BIG DATA IN ENVIRONMENTAL SCIENCE AND TOXICOLOGY

superfund.tamu.edu

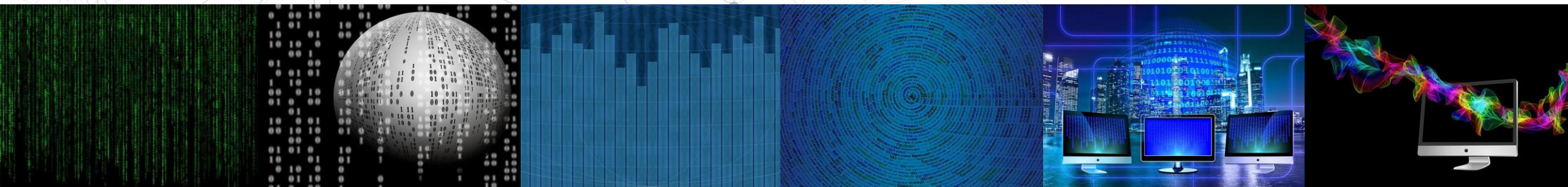
TEXAS A&M UNIVERSITY
Superfund Research Center

Welcome !

This Session will Begin at 2:00 pm Eastern US Time

"Manipulating Big(ish) Data in Excel, and Reading into R"

Fred Wright – North Carolina State University
Candice Brinkmeyer-Langford – Texas A&M University
Dillon Lloyd – North Carolina State University

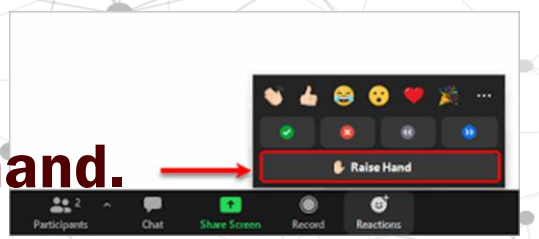


BIG DATA IN ENVIRONMENTAL SCIENCE AND TOXICOLOGY

superfund.tamu.edu

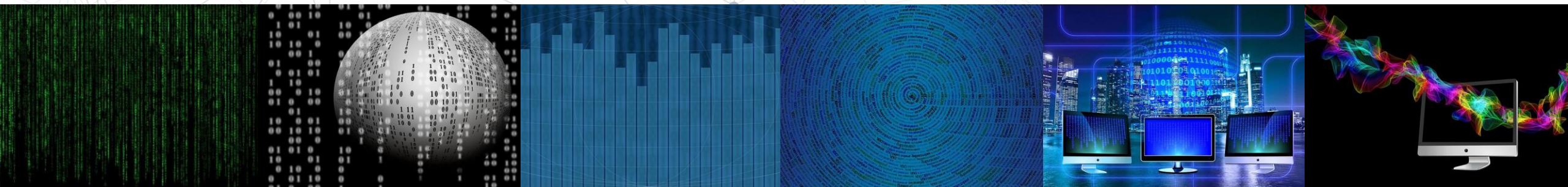
TEXAS A&M UNIVERSITY
Superfund Research Center

