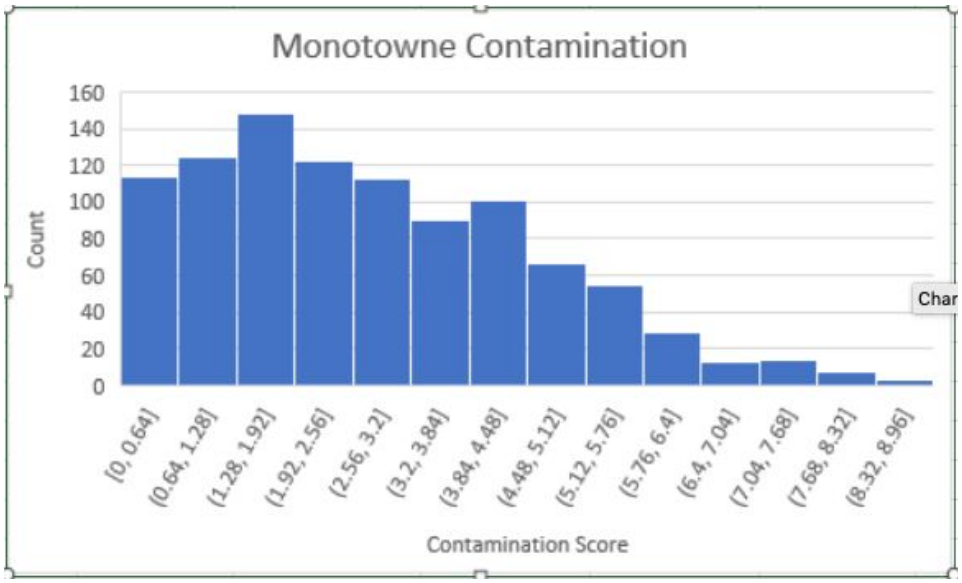
11. When the new histogram pops up, click directly on the chart and add labels:

    Horizontal axis title: **Chart Tools > Axis Titles > Primary Horizontal Axis Title > Edit Horizontal Axis Title**, then type "Contamination Score"

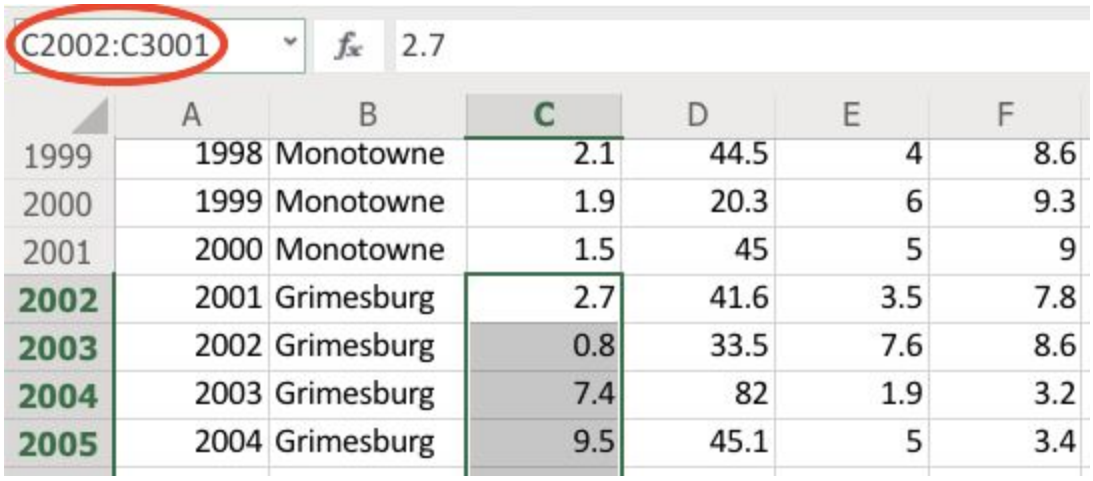
    Vertical axis title: **Chart Tools > Axis Titles > Primary Vertical Axis Title > Edit Vertical Axis Title**, then type "Count"

    Chart title: **Chart Tools > Chart Title > Edit Chart Title**, then type "Monotowne Contamination"

12. Your second histogram is all set!



13. One more histogram for Grimesburg. Select *only* the health scores for Person 2001 to Person 3000 (the people who live in Grimesburg) by clicking into the name box in the upper-left corner and typing in **C2002:C3001**. Then press Enter. Note: The program will probably jump around and move your view down to cell C2002 before it allows you to actually select that range of values. If necessary, **type C2002:C3001 into the name box again and press Enter a second time**. (You can also just highlight cells C2002 to C3001 manually by clicking and dragging, if you prefer.)



