



Applied Data Analysis

Module 4 Lab: Healthcare Data Walkthrough

Learning Objectives

- Rearrange and sort data in Excel.
- Use Excel's COUNTIF and COUNTIFS functions to turn categorical data into numerical counts.
- Use Excel's AVERAGE and AVERAGEIFS functions to compare averages across categories.
- Create and interpret bar charts and histograms in Excel.
- Find proportions and percentages in Excel.
- Ensure that visualizations are properly labeled and readable.

Data Set

Mod4Lab.csv

Note: We're using a large data set for this lab, so make sure you copy and paste **all 5426 rows** of data into a new Excel Online worksheet. (See Exercise 1 for special instructions on this.)

What You'll Need

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

Overview

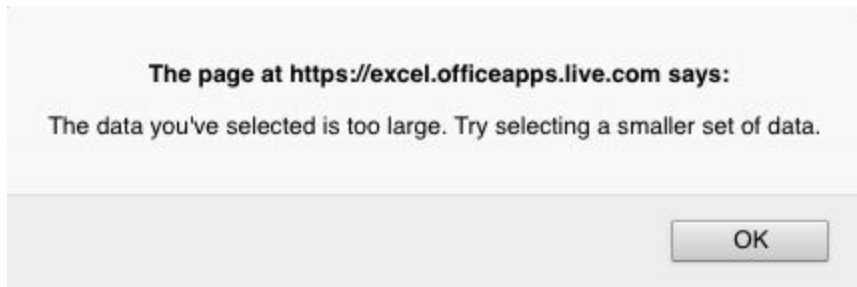
Imagine you're working as a data analyst for a hospital. Your boss wants you to use patient data to report on the effectiveness of a new, rapid surgical procedure and whether the hospital should replace the traditional procedure with the rapid version. In this lab, you'll prepare several different graphs and reports to answer this question.

Exercise 1: Traditional versus Rapid

We'll start things off by creating a quick, simple visualization to show how many patients got the traditional surgical procedure versus the new, rapid procedure.

1. Open the data set in Excel Online, which shows information about a surgical procedure for 5425 different medical patients at a hospital.

Note: You might get the following message if you try to copy and paste this large data set into Excel all at once:



If that happens, then carefully copy and paste the data about 2000 rows at a time until you have all 5426 rows in a new Excel Online worksheet (that's 5425 patients, plus the top row with the column titles—the bottom patient's ID number should be 5425). It's a little tedious, but it's very important to make sure you have the *entire* data set copied over. If you see any cells filled with pound signs (#####), then you can expand that column in the spreadsheet by clicking and dragging the line between two columns.

Here's a quick snapshot of the first few rows of the data set:

	A	B	C	D	E	F
1	id	duration (hours)	cost	inf	procedure	age
2	1	23	\$495.42	No Infection	Traditional	32.68
3	2	14	\$772.70	No Infection	Traditional	45.44
4	3	23	\$663.30	No Infection	Traditional	31.87
5	4	18	\$599.50	No Infection	Traditional	27.51
6	5	29	\$985.82	No Infection	Traditional	44.37
7	6	19	\$822.82	No Infection	Traditional	34.51
8	7	14	\$689.54	No Infection	Traditional	40.43
9	8	8	\$426.78	No Infection	Traditional	24.68
10	9	20	\$649.20	No Infection	Traditional	41.86
11	10	17	\$780.87	No Infection	Traditional	42.87
12	11	10	\$570.64	No Infection	Traditional	35.52

Each row represents one patient. Here's what each column/variable represents:

id = a number representing the individual for ID purposes (i.e. Patient 1, Patient 2, etc.)

duration = the amount of time that the patient's surgical procedure took, in hours

cost = the total cost of the procedure, in dollars, for that patient

inf = whether the patient got an infection in the hospital following the procedure (either No Infection or Infection)

procedure = the type of surgical procedure performed on that patient (either Traditional or Rapid)

age = the exact age of that individual, in years (as a decimal)

- Off to the side of the existing data set, create a new mini-table to show the counts for “Traditional” and “Rapid” from the procedure column. Your table should be in columns H and I, like this:

	A	B	C	D	E	F	G	H	I
1	id	duration (hours)	cost	inf	procedure	age			count
2	1	23	\$495.42	No Infection	Traditional	32.68		Traditional	
3	2	14	\$772.70	No Infection	Traditional	45.44		Rapid	
4	3	23	\$663.30	No Infection	Traditional	31.87			
5	4	18	\$599.50	No Infection	Traditional	27.51			
6	5	29	\$985.82	No Infection	Traditional	44.37			

- To find the number of patients in these two categories, use Excel’s COUNTIF function. The syntax is **=COUNTIF(range, criteria)**. The range takes the form **firstcell:lastcell** (with a colon in between) and the criteria is whatever value or text you’re looking for. **Note:** If the criteria is a text value instead of a number, it **MUST** be in quotation marks.

In this case, you’re counting the number of times “Traditional” shows up in the procedure column (that’s column E). So your range is E2:E5426. Note that even though there are 5425 patients in this data set (not 5426), the first row in the spreadsheet contains the column titles, which is why you’re starting at E2 and ending at E5426. Your criteria is “Traditional” (in quotation marks), so this is what you’ll type into cell I2 to get the first value in your mini-table:

I2 **=COUNTIF(E2:E5426, "Traditional")**

Hit Enter to get the count.

H	I
	count
Traditional	3334
Rapid	

There we go: 3334 patients got the Traditional surgical procedure.

- Use COUNTIF again to find the number of patients who got the Rapid procedure. Type this into cell I3 of your new mini-table:

I3 **=COUNTIF(E2:E5426, "Rapid")**

H	I
	count
Traditional	3334
Rapid	2091

Hint: To make sure you didn't miss anyone, add up those values:

$$3334 + 2091 = 5425$$

Sure enough, your data set had 5425 total patients, so you didn't miss any individuals or any categories.

- Now highlight everything in that new mini-table, including the category names (but do *not* use CTRL+A to highlight the entire spreadsheet).

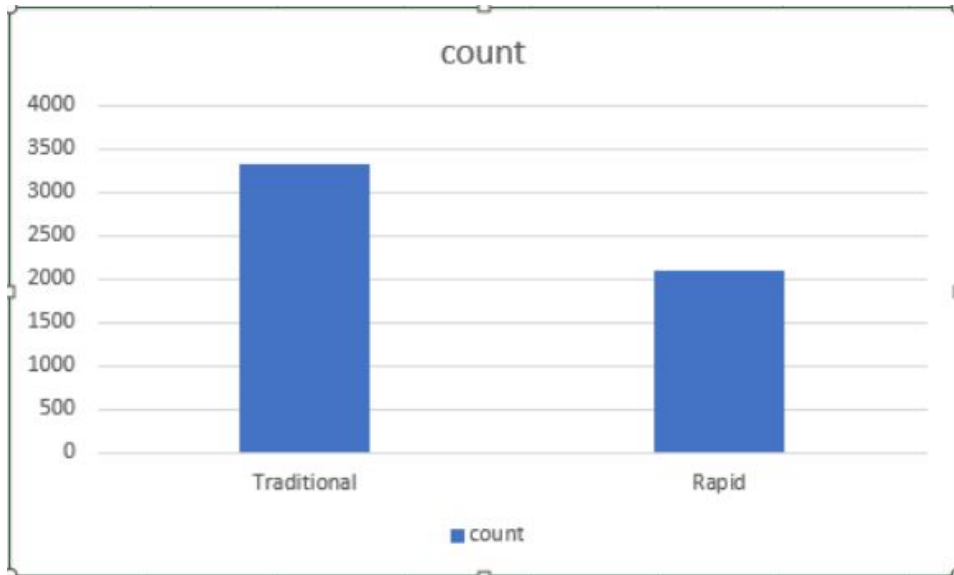
H	I
	count
Traditional	3334
Rapid	2091

- With the table values highlighted, click Insert > Column > Clustered Column (it's usually the icon on the left side of the 2-D Columns).