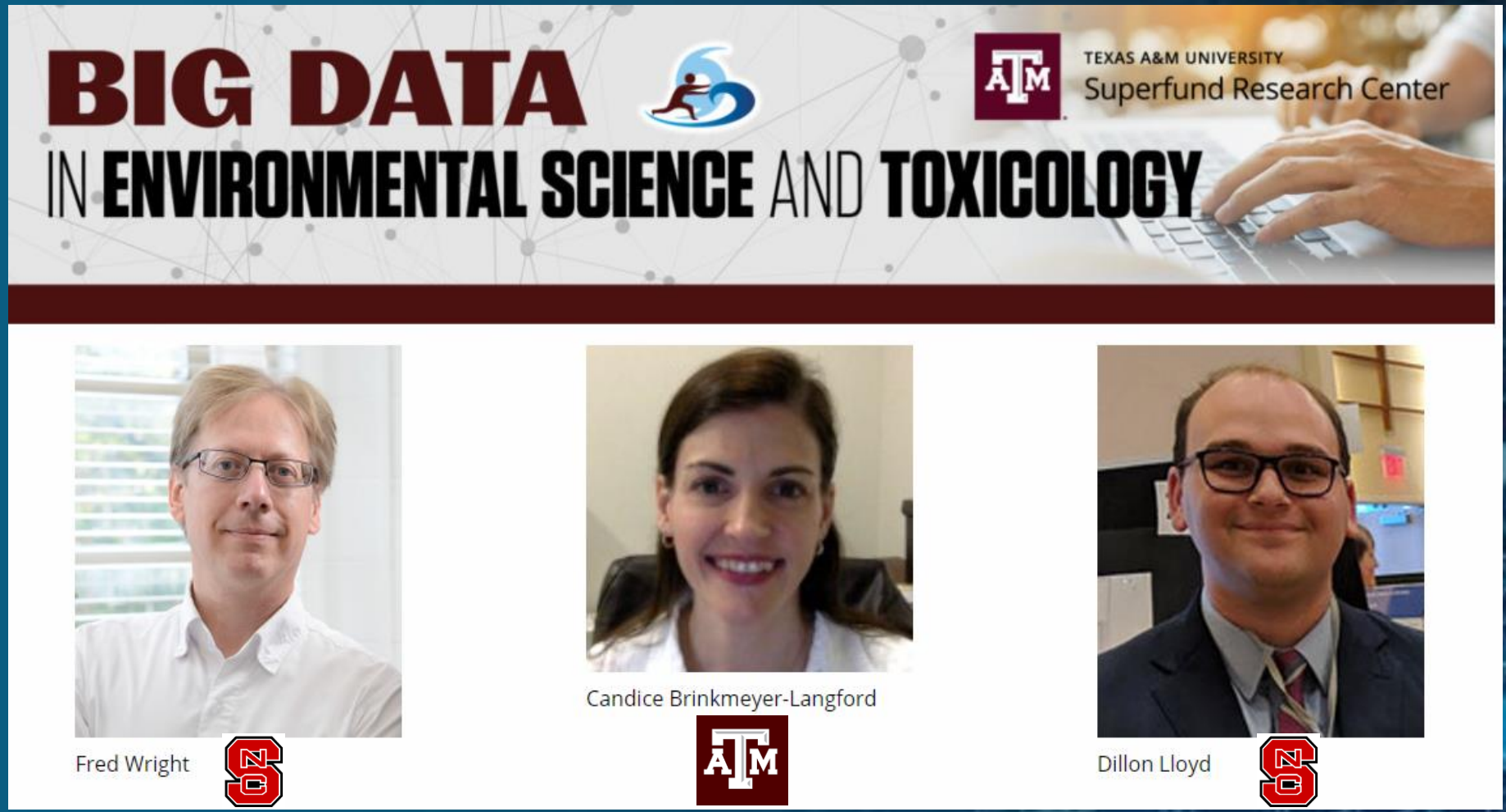
- All participants are muted to enable the speaker to present without interruption.
- Please rename yourself and designate Full Name and Affiliation.
- Last-minute installation issues? Use private chat to Dillon Lloyd.
- Use the reaction icon at the bottom of your screen to raise your hand.



However, due to time constraints, most questions will be answered in chat window, with answers posted later as an FAQ.

- This meeting will be recorded, and posted on the @tamusuperfund website https://superfund.tamu.edu/big-data-series-2021/ in the coming weeks.





Basics

Managing data and plotting

Data analysis

What is this session about?

- This session is NOT a comprehensive introduction to Excel
- We present a number of topics that have been useful to the speakers for Superfund-relevant data.
- In other words, what do we and collaborators do most often?
- We include links to a number of topics
- We assume only the most basic familiarity with Excel

What is this session about?

- Excel is not the main tool for data science
- It is a widely used tool for storing and working with data
- Best practices need to acknowledge reality that Excel is widely used

Basics

Fred Wright, NC State University

What is Big Data?

- Big Data invokes and evokes a lot of concepts
- Data we can hold in one spreadsheet is not *that* big

- However, data that we cannot perceive in one screenful is big enough that we need to let the software tell us things

Josh Lee, Medium

Soundcloud

Amazon

Excel data limits

- Much better than the old days!
- Roughly speaking, big enough for gene expression datasets
- Not big enough for genome scans or sequence data

Worksheet and workbook specifications and limits

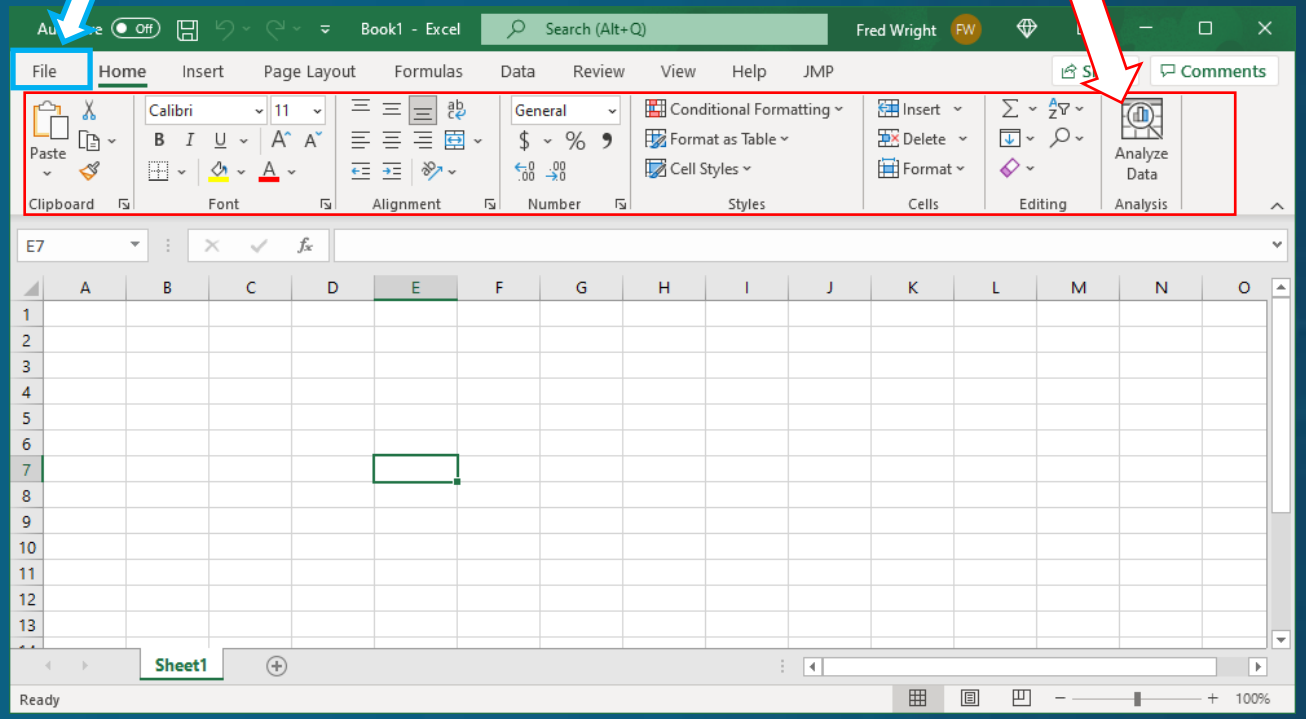
Feature	Maximum limit
Open workbooks	Limited by available memory and system resources
Total number of rows and columns on a worksheet	1,048,576 rows by 16,384 columns
Column width	255 characters
Row height	409 points
Page breaks	1,026 horizontal and vertical
Total number of characters that a cell can contain	32,767 characters
Characters in a header or footer	255