	A	B	C	D	E	F
1999	1998	Monotowne	2.1	44.5	4	8.6
2000	1999	Monotowne	1.9	20.3	6	9.3
2001	2000	Monotowne	1.5	45	5	9
2002	2001	Grimesburg	2.7	41.6	3.5	7.8
2003	2002	Grimesburg	0.8	33.5	7.6	8.6
2004	2003	Grimesburg	7.4	82	1.9	3.2
2005	2004	Grimesburg	9.5	45.1	5	3.4

14. With cells C2002:C3001 highlighted, click Insert > Other Charts > Histogram (the first icon under Statistical—to see what the icon looks like, scroll back up to Step 12 of Exercise 1).

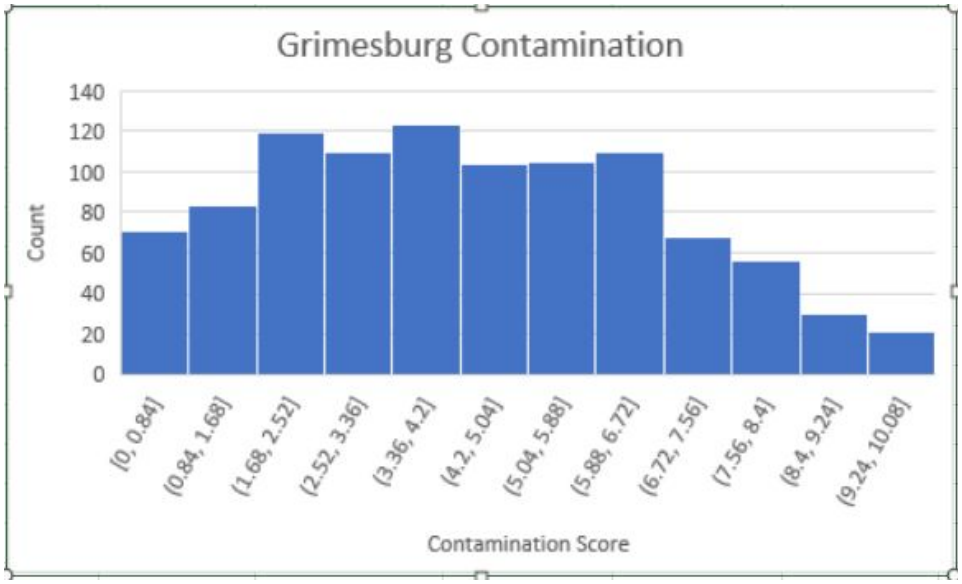
15. You know what's next, right? Click on the chart and do the following to add helpful labels:

    Horizontal axis title: **Chart Tools > Axis Titles > Primary Horizontal Axis Title > Edit Horizontal Axis Title**, then type "Contamination Score"

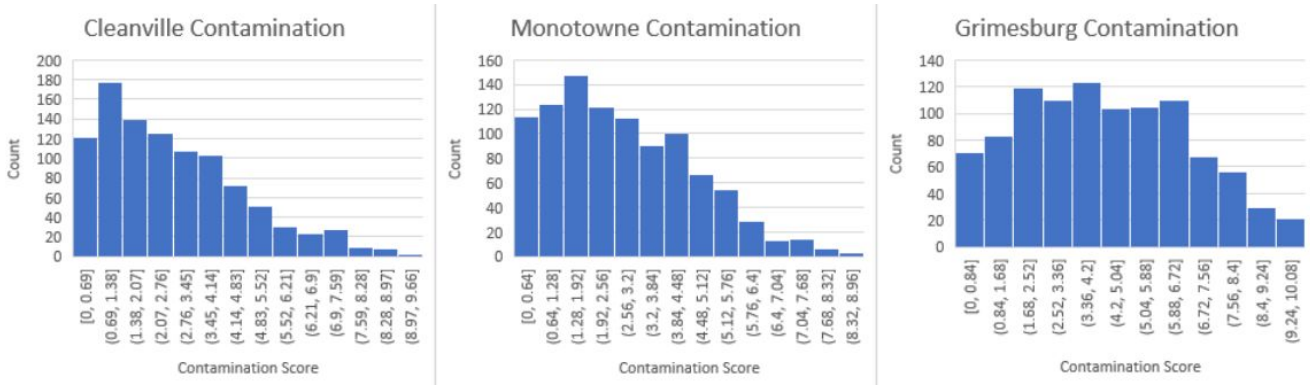
    Vertical axis title: **Chart Tools > Axis Titles > Primary Vertical Axis Title > Edit Vertical Axis Title**, then type "Count"