The screenshot shows the Excel Online interface. The 'Insert' tab is active, and the 'Column' button in the 'Charts' group is highlighted. A red arrow points to the '2-D Column' chart icon in the dropdown menu. Below the ribbon, a table is visible with columns A through K. The table data is as follows:

	A	B	C	D	E	F	G	H	I	J	K
1	id	duration (hours)	cost	inf	proc				count		
2	1	23	\$495.42	No Infection	Traditional						
3	2	14	\$772.70	No Infection	Traditional						
4	3	23	\$663.30	No Infection	Traditional						
5	4	18	\$599.50	No Infection	Traditional						
6	5	29	\$985.82	No Infection	Traditional						

A bar chart should pop up. It should look like this:



7. As usual, though, you should add some helpful titles to your chart. Click on the chart and do the following:

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

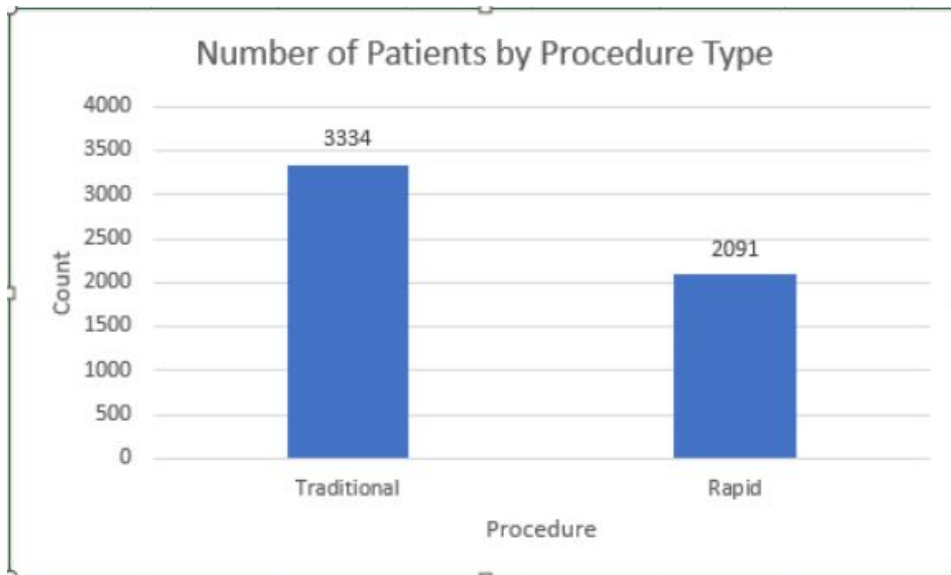
Vertical axis title: **Chart Tools > Axis Titles > Primary Vertical Axis Title > Edit Vertical Axis Title**, then type "Count"

Chart title: **Chart Tools > Chart Title > Edit Chart Title**, then type "Number of Patients by Procedure Type"

Data labels: **Chart Tools > Data Labels > Outside End**

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

8. Now your graph is clear and easy to read.



9. It's also helpful to see what *percentage* of patients got each procedure type. To do this, set up another mini-table below the one you just created (still in columns H and I). Like this:

H	I
	count
Traditional	3334
Rapid	2091
	percent
Traditional	
Rapid	

10. First, figure out the proportions as decimals (we'll convert these to percentages in a minute). Divide each of the counts from the other table by the total number of patients: 5425.

For example, in the "Traditional" row of your new table, divide the "Traditional" *count* from the other table by 5425 (use "=" and "/" to divide in Excel).

I6	fx	=I2/5425
----	-------------	----------

H	I
	count
Traditional	3334
Rapid	2091
	percent
Traditional	0.614562
Rapid	

11. Now do the same thing for the Rapid procedure: divide the “Rapid” count in the top table by 5425.

17	\sum	=I3/5425
----	--------	----------

H	I
	count
Traditional	3334
Rapid	2091
	percent
Traditional	0.614562
Rapid	0.385438

Those are your proportions as decimals, but they’ll be easier to parse (and read) if you convert them to percentages.

12. Highlight those two decimal values and click the percentage sign (%) in the Number bar of the Home tab.

Excel Online | OneDrive > Documents | Book 2

File Home Insert Data Review View Tell me what you want to do Open in E

Undo Clipboard Font Alignment

General % Percentage Number

16 fx =I2/5425

	A	B	C	D	E	F	G	H	I
1	d	duration (hours)	cost	inf	procedure	age			count
2	1	23	\$495.42	No Infection	Traditional	32.68		Traditional	3334
3	2	14	\$772.70	No Infection	Traditional	45.44		Rapid	2091
4	3	23	\$663.30	No Infection	Traditional	31.87			
5	4	18	\$599.50	No Infection	Traditional	27.51			percent
6	5	29	\$985.82	No Infection	Traditional	44.37		Traditional	0.614562
7	6	19	\$822.82	No Infection	Traditional	34.51		Rapid	0.385438
8	7	14	\$689.54	No Infection	Traditional	40.43			

Excel will convert them both to percentages.

	percent
Traditional	61.46%
Rapid	38.54%

Now you can see that about 61.46% of the patients from this data set got the Traditional procedure, while 38.54% got the Rapid version.

Exercise 2: Cost Distribution

Now we'll see how the costs of each type of procedure compare to each other.

1. First off, set up a column off to the side of the rest of your data for the average cost, like this (in column K):

	A	B	C	D	E	F	G	H	I	J	K
1	id	duration (hours)	cost	inf	procedure	age			count		avg cost
2	1	23	\$495.42	No Infection	Traditional	32.68		Traditional	3334		
3	2	14	\$772.70	No Infection	Traditional	45.44		Rapid	2091		
4	3	23	\$663.30	No Infection	Traditional	31.87					
5	4	18	\$599.50	No Infection	Traditional	27.51			percent		
6	5	29	\$985.82	No Infection	Traditional	44.37		Traditional	61.46%		
7	6	19	\$822.82	No Infection	Traditional	34.51		Rapid	38.54%		

- Use Excel's AVERAGE function to see what the average patient paid (including both types of procedure). The syntax is **=AVERAGE(firstcell:lastcell)**. The cost is in column C of this data set, so your first cell is C2 and your last cell is C5426. Type this into cell K2:

K2 **=AVERAGE(C2:C5426)**

Hit Enter.

K
avg cost
\$610.82

So the average procedure cost about \$610.82 (and remember, this includes *both* types of procedure). Now let's see how those costs are distributed.

- Click on the "C" above column C (which shows the cost) to highlight everything in that column.

	A	B	C	D	E	F
1	id	duration (hours)	cost	inf	procedure	age
2	1	23	\$495.42	No Infection	Traditional	32.68
3	2	14	\$772.70	No Infection	Traditional	45.44
4	3	23	\$663.30	No Infection	Traditional	31.87
5	4	18	\$599.50	No Infection	Traditional	27.51
6	5	29	\$985.82	No Infection	Traditional	44.37
7	6	19	\$822.82	No Infection	Traditional	34.51
8	7	14	\$689.54	No Infection	Traditional	40.43
9	8	8	\$426.78	No Infection	Traditional	24.68
10	9	20	\$649.20	No Infection	Traditional	41.86
11	10	17	\$780.87	No Infection	Traditional	42.87
12	11	10	\$570.64	No Infection	Traditional	35.52

- With everything in column C highlighted, click Insert > Other Charts > Histogram (usually the first icon under Statistical).