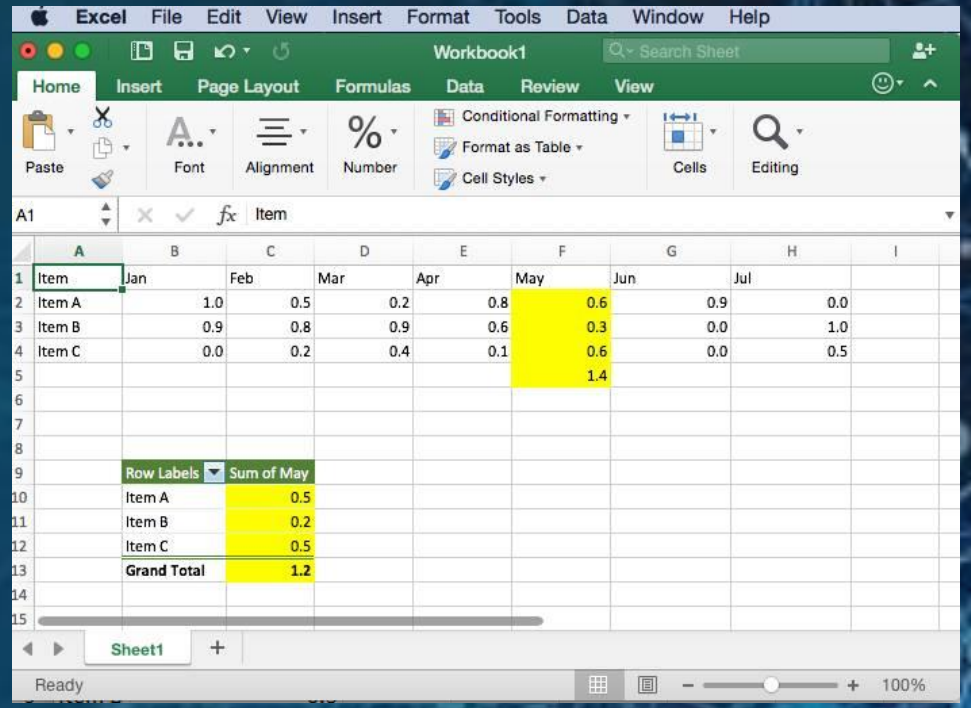
PC vs. Mac

Tab

Ribbon

(may need to turn on the ribbon using View)