    Chart title: **Chart Tools > Chart Title > Edit Chart Title**, then type "Grimesburg Contamination"

16. And there's chart number three:



17. Finish up by dragging and resizing the three histograms next to each other.



Now you can see that soil contamination is fairly similar in both Cleanville and Monotowne—both towns had a lot of low contamination scores. But Grimesburg appears to have higher contamination scores overall. It's possible that these higher soil contamination levels are at least partly responsible for the lower health scores in Grimesburg, which we saw in Exercise 1.

Exercise 3: Age by Region

It definitely seems logical that the higher soil contamination in Grimesburg is influencing the region's lower health scores, but there are a couple other factors that might be contributing as well. For example, what if there are just more elderly people represented in Grimesburg than the other cities? A larger number of older people could potentially explain the health difference, since the elderly tend to be less healthy than the young. Let's create a quick table to see if this is the case.

1. Directly below the tables you created in Exercises 1 and 2, create a third mini-table in columns H, I, J, and K. Include a row for each of the three regions (Cleanville, Monotowne, and

Grimesburg) and columns for each region's mean age, minimum age, and maximum age. Like this:

	A	B	C	D	E	F	G	H	I	J	K	L
1	id	region	contam	age	outdoors	health			residents	mean health	min. health	max. health
2	1	Cleanville	0.8	58.9	3.1	8.1		Cleanville	1000	6.9198	1.3	10
3	2	Cleanville	3.9	81.9	1.5	5		Monotowne	1000	7.7761	2.8	10
4	3	Cleanville	1.2	63.3	5.1	6.9		Grimesburg	1000	6.1113	1.3	10
5	4	Cleanville	1.2	29.1	5.7	8.7						
6	5	Cleanville	1.9	58.3	6	7.2			mean contam.	min. contam.	max. contam.	
7	6	Cleanville	0.3	49.8	5.8	8.4		Cleanville	2.7683	0	9.4	
8	7	Cleanville	3.3	30	4.3	7.7		Monotowne	2.7855	0	8.9	
9	8	Cleanville	0.7	55.4	3.8	7.9		Grimesburg	4.3205	0	10	
10	9	Cleanville	0.6	76.6	3	7.1						
11	10	Cleanville	0.6	65.3	5.6	7.6			mean age	min. age	max. age	
12	11	Cleanville	0.4	43.9	7.6	8.2		Cleanville				
13	12	Cleanville	4.8	26.1	3.8	7.1		Monotowne				
14	13	Cleanville	2.5	60.5	3.8	6.4		Grimesburg				

And here's a closer look at the new table:

	mean age	min. age	max. age
Cleanville			
Monotowne			
Grimesburg			

2. Use Excel's AVERAGEIFS function—exactly you did in Exercises 1 and 2—to find the mean/average age in each of the three regions. Once again, the syntax is **=AVERAGEIFS(average_range, criteria_range, criteria)**. (See Step 5 of Exercise 1 if you need a detailed refresher on how the syntax works for this function.)

   This time, the average_range is D2:D3001, because the residents' ages are all in column D. The criteria_range is B2:B3001 as usual, because that's the "region" column. The criteria is either "Cleanville" or "Monotowne" or "Grimesburg" (in quotes).