The screenshot shows the Excel Online ribbon with the 'Insert' tab selected. The 'Other Charts' dropdown menu is open, showing various chart types. The 'Statistical' category is expanded, and the 'Histogram' icon is circled in red. A red arrow points from the 'cost' column in the spreadsheet to the 'Histogram' icon.

	A	B	C	D	E	F	G	H
1	id	duration (hours)	cost	inf	procedure	age		
2	1	23	\$495.42	No Infection	Traditional	32.68		Traditional
3	2	14	\$772.70	No Infection	Traditional	45.44		Rapid
4	3	23	\$663.30	No Infection	Traditional	31.87		
5	4	18	\$599.50	No Infection	Traditional	27.51		
6	5	29	\$985.82	No Infection	Traditional	44.37		Traditional
7	6	19	\$822.82	No Infection	Traditional	34.51		Rapid
8	7	14	\$689.54	No Infection	Traditional	40.43		
9	8	8	\$426.78	No Infection	Traditional	24.68		
10	9	20	\$649.20	No Infection	Traditional	41.86		
11	10	17	\$780.87	No Infection	Traditional	42.87		
12	11	10	\$570.64	No Infection	Traditional	35.52		
13	12	13	\$827.86	No Infection	Traditional	38.33		
14	13	21	\$1,230.41	No Infection	Traditional	24.74		
15	14	12	\$445.03	No Infection	Traditional	30.53		
16	15	27	\$959.64	No Infection	Traditional	30.25		
17	16	24	\$1,021.72	No Infection	Traditional	37.17		

- When your new histogram pops up, click on the chart and do the following to add some quick labels:

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

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 "Cost Distribution"

- Your histogram should now look like this:



It might look a little confusing at first, but here's how it works: The horizontal axis along the bottom shows different ranges, or **bins**, of the cost variable. For example, the bin on the far left (the one labeled [\$123.95 ,...]) represents all the patients who paid between \$123.95 and \$185.95 for the procedure.

The vertical axis on the left shows the *number* of patients who fall into each bin category. For example, the blue bar on the far left is slightly below the marker for 200, which means that just under 200 people paid between \$123.95 and \$185.95 for the procedure.

Notice how the program automatically chooses the bins for you based on an even distribution of the data. Sadly, there's no way to change the bin size in Excel Online (though 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).

Our main takeaway from this graph is that the cost for these procedures varies pretty widely, but there's a big spike around the \$200 range.

- Now let's see how the average cost differs by procedure type. Next to the column you set up earlier for the average cost, set up two more columns for the Traditional procedure average cost and the Rapid procedure average cost, in columns L and M. Like this:

	K	L	M
1	avg cost	Trad avg	Rapid avg
2	\$610.82		

8. Use Excel's AVERAGEIFS function to find the average cost of the Traditional procedure. It's similar to the AVERAGE function, but AVERAGEIFS lets you find the average of cells that meet specific criteria (like finding the average of the costs in column C, but *only* for those patients who also have "Traditional" in column E).

The syntax is **=AVERAGEIFS(average_range, criteria_range, criteria)**. The average_range is the range of cells you want to pull the average from. The criteria_range is the range of cells that dictate which rows to include in the average. 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). So for this particular exercise, the average_range will be C2:C5426, because that covers everything in the "cost" column. The criteria_range will be E2:E5426, because you want to narrow down the patients and only look at those with "Traditional" in the "procedure" column. The criteria is "Traditional" (and it needs to be in quotation marks because it's text, not a number).

So here's what you should type into cell L2:

	K	L	M	N	O	P
	avg cost	Trad avg	Rapid avg			
	\$610.82	=AVERAGEIFS(C2:C5426, E2:E5426, "Traditional")				

It pulls from the original data like this:

L2	=AVERAGEIFS(C2:C5426, E2:E5426, "Traditional")					
	A	B	C	D	E	F
1	id	duration (hours)	cost	inf	procedure	age
2	1	23	\$495.42	No Infection	Traditional	32.68
3	2	14	\$772.70	No Infection	Traditional	45.44
4	3	23	\$663.30	No Infection	Traditional	31.87
5	4	18	\$599.50	No Infection	Traditional	27.51
6	5	29	\$985.82	No Infection	Traditional	44.37
7	6	19	\$822.82	No Infection	Traditional	34.51
8	7	14	\$689.54	No Infection	Traditional	40.43
9	8	8	\$426.78	No Infection	Traditional	24.68
10	9	20	\$649.20	No Infection	Traditional	41.86
11	10	17	\$780.87	No Infection	Traditional	42.87

Hit Enter, and Excel will calculate the average cost *only* for those patients with “Traditional” in column E.

K	L	M
avg cost	Trad avg	Rapid avg
\$610.82	806.2006	

You want that as a dollar amount, though, so highlight that cell and click on Home, then click the dropdown menu that says “General” and select “Currency” from the options.

The screenshot shows the Excel Online interface. The formula bar at the top displays the formula: `=AVERAGEIFS(C2:C5426, E2:E5426, "Traditional")`. The 'Home' tab is selected in the ribbon. The 'Number' dropdown menu is open, showing options: General, Number, Currency (highlighted with a red circle), and Accounting. The spreadsheet below shows columns A through M. Column L is highlighted, and cell L2 contains the value 806.20.

Now you can see that the average cost for the Traditional procedure was \$806.20.

K	L	M
avg cost	Trad avg	Rapid avg
\$610.82	\$806.20	

That’s quite a bit higher than the total average we found earlier for *both* procedure types.

- Repeat Step 8 and use Excel’s AVERAGEIFS function again to find the average cost of the Rapid procedure in cell M2. You can use the exact same syntax as before, but change the criteria to “Rapid” (in quotes):

M2 `=AVERAGEIFS(C2:C5426, E2:E5426, "Rapid")`

Hit Enter again.

K	L	M
avg cost	Trad avg	Rapid avg
\$610.82	\$806.20	299.3043

And once again, change that cell to a dollar amount by clicking Home > General > Currency (in the Number bar).

K	L	M
avg cost	Trad avg	Rapid avg
\$610.82	\$806.20	\$299.30

Now you have some interesting data! It turns out that the Rapid procedure had a far lower average cost than the Traditional procedure. In fact, the Traditional procedure was well over *twice* as expensive as the Rapid procedure, on average.

10. Create another histogram, but this time you want to see the cost distribution for *only* the Traditional procedures. Luckily, the data are already sorted by procedure type—in other words, all of the patients with a “Traditional” value in column E are listed first, while all of the “Rapid” entries are at the bottom of the spreadsheet. (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 that particular column.)

Take another look at the counts for each procedure type that you found earlier (in column I):

	A	B	C	D	E	F	G	H	I
1	id	duration (hours)	cost	inf	procedure	age			count
2	1	23	\$495.42	No Infection	Traditional	32.68		Traditional	3334
3	2	14	\$772.70	No Infection	Traditional	45.44		Rapid	2091
4	3	23	\$663.30	No Infection	Traditional	31.87			

There were 3334 patients who got the Traditional procedure. And since the patient data are sorted already by procedure type, that means all the Traditional folks are found in the range from row 2 to row 3335 (we added 1 because the first row of the spreadsheet contains the column titles).

So, to select *only* the costs for Patient 1 to Patient 3334 (those are the Traditional folks), click into the name box in the upper-left corner and type in **C2:C3335**, like this:

C2:C3335		fx		495.42		
	A	B	C	D	E	F
1	id	duration (hours)	cost	inf	procedure	age
2	1	23	\$495.42	No Infection	Traditional	32.68
3	2	14	\$772.70	No Infection	Traditional	45.44
4	3	23	\$663.30	No Infection	Traditional	31.87

Hit Enter, and that range of values will be highlighted.

- With cells C2:C3335 highlighted, click Insert > Other Charts > Histogram (the first icon under Statistical).