CTRL* vs. CMD*

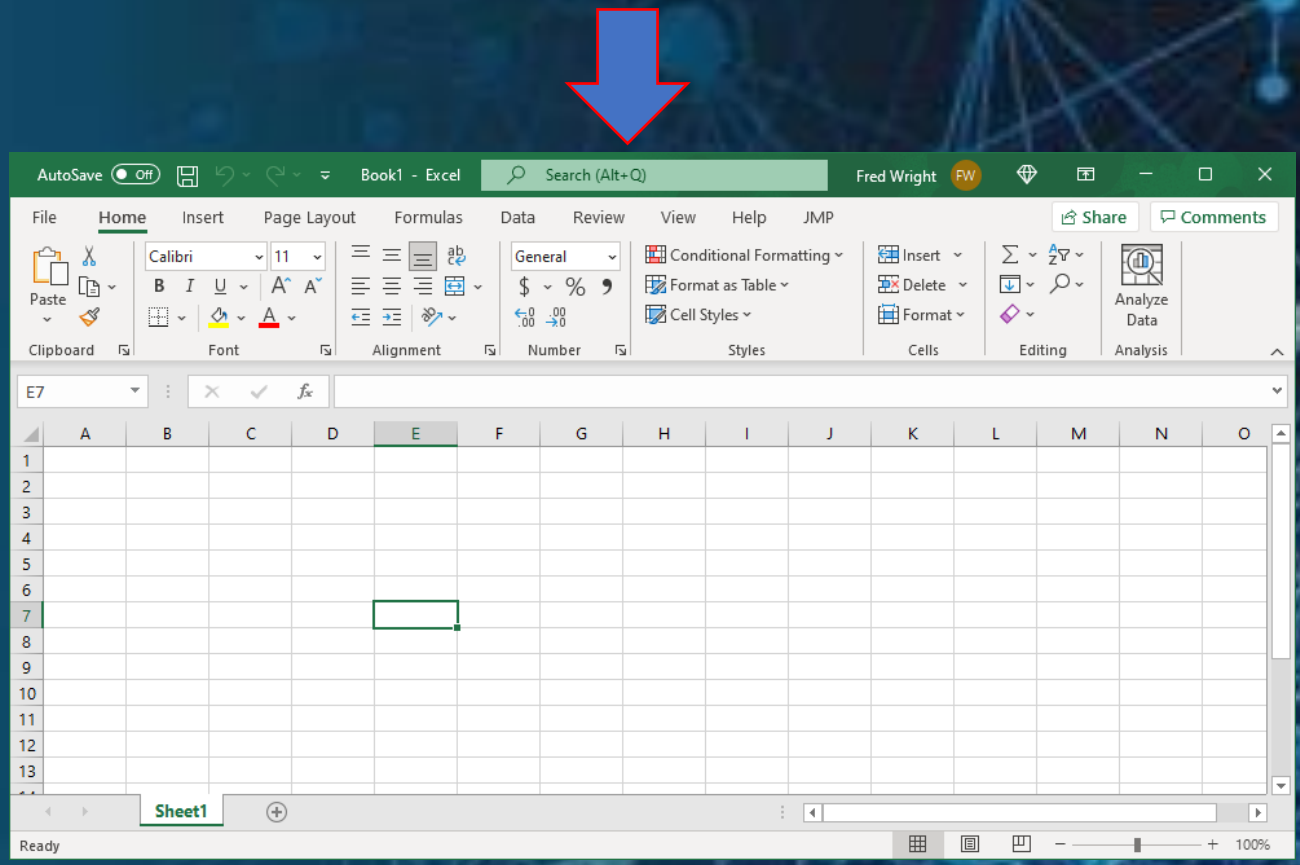
* means whatever you are typing!

Got an Excel question?

- The search bar is pretty good!
- If you're still not sure, try googling the question, which often leads to a keyword
- Examples

- format values in Euro currency

- how do I turn my rows into columns…. what is that called?



Importing data from a text file

https://support.microsoft.com/en-us/office/import-or-export-text-txt-or-csv-files-5250ac4c-663c-47ce-937b-339e391393ba

- Use file>open>browse>(correct format)>open
- But… this is not an Excel file
- The text import wizard saves the day. Pay attention to delimiters (character that separates variables)
- Need to skip the first few lines? There's a box for that!

I goofed!

https://support.microsoft.com/en-us/office/undo-redo-or-repeat-an-action-84bdb9bc-4e23-4f06-ba78-f7b893eb2d28#:~:text=To%20undo%20an%20action%20press,want%20to%20undo%20multiple%20steps.

- Never fear, CTRL-z is here
- or undo curvy arrow
- Wait, I didn't goof!

  Other curvy arrow (or CTRL-y)

lowes.com

Jump around

https://support.microsoft.com/en-us/office/move-or-scroll-through-a-worksheet-06fc34b8-64bb-4d78-9b62-34656d700f82

- You can scroll around
- Or you can jump using CTRL* (e.g. CTRL ↓)
- Try CTRL END or CTRL HOME. What do they do?
- These are called “shortcuts”
- Select stuff using SHIFT
- Combine stuff, e.g. CTRL SHIFT *

hot147.com

Freeze panes!

- Useful for scrolling around while keeping some rows/columns fixed
- View>freeze panes >freeze panes
- Top row, first column, and what else?

EMI and Georgann Deen

Pixar

Appearance and formatting of cells (basic)

https://support.microsoft.com/en-us/office/available-number-formats-in-excel-0afe8f52-97db-41f1-b972-4b46e9f1e8d2

- Cell appearance similar to other MS products
- Boldface, italics, text, size, etc.

- Formatting may appear on the Home action bar
- Or Design>formatting
- Check Alignment and Number
- The formatting doesn't really change the underlying values (look at formula bar)

Sorting

https://support.microsoft.com/en-us/office/sort-data-in-a-range-or-table-62d0b95d-2a90-4610-a6ae-2e545c4a4654

- select value in a column, hope for the best!

OR

- do it deliberately, Home>Editing>Sort and Filter>Custom Sort

- But I have two header rows!