   To get the mean age in Cleanville, type this into cell I12 of your new mini-table and press Enter:

   Here's what you'll type into cell I13 to find the mean age in Monotowne:

   And here's what you'll type into cell I14 to find the mean age in Grimesburg:

Now your table should have all three averages:

	mean age	min. age	max. age
Cleanville	52.0368		
Monotowne	53.3964		
Grimesburg	49.1465		

3. In the second column of your mini-table, use Excel's MINIFS function to find the lowest (minimum) age for each region, exactly like you did in Exercises 1 and 2. (See Step 7 of Exercise 1 for a refresher on how this function works.)

   You'll essentially type in **=MINIFS(D2:D3001, B2:B3001, "Region")** every time, but swap out **Region** for either "Cleanville" or "Monotowne" or "Grimesburg" (in quotes).

   Minimum age in Cleanville:

   Minimum age in Monotowne:

   Minimum age in Grimesburg:

   Table with all minimums filled in:

	mean age	min. age	max. age
Cleanville	52.0368	18.1	
Monotowne	53.3964	18.4	
Grimesburg	49.1465	18.1	

4. In the last column of your mini-table, use Excel's MAXIFS function to find the highest (maximum) age for each region, exactly like you did in Exercises 1 and 2. (See Step 9 of Exercise 1 for further details on how this function works.)

You'll essentially type in **=MAXIFS(D2:D3001, B2:B3001, "Region")** every time, but swap out **Region** for either "Cleanville" or "Monotowne" or "Grimesburg" (in quotes).

Maximum age in Cleanville:

Maximum age in Monotowne:

Maximum age in Grimesburg:

Table with all maximums filled in:

	mean age	min. age	max. age
Cleanville	52.0368	18.1	101.8
Monotowne	53.3964	18.4	100.9
Grimesburg	49.1465	18.1	92.1

We don't need to make a histogram to see that Grimesburg's health problems aren't being caused by a larger number of older people. In fact, Grimesburg seems to skew a little *younger* than the other two towns, so age probably isn't a factor here.

Exercise 4: The Great Outdoors

One other factor that could be contributing to Grimesburg's lower health scores is the amount of time people are spending outdoors. It's possible that people who go outside less often are less healthy. Let's see if the data support this.

1. Directly below the tables you created in Exercises 1, 2, and 3, create one final mini-table in columns H, I, J, and K. Include a row for each of the three regions (Cleanville, Monotowne, and Grimesburg) and columns for each region's mean outdoors score, minimum outdoors score, and maximum outdoors score. Like this:

	A	B	C	D	E	F	G	H	I	J	K	L
1	id	region	contam	age	outdoors	health			residents	mean health	min. health	max. health
2	1	Cleanville	0.8	58.9	3.1	8.1		Cleanville	1000	6.9198	1.3	10
3	2	Cleanville	3.9	81.9	1.5	5		Monotowne	1000	7.7761	2.8	10
4	3	Cleanville	1.2	63.3	5.1	6.9		Grimesburg	1000	6.1113	1.3	10
5	4	Cleanville	1.2	29.1	5.7	8.7						
6	5	Cleanville	1.9	58.3	6	7.2			mean contam.	min. contam.	max. contam.	
7	6	Cleanville	0.3	49.8	5.8	8.4		Cleanville	2.7683	0	9.4	
8	7	Cleanville	3.3	30	4.3	7.7		Monotowne	2.7855	0	8.9	
9	8	Cleanville	0.7	55.4	3.8	7.9		Grimesburg	4.3205	0	10	
10	9	Cleanville	0.6	76.6	3	7.1						
11	10	Cleanville	0.6	65.3	5.6	7.6			mean age	min. age	max. age	
12	11	Cleanville	0.4	43.9	7.6	8.2		Cleanville	52.0368	18.1	101.8	
13	12	Cleanville	4.8	26.1	3.8	7.1		Monotowne	53.3964	18.4	100.9	
14	13	Cleanville	2.5	60.5	3.8	6.4		Grimesburg	49.1465	18.1	92.1	
15	14	Cleanville	0.6	59.9	6.2	7.7						
16	15	Cleanville	3.4	37.4	1.3	7.7			mean outdoors	min. outdoors	max. outdoors	
17	16	Cleanville	3	81.3	3.4	5.6		Cleanville				
18	17	Cleanville	1.7	37.5	3	9		Monotowne				
19	18	Cleanville	4.7	21.2	2	7.7		Grimesburg				

Here's a close-up:

	mean outdoors	min. outdoors	max. outdoors
Cleanville			
Monotowne			
Grimesburg			

2. Just like you did in the other exercises, use the AVERAGEIFS function to calculate the mean/average outdoors score in the three regions. As usual, the syntax is **=AVERAGEIFS(average_range, criteria_range, criteria)**. (See Step 5 of Exercise 1 if you need a detailed refresher on how the syntax works for this function.)