Excel Online | OneDrive > Documents | Book 2

File Home Insert Data Review View Tell me what you want to do Open in Excel

Function Survey Table PivotTable Picture Shapes Office Add-ins Column Line Pie Bar Area Scatter Other Charts Hyperlink Comment

Functions Tables Illustrations Add-ins Charts

C2 495.42

	A	B	C	D	E	F	G	H
1	id	duration (hours)	cost	inf	procedure	age		
2	1	23	\$495.42	No Infection	Traditional	32.68		Traditional
3	2	14	\$772.70	No Infection	Traditional	45.44		Rapid
4	3	23	\$663.30	No Infection	Traditional	31.87		
5	4	18	\$599.50	No Infection	Traditional	27.51		
6	5	29	\$985.82	No Infection	Traditional	44.37		Traditional
7	6	19	\$822.82	No Infection	Traditional	34.51		Rapid
8	7	14	\$689.54	No Infection	Traditional	40.43		
9	8	8	\$426.78	No Infection	Traditional	24.68		
10	9	20	\$649.20	No Infection	Traditional	41.86		
11	10	17	\$780.87	No Infection	Traditional	42.87		
12	11	10	\$570.64	No Infection	Traditional	35.52		
13	12	13	\$827.86	No Infection	Traditional	38.33		
14	13	21	\$1,230.41	No Infection	Traditional	24.74		
15	14	12	\$445.03	No Infection	Traditional	30.53		
16	15	27	\$959.64	No Infection	Traditional	30.25		

Waterfall

Funnel

Hierarchical

Statistical

- When that new histogram pops up, click on the chart and do the following to add some labels:

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

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 “Traditional Cost Distribution”

13. Your histogram should now look like this:



14. Now create one more histogram for the Rapid procedure cost distribution. We saw earlier that the Traditional cost data ran down to C3335. Since the data are already sorted by procedure type, that means the Rapid cost data starts at the next cell, C3336. It ends with the final patient’s cost data, all the way down at C5426.

So, to select *only* the costs for the patients who got the Rapid procedure, click into the name box in the upper-left corner and type in **C3336:C5426**, like this:

C3336:C5426						
242.16						
	A	B	C	D	E	F
1	id	duration (hours)	cost	inf	procedure	age
2	1	23	\$495.42	No Infection	Traditional	32.68
3	2	14	\$772.70	No Infection	Traditional	45.44
4	3	23	\$663.30	No Infection	Traditional	31.87

Press Enter. Note: The program might jump around a bit as it tries to select those cells, and it might take your view down to cell C3336 before it allows you to actually select that range of values. If necessary, **type C3336:C5426 into the name box a second time**, and press Enter again. (You can also just highlight cells C3336 to C5426 manually by clicking and dragging, but that takes a long time.) You should see this:

	A	B	C	D	E	F
3332	3331	24	\$684.73	No Infection	Traditional	30.63
3333	3332	23	\$1,044.28	No Infection	Traditional	25.93
3334	3333	26	\$988.96	No Infection	Traditional	58.92
3335	3334	15	\$843.07	No Infection	Traditional	27
3336	3335	10	\$242.16	No Infection	Rapid	32.4
3337	3336	3	\$202.11	No Infection	Rapid	22.38
3338	3337	3	\$311.08	No Infection	Rapid	25.53
3339	3338	7	\$427.71	No Infection	Rapid	40.81
3340	3339	3	\$233.46	No Infection	Rapid	27.35

15. With cells C3336:C5426 highlighted, click Insert > Other Charts > Histogram (the first icon under Statistical), just like you did in Step 11.

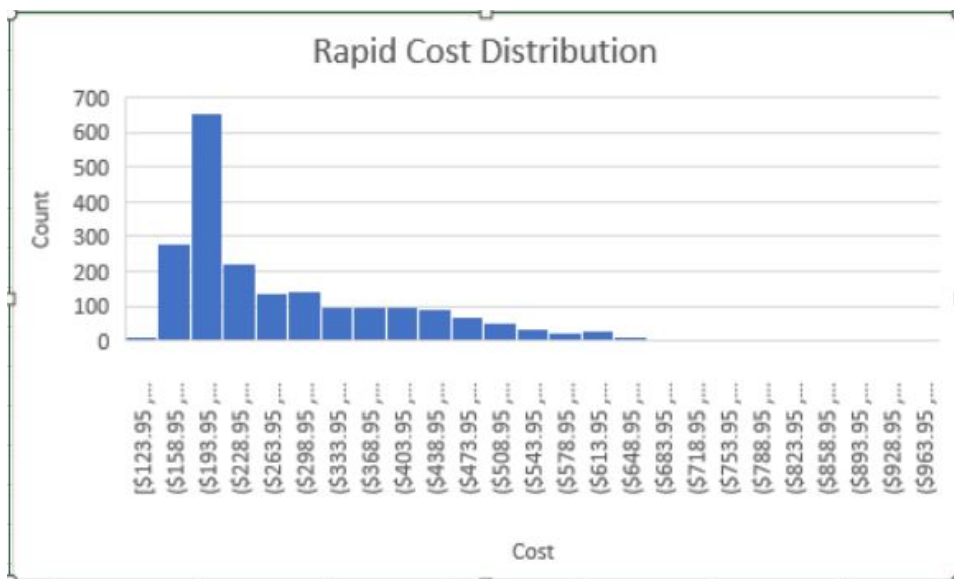
16. When your new histogram pops up, click on the chart and follow these directions to add labels again:

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

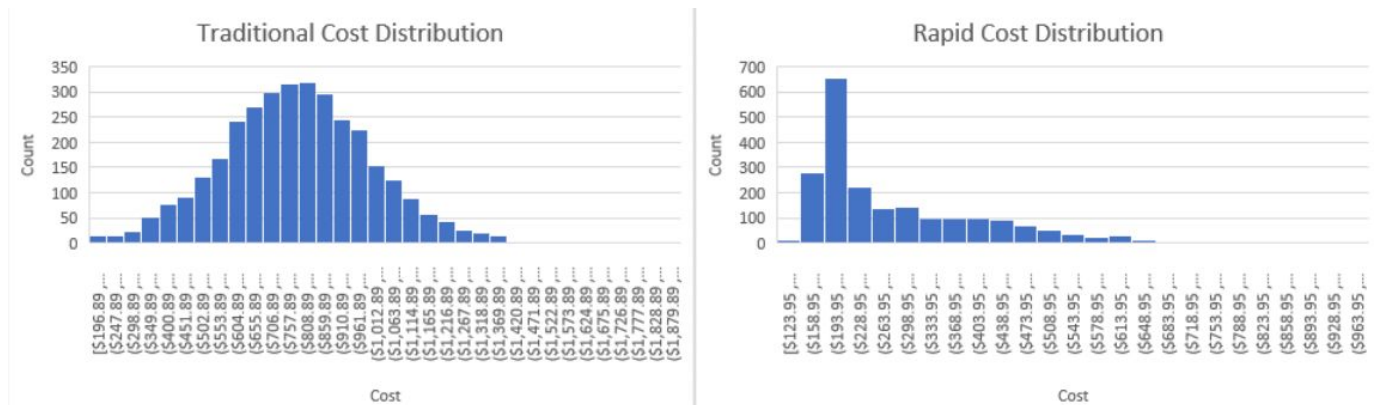
Vertical axis title: **Chart Tools > Axis Titles > Primary Vertical Axis Title > Edit Vertical Axis Title**, then type "Count"

Chart title: **Chart Tools > Chart Title > Edit Chart Title**, then type "Rapid Cost Distribution"

17. Now your graph should look like this:



18. If you look at the two histograms you just created side by side (you can literally just drag the two graphs next to each other in Excel), it's clear that there's a marked difference in cost for Traditional versus Rapid procedures.