- Sorting is dangerous! Always do a sanity check. Did I sort the region intended, and not just one column?

Adding, deleting, hiding

https://support.microsoft.com/en-us/office/insert-or-delete-rows-and-columns-6f40e6e4-85af-45e0-b39d-65dd504a3246#:~:text=To%20insert%20a%20single%20column,and%20then%20select%20Insert%20Columns.

https://support.microsoft.com/en-us/office/hide-or-show-rows-or-columns-659c2cad-802e-44ee-a614-dde8443579f8

- Logical for the most part
- The right mouse button is your friend

Highlighting info by conditional formatting

https://support.microsoft.com/en-us/office/use-conditional-formatting-to-highlight-information-fed60dfa-1d3f-4e13-9ecb-f1951ff89d7f

- Highlight interesting cells or ranges of cells

- Select cells
- Home>Conditional formatting
- Take a look and explore

Simple functions and formulas

- Make new columns that are log10 of the previous columns
- Note how the cells are relative in the formula
- You can change them to absolute (sometimes called “anchor”) values by using $ or F4 (on PC)

Transforming data (Power Query)

https://www.youtube.com/watch?v=yE8sLnywPAg

- Data tab
- From table/range (make sure correct range is selected)

Drop-down lists

https://support.microsoft.com/en-us/office/create-a-drop-down-list-7693307a-59ef-400a-b769-c5402dce407b

1. Select the cells that you want to contain the lists.
2. On the ribbon, DATA -> Data Tools->Data Validation.
3. Allow to List.
4. Click in Source, type the text or numbers (separated by commas, for a comma-delimited list) that you want in your drop-down list, and click OK.

Good naming and formatting practices for sharing data via spreadsheet

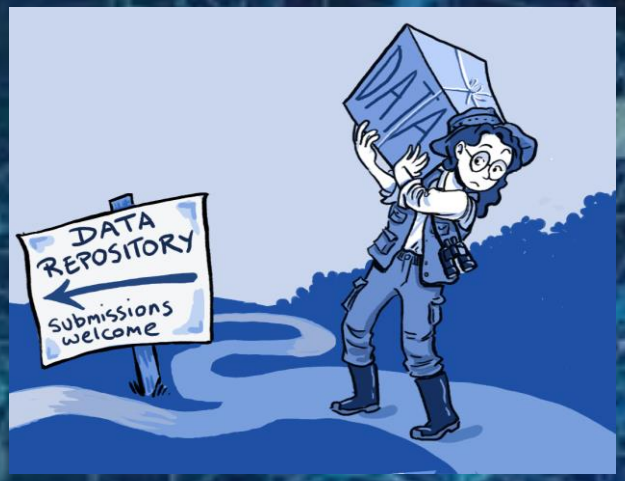
concept

reality

- Data science is 1% inspiration, 99% exasperation (mainly caused by extra spaces, poor heading labels, and unnecessary detail in missing data)

Good naming and formatting practices for sharing data via spreadsheet

- DO keep all the data as one row/column rectangle (if possible)
- DO keep things as simple as possible
- DO keep column headings descriptive and CONSISTENT, including utterly consistent capitalization
- DO use a single code (e.g. NA) for missing data if possible
- DO NOT merge cells
- DO NOT add actual commas to numerical values
- DO NOT add hidden spaces in data
- DO NOT use spaces in column headings. Underscore is preferred.



Ainsley Seago. doi:10.1371/journal.pbio.1001779.g001

Transposing data

- Changes rows to columns
- Not exactly “turning the rectangle of data on its side”
- Copy and Paste Special, select Transpose

- See https://support.microsoft.com/en-us/office/transpose-function-ed039415-ed8a-4a81-93e9-4b6dfac76027 to learn how to do this using the TRANSPOSE function, while preserving formulas

Create tables

https://support.microsoft.com/en-us/office/create-and-format-tables-e81aa349-b006-4f8a-9806-5af9df0ac664

- Don’t we already have a table?’
- A table is a specific object for a rectangular data range that Excel can use for analysis
- Hallmark of a table is the coloring/banding, and the drop down button for column headings
- Home > Format as Table

Pivot tables

https://support.microsoft.com/en-us/office/create-a-pivottable-to-analyze-worksheet-data-a9a84538-bfe9-40a9-a8e9-f99134456576

- Useful for summarizing totals, etc. across variables
- Similar to cross-tabulation
- Somewhat restricted to certain functions, etc.
- Insert>Pivot Table
- The next speaker will provide different ways of summarizing what you have

Next up: Manipulating data and counting stuff!