   This time around, the average_range is E2:E3001, because column E contains all the outdoors data. The criteria_range is still B2:B3001, because that's the "region" column. The criteria is either "Cleanville" or "Monotowne" or "Grimesburg" (in quotes).

   In the first cell of your mini-table, use this syntax for Cleanville's mean outdoors score:

   I17 | fx | =AVERAGEIFS(E2:E3001, B2:B3001, "Cleanville")

   Here's Monotowne's mean outdoors score:

   I18 | fx | =AVERAGEIFS(E2:E3001, B2:B3001, "Monotowne")

   And here's the mean outdoors score for Grimesburg:

The table now looks like this:

	mean outdoors	min. outdoors	max. outdoors
Cleanville	4.301		
Monotowne	4.4531		
Grimesburg	4.7854		

Interesting... it looks like the residents of all three towns spend a similar amount of time outdoors, on average.

3. Next up, use Excel's MINIFS function to find the lowest (minimum) outdoors score for each region, exactly like you did in the other exercises. (See Step 7 of Exercise 1 for a refresher on how this function works.)

   You can type in **=MINIFS(E2:E3001, B2:B3001, "Region")** every time, but swap out **Region** for either "Cleanville" or "Monotowne" or "Grimesburg" (in quotes).

   Minimum outdoors score in Cleanville:

   Minimum outdoors score in Monotowne:

   Minimum outdoors score in Grimesburg:

   Table with all minimums filled in:

	mean outdoors	min. outdoors	max. outdoors
Cleanville	4.301	0	
Monotowne	4.4531	0	
Grimesburg	4.7854	0	

Zeros across the board! Unsurprisingly, it looks like there are some people in each town who never go outside at all (this could be due to any number of factors: illness, etc.).

4. Maximum time! Use Excel's MAXIFS function to find the highest (maximum) outdoors score for each region, like in the other exercises. (See Step 9 of Exercise 1 for further details on how this function works.)

   This time, type **=MAXIFS(E2:E3001, B2:B3001, "Region")** for each cell, but swap out **Region** for either "Cleanville" or "Monotowne" or "Grimesburg" (in quotes).

   Maximum outdoors score in Cleanville:

   Maximum outdoors score in Monotowne:

   Maximum outdoors score in Grimesburg:

   And here's the final table with everything filled in:

	mean outdoors	min. outdoors	max. outdoors
Cleanville	4.301	0	9.4
Monotowne	4.4531	0	9.5
Grimesburg	4.7854	0	9.7

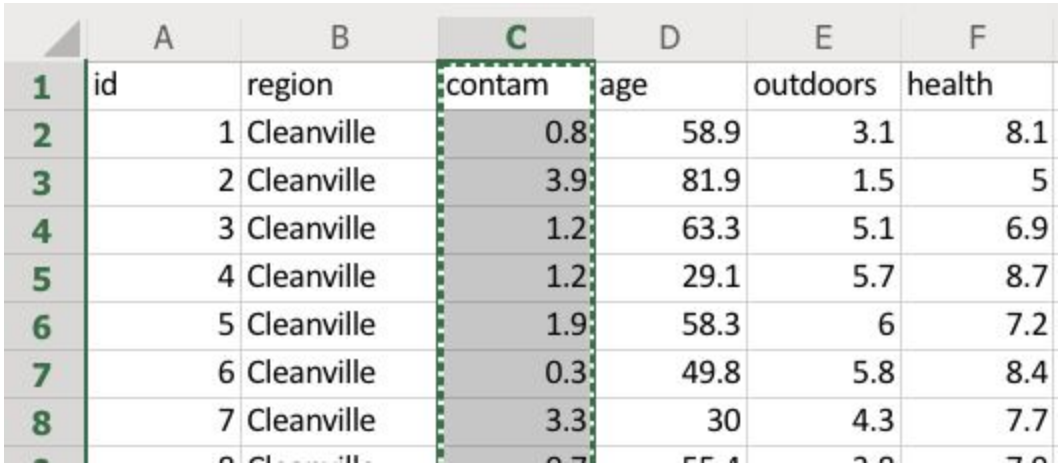
   There we go. The mean, minimum, and maximum were all very similar across the three regions. That means there was no significant difference in the amount of time people spent outside in the three different towns. So it's safe to conclude that Grimesburg's lower health scores aren't being caused by the amount of time people spend outdoors.