Financially, at least, people do seem to benefit from the new Rapid procedure. It's less expensive, pretty much across the board.

Exercise 3: Duration Distribution

Now we'll use the same tricks from Exercise 2 to focus on the *duration* of each type of procedure.

- Set up another mini-table off to the side of the original data, directly beneath your table from Exercise 2 (still in columns K, L, and M). You'll need columns for the total average duration, the average duration of Traditional procedures, and the average duration of Rapid procedures, like this:

	K	L	M
1	avg cost	Trad avg	Rapid avg
2	\$610.82	\$806.20	\$299.30
3			
4	avg duration	Trad avg	Rapid avg
5			

- Use the AVERAGE function to find the average duration of procedures (both types). The duration data is in column B, so your total range runs from cell B2 all the way down to B5426. Type this into cell K5:

K5 fx =AVERAGE(B2:B5426)

Hit Enter.

avg duration	Trad avg	Rapid avg
13.2611982		

The average procedure in this data set took about 13.26 hours.

- Use Excel's AVERAGEIFS function to find the average duration of the Traditional procedure. (Basically, you're finding the average of the durations in column B, but *only* for those patients with "Traditional" in column E).

Again, the syntax is **=AVERAGEIFS(average_range, criteria_range, criteria)**. The *average_range* is the range of cells you want to pull the average from. The *criteria_range* is the range of cells that dictate which rows to include in the average. 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). So this time, the *average_range* will be B2:B5426, because that covers everything in the "duration" column. The *criteria_range* will still be E2:E5426, because you want to narrow down the patients and only look at those with "Traditional" in the "procedure" column. The *criteria* is "Traditional" (in quotation marks).

So here's what you should type into cell L5:

L5		=AVERAGEIFS(B2:B5426, E2:E5426, "Traditional")
----	---	--

When you press Enter, you'll see that the average Traditional procedure took about 18.28 hours.

avg duration	Trad avg	Rapid avg
13.2611982	18.27894	

Yikes—that's a really long time to be in surgery.

- Repeat Step 3, using the AVERAGEIFS function again to find the average duration of the Rapid procedure in cell M5. You can use the exact same syntax as in Step 3, but change the criteria to "Rapid" (in quotes):

M5		=AVERAGEIFS(B2:B5426, E2:E5426, "Rapid")
----	---	--

You know what's next: hit Enter.

avg duration	Trad avg	Rapid avg
13.2611982	18.27894	5.260641

The Rapid procedure lives up to its name! With an average of 5.26 hours per patient, it's much, much quicker than the Traditional procedure.

- Now let's churn out three quick histograms to see how the cost distributions look for both types together and each type individually.

Click on the "B" above column B (which shows the duration) to highlight everything in that column. With column B highlighted, click Insert > Other Charts > Histogram (the first icon under Statistical). It's the same icon you've been using for the past couple exercises.

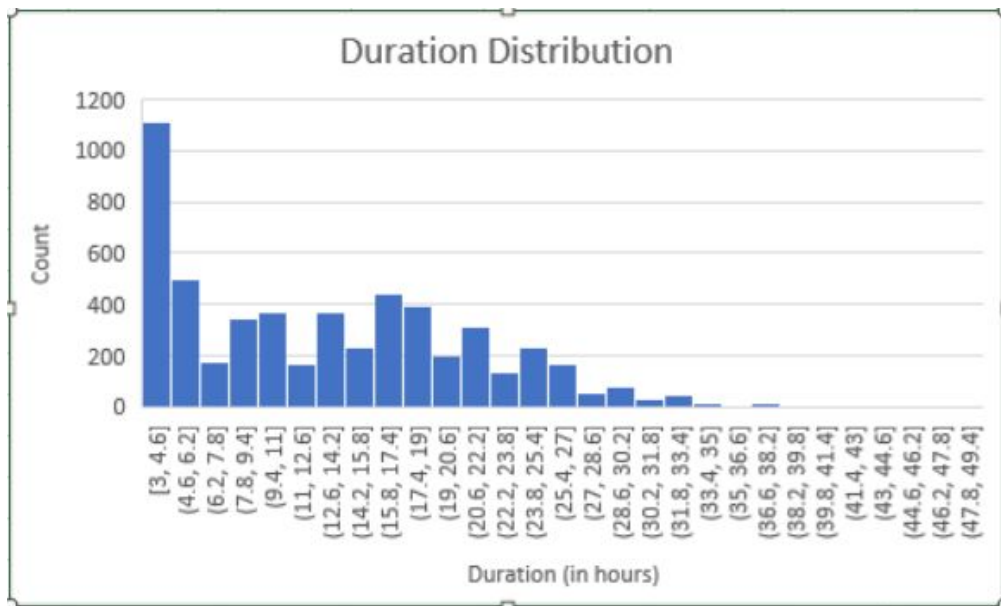
- When the new graph pops up, click on it and do the following to add important labels:

Horizontal axis title: **Chart Tools > Axis Titles > Primary Horizontal Axis Title > Edit Horizontal Axis Title**, then type "Duration (in hours)"

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 "Duration Distribution"

- Feast your eyes:

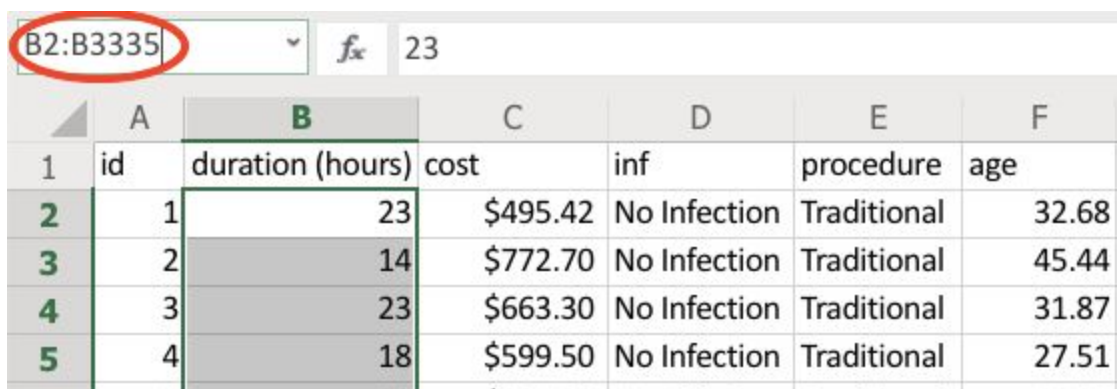


It looks like there was a big spike on surgeries that took between 3 and 4.6 hours.

8. Your second histogram will show the duration distribution for *only* the Traditional procedures. Like we mentioned before, the data are already sorted by procedure type—in other words, all of the patients with a “Traditional” value in column E are listed first. (Note: If they weren’t sorted already, you could click **Data > Filter** to bring up dropdown menus at the top of each column, which let you sort all of the data by that particular column.)

We also mentioned this before, but there were 3334 patients who got the Traditional procedure. Since the patient data are sorted already by procedure type, that means all the Traditional folks are found in the range from row 2 to row 3335 (we added 1 because the first row contains the column titles).

So, to select *only* the durations for Patient 1 to Patient 3334 (those are the Traditional folks), click into the name box in the upper-left corner and type in **B2:B3335**, like this:



	A	B	C	D	E	F
1	id	duration (hours)	cost	inf	procedure	age
2	1	23	\$495.42	No Infection	Traditional	32.68
3	2	14	\$772.70	No Infection	Traditional	45.44
4	3	23	\$663.30	No Infection	Traditional	31.87
5	4	18	\$599.50	No Infection	Traditional	27.51

Hit Enter to highlight that range of cells.