http://publicdomainaudiovideo.blogspot.com/

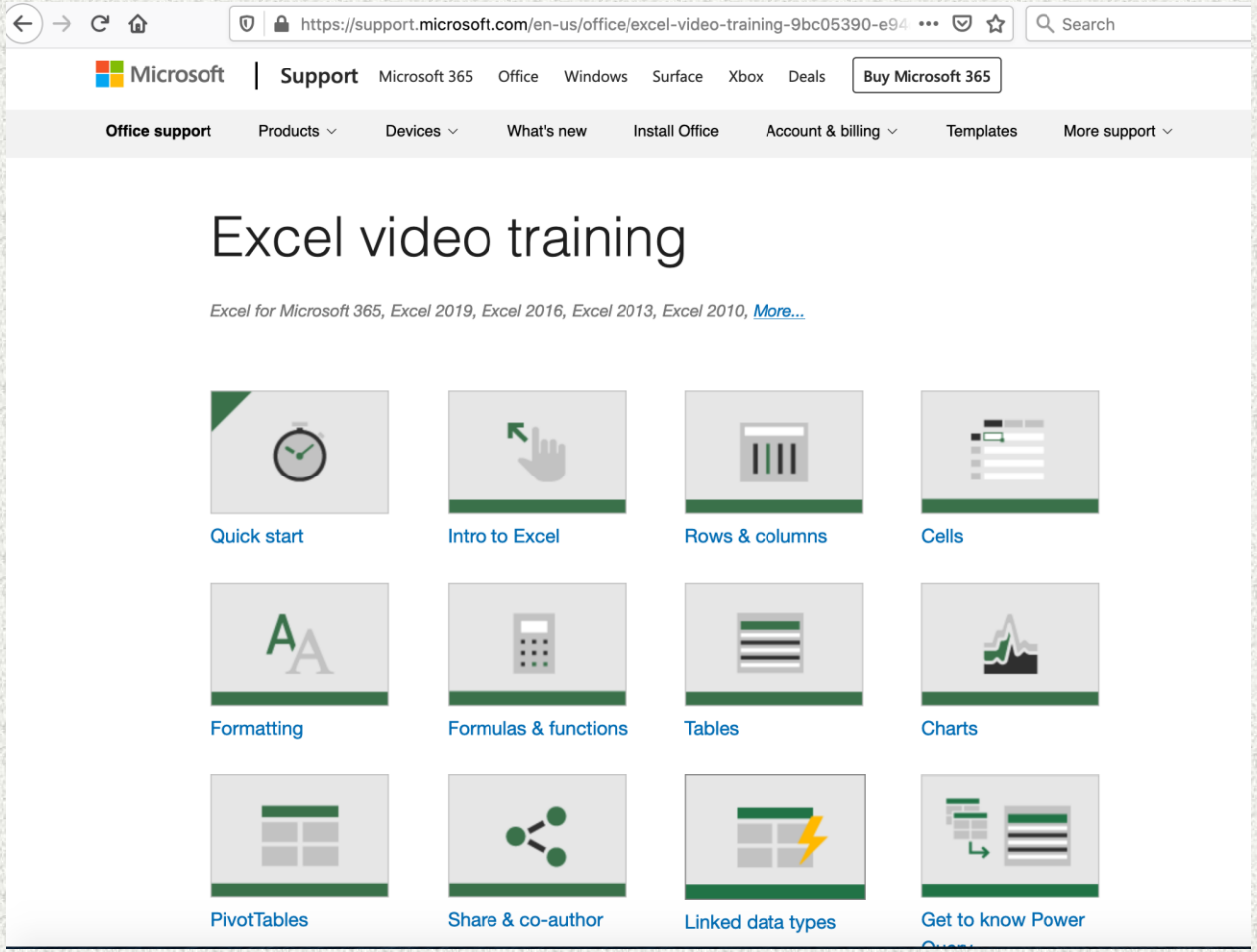
Manipulating Big(ish) Data in Excel

Candice Brinkmeyer-Langford

Texas A&M University

Excel version and support

- Excel version being used: Microsoft Excel for Mac, version 16.51
- https://support.microsoft.com/en-us/office/excel-video-training-9bc05390-e94c-46af-a5b3-d7c22f6990bb



Answering questions with Excel

IF
COUNTIF
AVERAGEIF
LOOKUP
VLOOKUP

Charts:
Bar chart
Scatterplot
Secondary axis
Stacked bar chart
Box-and-whisker plot