Exercise 5: Health versus Contamination

Now that we've shown that the lower health scores in Grimesburg probably aren't being affected by the age of the residents or the amount of time they spend in the great outdoors, we can create one final graph (a scatterplot this time) to see if there's any correlation between the soil contamination in Grimesburg and the lower health scores of the people who live there.

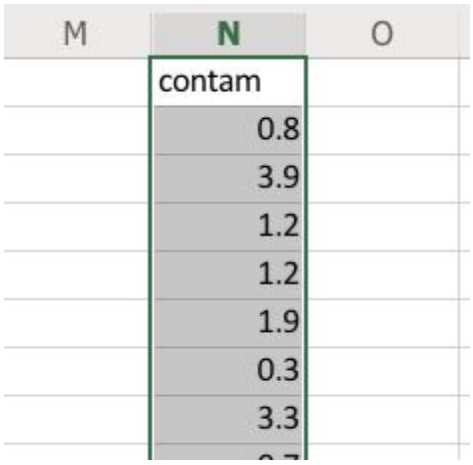
1. Before creating a scatterplot in Excel Online, you need to make sure the two columns you're interested in are directly adjacent to each other. (Note: You don't need to do this in the desktop versions of Excel.)

   Click on the "C" above column C to highlight the entire "contam" column. Then press Ctrl+C (or Command+C on a Mac) to copy all the data from the column.



	A	B	C	D	E	F
1	id	region	contam	age	outdoors	health
2	1	Cleanville	0.8	58.9	3.1	8.1
3	2	Cleanville	3.9	81.9	1.5	5
4	3	Cleanville	1.2	63.3	5.1	6.9
5	4	Cleanville	1.2	29.1	5.7	8.7
6	5	Cleanville	1.9	58.3	6	7.2
7	6	Cleanville	0.3	49.8	5.8	8.4
8	7	Cleanville	3.3	30	4.3	7.7

   Move over to column N, click the "N" at the top of the column to highlight it, and press Ctrl+V (or Command+V on a Mac) to paste the data you copied.



M	N	O
	contam	
	0.8	
	3.9	
	1.2	
	1.2	
	1.9	
	0.3	
	3.3	

   Hint: To make sure you've pasted *all* the data, look at the bottom-right corner of your browser window. It should show that your "count" is 3001, which means you successfully pasted all 3001 rows of data.

2. Now copy and paste the health score data from column F over into column O. To do this, click on the "F" at the top of column F, copy the data (press Ctrl+C or Command+C), and paste it into column O (press Ctrl+V or Command+V).

	A	B	C	D	E	F
1	id	region	contam	age	outdoors	health
2	1	Cleanville	0.8	58.9	3.1	8.1
3	2	Cleanville	3.9	81.9	1.5	5
4	3	Cleanville	1.2	63.3	5.1	6.9
5	4	Cleanville	1.2	29.1	5.7	8.7
6	5	Cleanville	1.9	58.3	6	7.2
7	6	Cleanville	0.3	49.8	5.8	8.4
8	7	Cleanville	3.3	30	4.3	7.7