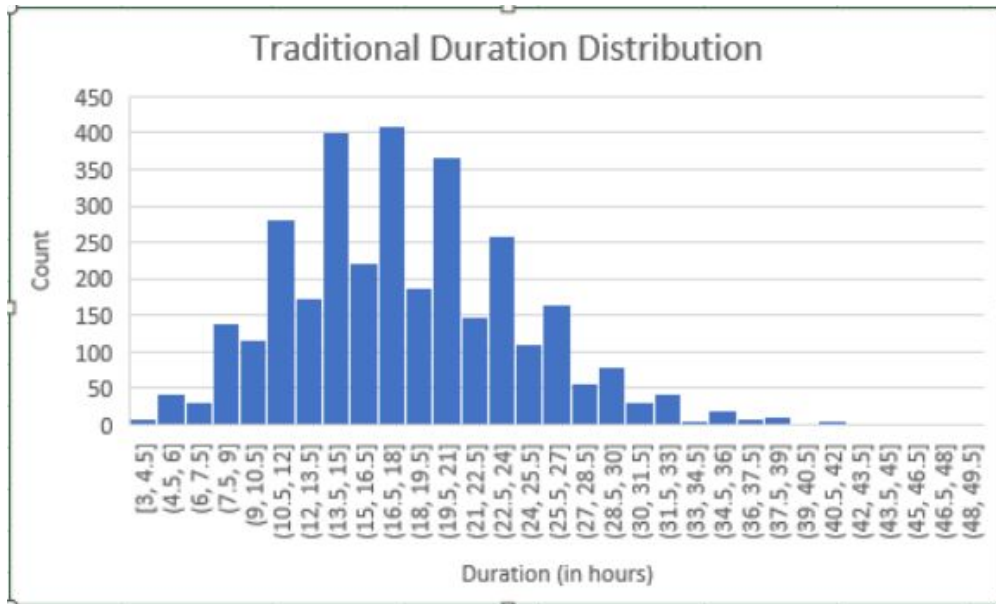
9. With cells B2:B3335 highlighted, click Insert > Other Charts > Histogram (the first icon under Statistical).
10. You probably know the drill by now. When the histogram pops up, click on the chart and do the following to add some helpful labels:

Horizontal axis title: **Chart Tools > Axis Titles > Primary Horizontal Axis Title > Edit Horizontal Axis Title**, then type “Duration (in hours)”

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 “Traditional Duration Distribution”

11. Check out your new chart:



12. Two down, one to go. Finish up by creating a histogram for the Rapid procedure's duration. We saw earlier that the Traditional duration data ran down to B3335. That means the Rapid duration data starts at the next cell, B3336. It ends with the final patient's duration data, all the way down at B5426.

So, to select *only* the durations for the patients who got the Rapid procedure, click into the name box in the upper-left corner and type in **B3336:B5426**. Hit Enter and wait a minute while the program jumps down to that point in the spreadsheet. For this range, you'll probably need to **type B3336:B5426 into the name box again, and press Enter again**.

B3336:B5426		fx		10		
	A	B	C	D	E	F
3333	3332	25	\$1,044.28	No Infection	Traditional	25.33
3334	3333	26	\$988.96	No Infection	Traditional	58.92
3335	3334	15	\$843.07	No Infection	Traditional	27
3336	3335	10	\$242.16	No Infection	Rapid	32.4
3337	3336	3	\$202.11	No Infection	Rapid	22.38
3338	3337	3	\$311.08	No Infection	Rapid	25.53
3339	3338	7	\$427.71	No Infection	Rapid	40.81

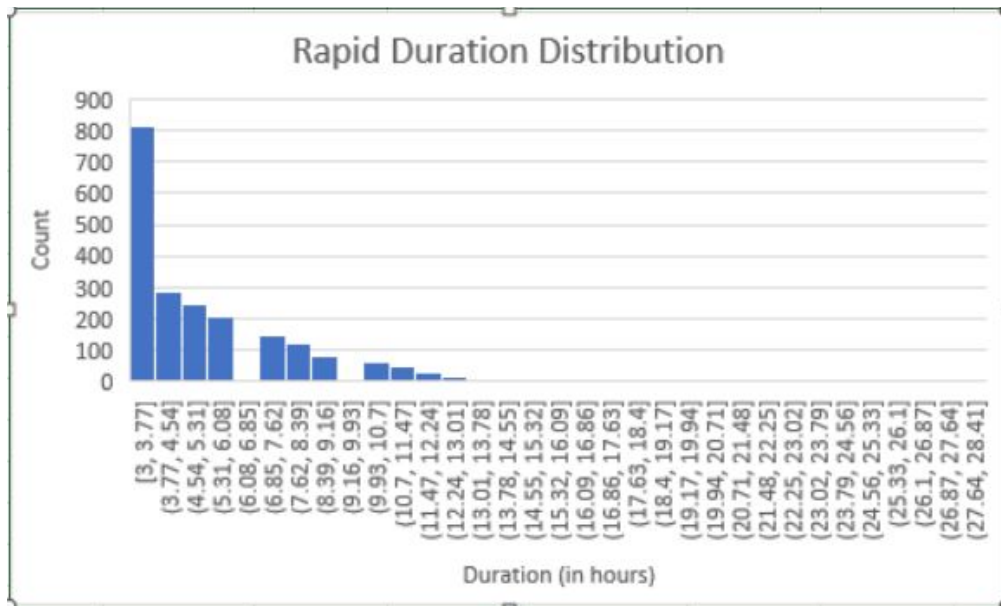
13. With cells B3336:B5426 highlighted, click Insert > Other Charts > Histogram (the first icon under Statistical), exactly like you did before.
14. When the histogram pops up, click on the chart and follow these directions to add labels again:

Horizontal axis title: **Chart Tools > Axis Titles > Primary Horizontal Axis Title > Edit Horizontal Axis Title**, then type “Duration (in hours)”

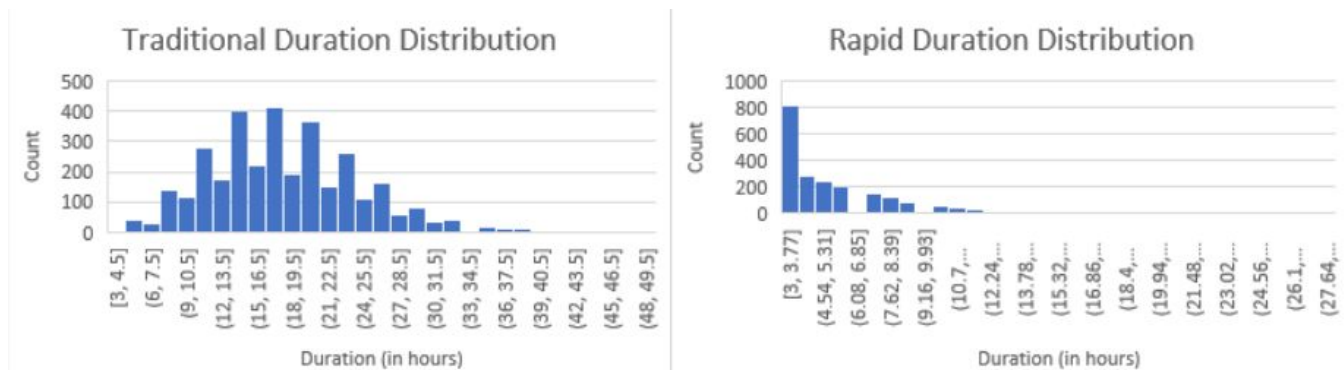
Vertical axis title: **Chart Tools > Axis Titles > Primary Vertical Axis Title > Edit Vertical Axis Title**, then type “Count”

Chart title: **Chart Tools > Chart Title > Edit Chart Title**, then type “Rapid Duration Distribution”

15. Here’s that histogram:



16. With a little dragging and resizing, you can see those last two graphs side by side:



The Rapid procedure is indeed rapid—patients are spending far less time in surgery than they do with the Traditional procedure.