Miscellaneous tips and tricks

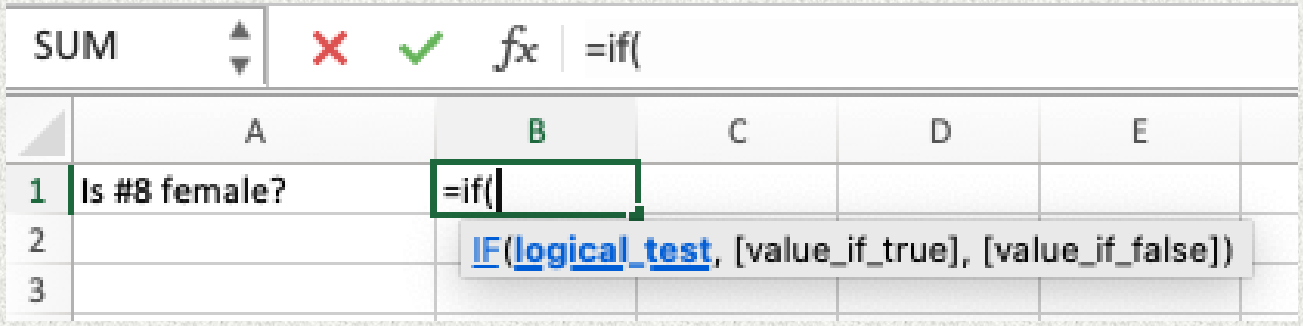
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Sample_ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2	Sex	F	F	M	F	M	M	M	F	F	M	F	M	M	F	M	F	M	F	M	F
3	Status	N	Y	Y	N	N	N	Y	Y	N	Y	Y	N	Y	N	N	Y	Y	Y	N	N
4	Cat1	3	7	9	3	1	4	7	10	6	9	1	8	8	1	1	2	5	2	8	1
5	Cat2	0.39980403	0.131167	0.83797435	0.99121998	0.37915433	0.4932649	0.31730995	0.78559809	0.29231847	0.22800014	0.12970948	0.20895121	0.09570753	0.57328551	0.74719656	0.97775537	0.90570917	0.23738616	0.63079488	0.03630628
6	Cat3	0.01765164	0.74075219	0.4243295	0.84309817	0.41803328	0.36549674	0.08057165	0.73402944	0.9952279	0.6285995	0.66614321	0.15886006	0.29619534	0.522676	0.49873951	0.499717	0.57727393	0.35131784	0.24743548	0.36314088
7	Cat4	76	75	65	53	89	66	88	78	98	67	75	96	90	65	70	89	73	84	85	71
8	Cat5	3	10	3	1	7	5	5	3	3	3	8	6	7	5	8	7	4	2	9	9
9	Cat6	0.27817695	0.07110137	0.37747939	0.13183529	0.31935773	0.1925463	0.63984189	0.64802318	0.26027957	0.93419475	0.52918677	0.2509355	0.39102529	0.77463581	0.59241452	0.48270423	0.83625265	0.41941355	0.83425559	0.93651799
10	Cat7	0.47111146	0.89977883	0.42504482	0.09046157	0.43938449	0.85181114	0.7906725	0.03804672	0.98191601	0.06929245	0.17350784	0.00365075	0.62290958	0.10161224	0.129711	0.87430617	0.63898328	0.72340885	0.87721221	0.74542847
11	Cat8	59	79	81	65	54	91	50	92	96	51	80	63	78	70	91	60	80	59	81	92
12	Cat9	0.15067873	0.69674031	0.2607653	0.54603649	0.94019016	0.63397075	0.31429126	0.1763019	0.87520766	0.0101267	0.8489138	0.81448124	0.14730753	0.47184386	0.97195415	0.10610593	0.55886697	0.34184183	0.38806705	0.24976963
13	Cat10	1646	1436	1486	1218	1151	1278	1329	1942	1085	1662	1748	1922	1061	1850	1160	1166	1049	1419	1275	1207
14	Cat11	535	586	543	470	479	502	582	537	412	514	462	487	543	405	504	478	466	419	467	591
15	Cat12	13	19	15	18	11	19	18	10	16	11	15	19	20	18	13	17	16	14	14	14
16																					

IF

- I want to know if individual 8 is female.
  - I could just look at this little table, obviously, but pretend it's a much bigger dataset.

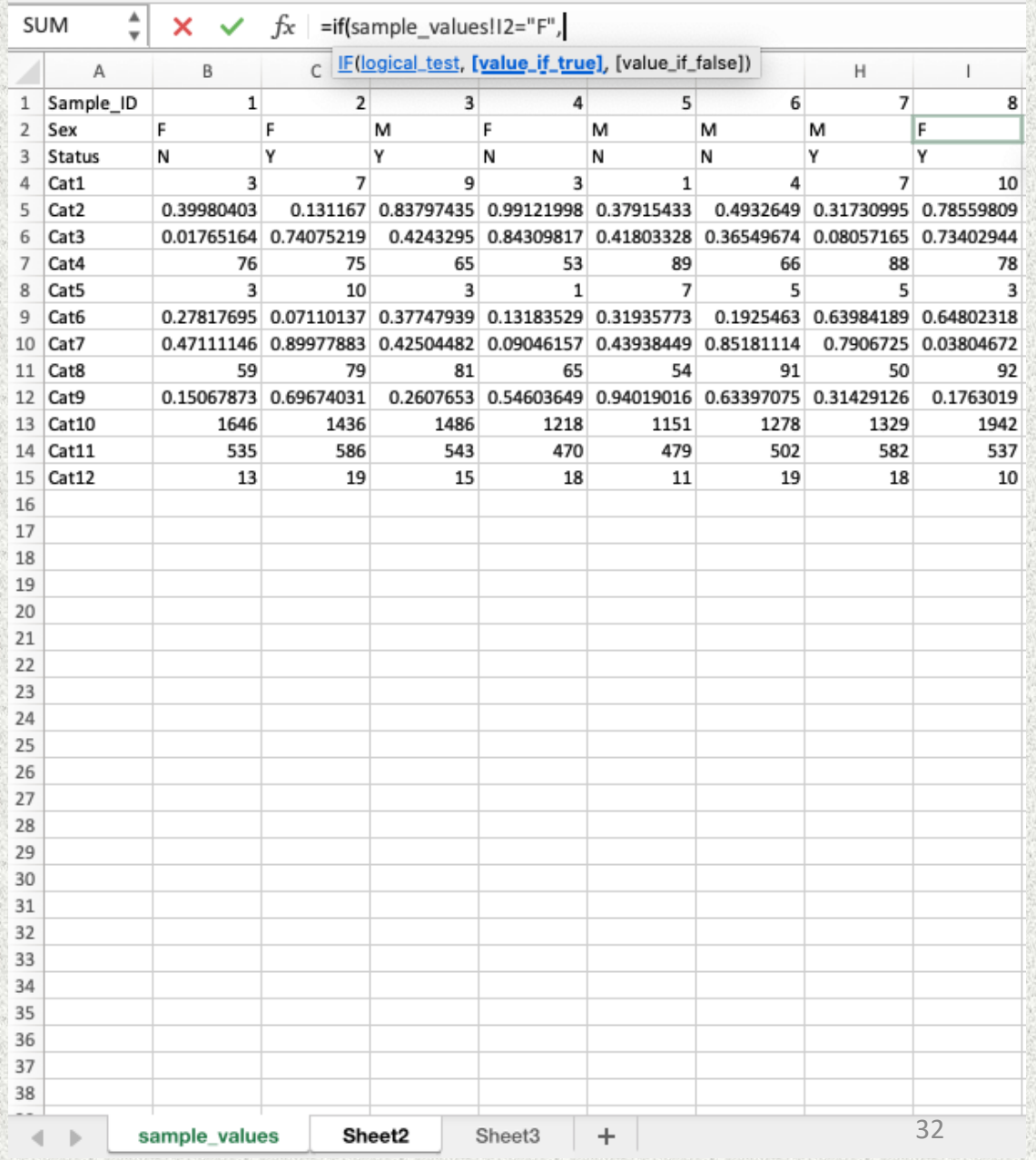
- The **IF** function can have 2 results (True or False)