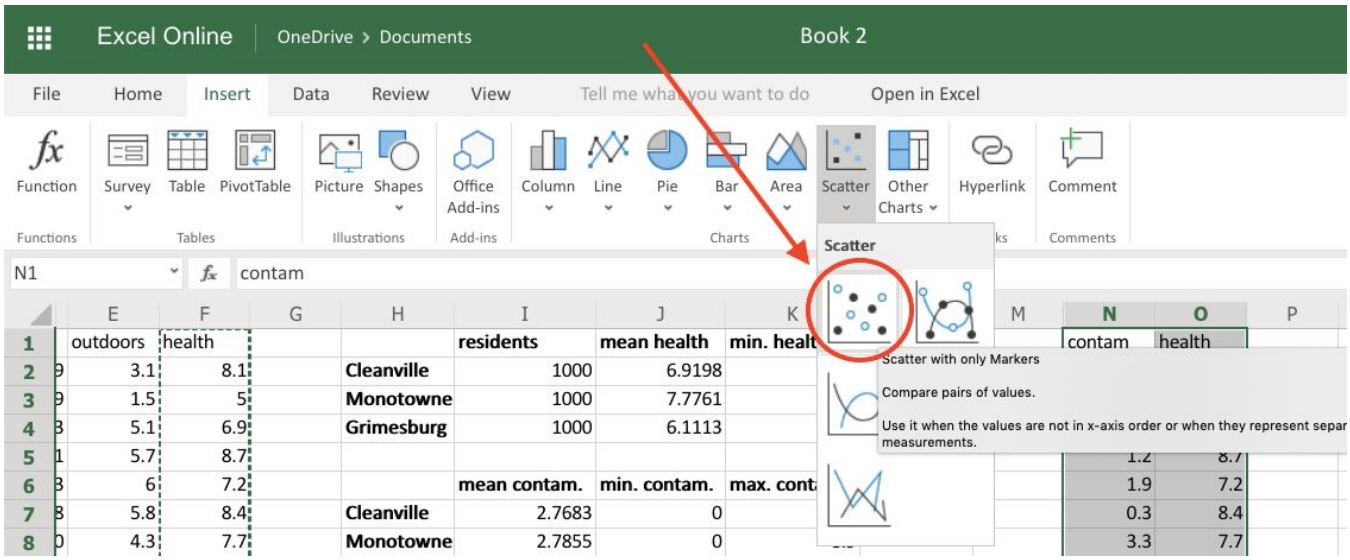
	N	O
	contam	health
	0.8	8.1
	3.9	5
	1.2	6.9
	1.2	8.7
	1.9	7.2
	0.3	8.4
	3.3	7.7

Now that you have the "contam" and "health" variables right next to each other, it's scatterplot time.

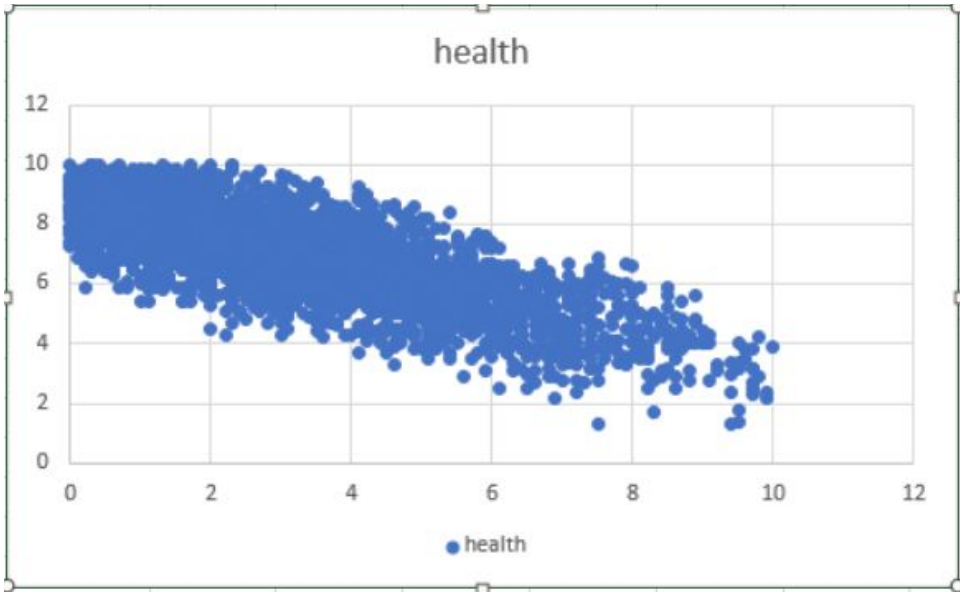
3. Highlight everything in columns N and O at the same time. To highlight both, click on the letter at the top of one column, hold the Shift key, and then click the letter at the top of the other column.