Remember how much lower the average cost was for the Rapid procedure versus the Traditional? Now you can see a possible cause for this (or at least a correlation): the faster

procedure may be leading to a cheaper cost for the Rapid patients. Less time in surgery probably means less cost for the patient.

Exercise 4: Infection Rates

For our final exercise, we'll take a look at how the two procedure types affected the patients' infection rates.

1. Off to the side of the existing data set, create a new mini-table to show the counts for "Infection" and "No Infection" in the infection column. Your table should be in columns O and P, like this:

	O	P
1		count
2	Infection	
3	No Infection	

2. Use Excel's COUNTIF function to find those counts. 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:** If the criteria is a text value instead of a number, it **MUST** be in quotation marks.

In cell P2, for instance, you're counting the number of times "Infection" shows up in the "inf" column (that's column D). So your range is D2:D5426. Your criteria is "Infection" (don't forget the quotation marks).

P2	<i>fx</i>	=COUNTIF(D2:D5426, "Infection")
----	-----------	---------------------------------

Hit Enter.

O	P
	count
Infection	72
No Infection	

Yep, you're reading that right: Out of over 5000 patients, only 72 got an infection.

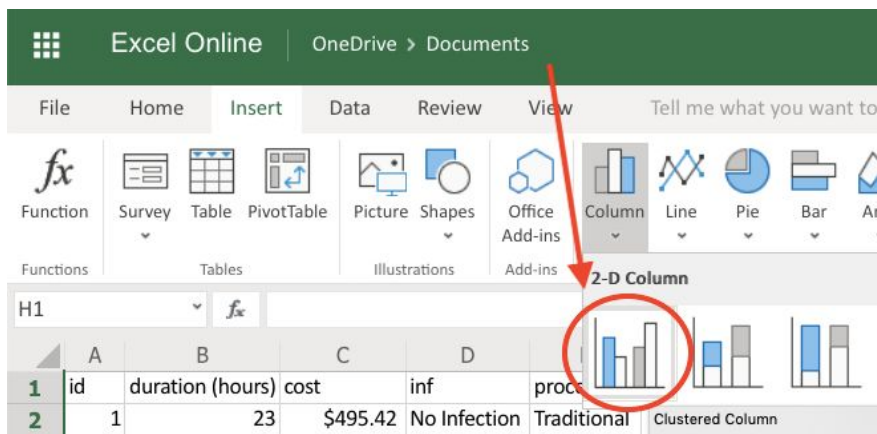
3. To find the number of patients who did not get an infection, use COUNTIF again. Click into cell P3 and use the same syntax as Step 2—the only difference is changing the criteria to "No Infection":

P3	=COUNTIF(D2:D5426, "No Infection")
----	------------------------------------

O	P
	count
Infection	72
No Infection	5353

The remaining 5353 patients did not get an infection. Good job, hospital!

- Now let's visualize these infection rates in a graph. Highlight everything in that new mini-table, including the category names (but do *not* highlight the entire spreadsheet). Then click Insert > Column > Clustered Column (usually the icon on the left side of the 2-D Columns).



- A bar chart should pop up. Click on the chart and add titles:

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

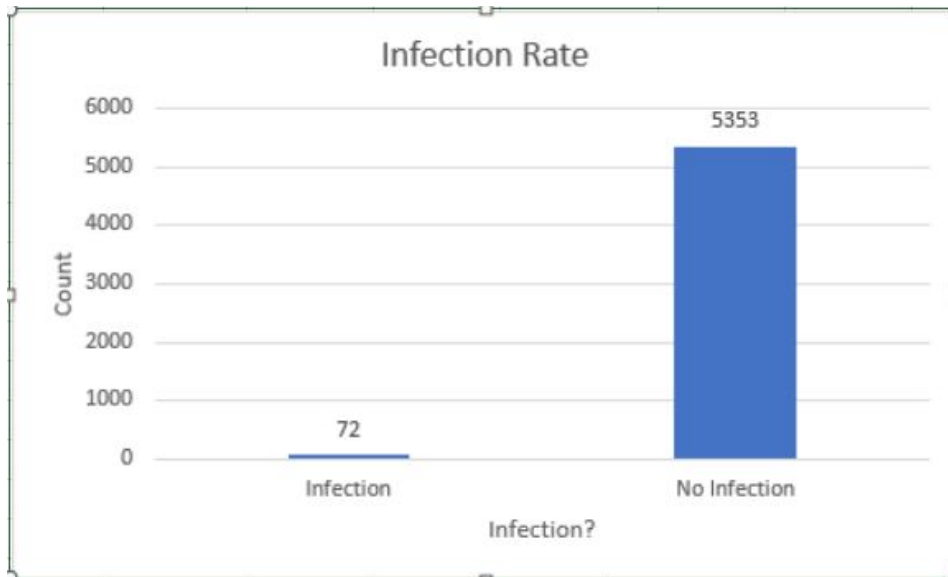
Vertical axis title: **Chart Tools > Axis Titles > Primary Vertical Axis Title > Edit Vertical Axis Title**, then type "Count"

Chart title: **Chart Tools > Chart Title > Edit Chart Title**, then type "Infection Rate"

Data labels: **Chart Tools > Data Labels > Outside End**

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

- Your graph should look like this now:



That's quite the difference! The infection rate was very low—always a good thing.

- Now break out the infection rate by the type of procedure. Do this by setting up—you guessed it—another mini-table off to the side. (Your graphs and charts are probably starting to stack up at this point, so feel free to drag and move some things around to make room in the spreadsheet.) Off to the right, in columns R, S, and T, make a 2-by-2 table with Traditional/Rapid labels for the rows, and Infection/No Infection labels for the columns.

	R	S	T
1		Infection	No Infection
2	Traditional		
3	Rapid		

- To find the counts for this table, use Excel's COUNTIFS function, which we saw back in the previous labs.

The syntax is **=COUNTIFS(criteria_range1, criteria1, criteria_range2, criteria2)**. It's like using two COUNTIF functions at the same time.

Both ranges take the form **firstcell:lastcell** (with a colon in between). The first range will be the "inf" data in column D (D2:D5426), and the second range will be the "procedure" data in column E (E2:E5426).

The first criteria (criteria1) is either "Infection" or "No Infection" (in quotation marks). The second criteria (criteria2) is either "Traditional" or "Rapid" (also in quotation marks).

This time, the first cell in your new mini-table (cell S2) is for the number of patients who got an infection *and* who got the Traditional procedure. Here's what you should type into cell S2:

S2 *fx* =COUNTIFS(D2:D5426, "Infection", E2:E5426, "Traditional")

Hit Enter and Excel will do the counting for you.

R	S	T
	Infection	No Infection
Traditional	63	
Rapid		

There were 63 patients who got an infection after getting the Traditional procedure.

- Repeat Step 8 for Traditional and No Infection. Just swap out the second criteria for “No Infection”:

T2 *fx* =COUNTIFS(D2:D5426, "No Infection", E2:E5426, "Traditional")

R	S	T
	Infection	No Infection
Traditional	63	3271
Rapid		

- Repeat Step 8 again for Rapid and Infection.

S3 *fx* =COUNTIFS(D2:D5426, "Infection", E2:E5426, "Rapid")

R	S	T
	Infection	No Infection
Traditional	63	3271
Rapid	9	

- Repeat Step 8 one last time for Rapid and No Infection.

T3 *fx* =COUNTIFS(D2:D5426, "No Infection", E2:E5426, "Rapid")

R	S	T
	Infection	No Infection
Traditional	63	3271
Rapid	9	2082

Well, it looks like there were far fewer infections for the Rapid procedure, but there were also fewer overall patients who *got* the Rapid procedure. To really compare how the procedures stacked up against each other in terms of infection rate, you'll need to break things down by percentage.

12. Time for one more mini-table. Directly below the table you just made (and still in columns R, S, and T), set up a new table with the exact same labels.

	R	S	T
1		Infection	No Infection
2	Traditional	63	3271
3	Rapid	9	2082
4			
5		Infection	No Infection
6	Traditional		
7	Rapid		

13. Figure out the proportions as decimals first. Divide each of the counts from the other table by the number of patients *who got that type of procedure*.

As a quick reminder, you already found the number of Traditional and Rapid patients in a different mini-table from way back in Exercise 1 of this lab. That table should still be over in columns H and I:

	H	I
1		count
2	Traditional	3334
3	Rapid	2091

So, for example, in the Traditional/Infection cell of your new table (cell S6), divide the Traditional/Infection *count* from the table above it (that's cell S2) by the number of patients who got the Traditional procedure (way over in cell I2). Use "=" and "/" to divide in Excel.

S6 fx =S2/I2

	H	I	J	K	L	M	N	O	P	Q	R	S	T
1		count		avg cost	Trad avg	Rapid avg			count			Infection	No Infection
2	Traditional	3334		\$610.82	\$806.20	\$299.30		Infection	72		Traditional	63	3271
3	Rapid	2091						No Infection	5353		Rapid	9	2082
4				avg duration	Trad avg	Rapid avg							
5		percent		13.2611982	18.27894	5.260641						Infection	No Infection
6	Traditional	61.46%									Traditional	=S2/I2	
7	Rapid	38.54%									Rapid		

Note: Do NOT divide by the total number of patients in the entire data set (5425) because you're not looking for the *total* percentage of patients who got an infection; you're looking for the percentage of Traditional patients who got an infection. That's why you're dividing by the total number of Traditional folks instead of by the grand total.

Hit Enter, and your new table should show the following decimal:

	Infection	No Infection
Traditional	0.018896	
Rapid		

14. To get the Traditional/No Infection proportion, divide the count for Traditional/No Infection (cell T2) by the number of patients who got the Traditional procedure (cell I2 again).

T6 fx =T2/I2

	H	I	J	K	L	M	N	O	P	Q	R	S	T
1		count		avg cost	Trad avg	Rapid avg			count			Infection	No Infection
2	Traditional	3334		\$610.82	\$806.20	\$299.30		Infection	72		Traditional	63	3271
3	Rapid	2091						No Infection	5353		Rapid	9	2082
4				avg duration	Trad avg	Rapid avg							
5		percent		13.2611982	18.27894	5.260641						Infection	No Infection
6	Traditional	61.46%									Traditional	0.018896	=T2/I2
7	Rapid	38.54%									Rapid		

Hit Enter!

	Infection	No Infection
Traditional	0.018896	0.98110378
Rapid		

15. To get the Rapid/Infection proportion, divide the count for Rapid/Infection (cell S3) by the number of patients who got the Rapid procedure (cell I3 this time).

S7 fx =S3/I3

	H	I	J	K	L	M	N	O	P	Q	R	S	T
1		count		avg cost	Trad avg	Rapid avg			count			Infection	No Infection
2	Traditional	3334		\$610.82	\$806.20	\$299.30		Infection	72		Traditional	63	3271
3	Rapid	2091						No Infection	5353		Rapid	9	2082
4				avg duration	Trad avg	Rapid avg						Infection	No Infection
5		percent		13.2611982	18.27894	5.260641						Infection	No Infection
6	Traditional	61.46%									Traditional	0.018896	0.98110378
7	Rapid	38.54%									Rapid	=S3/I3	

Press Enter to calculate.

	Infection	No Infection
Traditional	0.018896	0.98110378
Rapid	0.004304	

16. And finally, to find the Rapid/No Infection proportion, divide the count for Rapid/No Infection (cell T3) by the number of patients who got the Rapid procedure (cell I3 again).

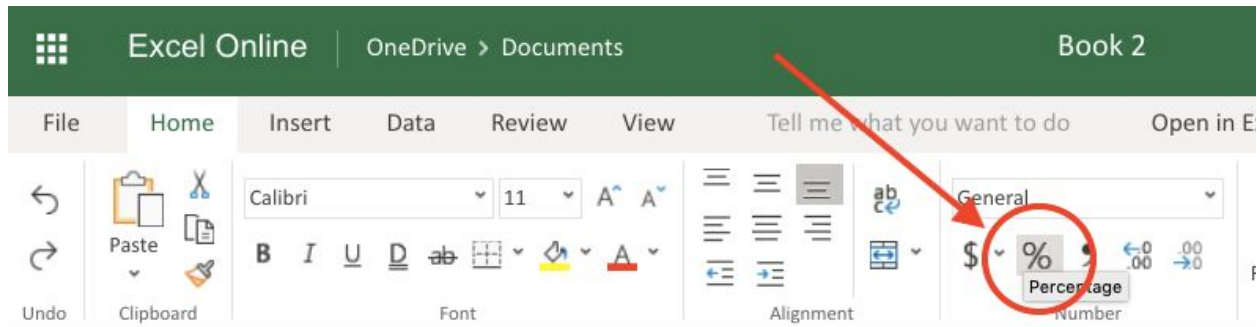
T7 fx =T3/I3

	H	I	J	K	L	M	N	O	P	Q	R	S	T
1		count		avg cost	Trad avg	Rapid avg			count			Infection	No Infection
2	Traditional	3334		\$610.82	\$806.20	\$299.30		Infection	72		Traditional	63	3271
3	Rapid	2091						No Infection	5353		Rapid	9	2082
4				avg duration	Trad avg	Rapid avg						Infection	No Infection
5		percent		13.2611982	18.27894	5.260641						Infection	No Infection
6	Traditional	61.46%									Traditional	0.018896	0.98110378
7	Rapid	38.54%									Rapid	0.004304	=T3/I3

Hit Enter to finish up.

	Infection	No Infection
Traditional	0.018896	0.98110378
Rapid	0.004304	0.99569584

17. Highlight those four decimals from the new table and click the percentage sign (%) in the Number bar of the Home tab.



Now you can see how the infection rates compare.

	Infection	No Infection
Traditional	1.89%	98.11%
Rapid	0.43%	99.57%

Both procedures had a fairly low infection rate, but for the Rapid procedure, the infection rate (0.43%) was less than *one-fourth* what it was for the Traditional procedure (1.89%). That's a really great outcome, and a good sign that the Rapid procedure seems quite a bit safer than the Traditional method.

Of course, a good theory is that the infection rate is lower for the Rapid procedure because those patients spend less time in the hospital.

- Quick side note: You might recall from the Business Data Walkthrough Lab in Module 2 that we briefly mentioned something called a “chi-squared” independence test (or χ^2 , after the Greek letter). Again, we won't delve into the actual math here, but what this test does, basically, is tell us whether the difference between any two rows in our percentage table (i.e. Traditional versus Rapid) is “statistically significant.” If so, it would mean that the difference is probably a real thing and not just a fluke of our sample size. If we ran this test for this particular percentage table, we would see that the difference *is* statistically significant. In other words, there is a real difference between the infection rates for the Traditional and Rapid procedures. And that difference is likely to continue.

18. So, since we're data analysts here, let's summarize our findings for the hospital:

The new Rapid procedure appears to be cheaper, faster, and safer. (We know it's cheaper from Exercise 2, faster from Exercise 3, and safer from Exercise 4.) Based on these metrics, we can safely recommend the Rapid procedure over the Traditional one.