N	O
contam	health
0.8	8.1
3.9	5
1.2	6.9
1.2	8.7
1.9	7.2
0.3	8.4
3.3	7.7

4. With both columns highlighted, click Insert > Scatter > Scatter with only Markers (it's usually the first icon under Scatter—the one with dots and no lines).



5. A new scatterplot should pop up. Excel Online will show the left column along the horizontal *x*-axis (that's column N in this case) and the right column along the vertical *y*-axis (column O).



6. As usual, this graph is kind of hard to read like this. So let's add some labels. Click on the scatterplot and follow these steps:

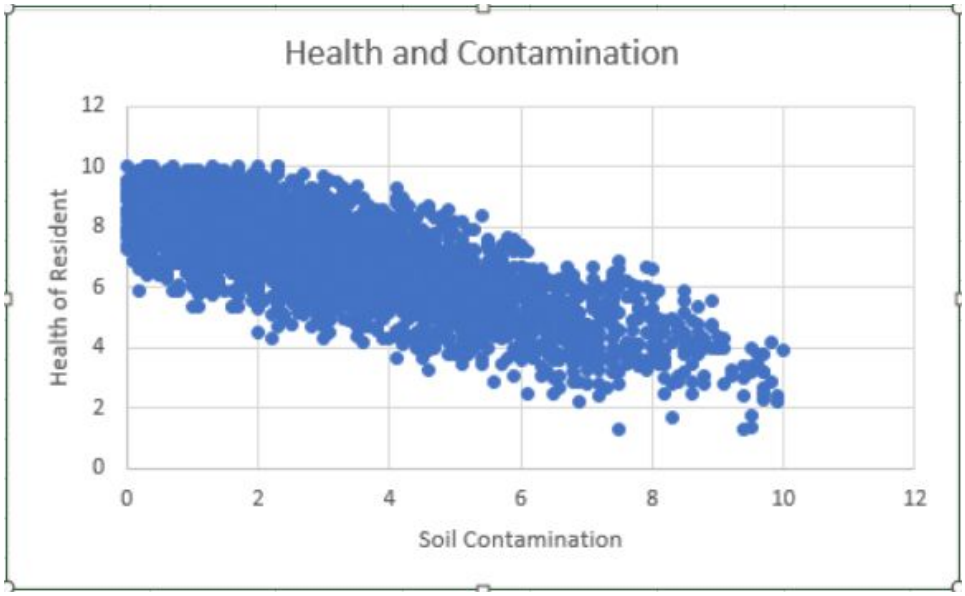
   Horizontal axis title: **Chart Tools > Axis Titles > Primary Horizontal Axis Title > Edit Horizontal Axis Title**, then type "Soil Contamination"

   Vertical axis title: **Chart Tools > Axis Titles > Primary Vertical Axis Title > Edit Vertical Axis Title**, then type "Health of Resident"

   Chart title: **Chart Tools > Chart Title > Edit Chart Title**, then type "Health and Contamination"

   You also don't need that annoying blue "health" legend at the bottom, so hide it by clicking **Chart Tools > Legend > None**.

7. Now your scatterplot should look like this:



Each blue dot represents one person from our data set. The higher the dot is, the better that person's health score is. The further to the right the dot is, the higher the soil contamination level at their home. Notice how the dots tend to move further down along the health axis as they move to the right along the contamination axis.

Reading the scatterplot, it's now clear that higher soil contamination levels are definitely correlated with lower health scores. In other words, higher contamination = lower health, which makes sense.

What's more, this correlation is true across the board, for all three regions. If we were to create separate scatterplots for each town—which we won't do right now—we'd see pretty much the same relationship: more soil contamination correlates to lower health scores in all three towns.

This is consistent with our explanation from earlier that one possible reason for Grimesburg's lower health scores is that Grimesburg had higher soil contamination scores (remember, Grimesburg's mean contamination score was about 4.3, whereas the other two towns were around 2.8).

By the way, the reason we say "correlates" instead of "causes" is that we don't have nearly enough data to say *for sure* that the soil contamination is *causing* lower health scores. It seems likely that the soil contamination is at least part of the reason for the lower health scores, but we'd need to analyze a lot more data using more sophisticated methods to really be certain. For now, we can say that higher soil contamination is correlated to lower health scores across the board.