- =IF(test,value_if_true,value_if_false)



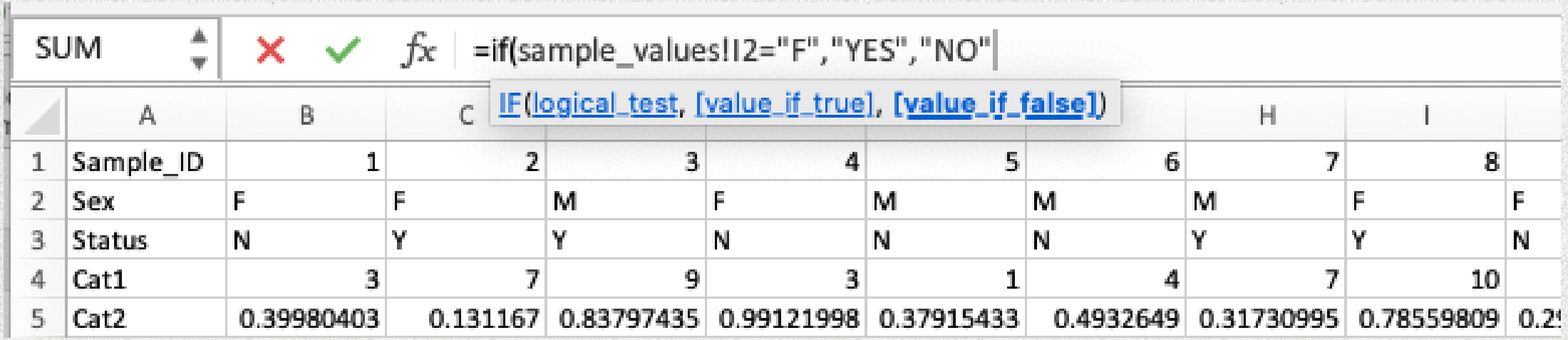
IF

- We are looking on a different sheet for our answer, hence the sample_values! before the specific cell I2
- Because we want to know if the contents of cell I2 are a specific value, we use quotation marks: "F"
- Don't forget to add the comma!



	A	B	C	D	E	F	G	H	I
1	Sample_ID	1	2	3	4	5	6	7	8
2	Sex	F	F	M	F	M	M	M	F
3	Status	N	Y	Y	N	N	N	Y	Y
4	Cat1	3	7	9	3	1	4	7	10
5	Cat2	0.39980403	0.131167	0.83797435	0.99121998	0.37915433	0.4932649	0.31730995	0.78559809
6	Cat3	0.01765164	0.74075219	0.4243295	0.84309817	0.41803328	0.36549674	0.08057165	0.73402944
7	Cat4	76	75	65	53	89	66	88	78
8	Cat5	3	10	3	1	7	5	5	3
9	Cat6	0.27817695	0.07110137	0.37747939	0.13183529	0.31935773	0.1925463	0.63984189	0.64802318
10	Cat7	0.47111146	0.89977883	0.42504482	0.09046157	0.43938449	0.85181114	0.7906725	0.03804672
11	Cat8	59	79	81	65	54	91	50	92
12	Cat9	0.15067873	0.69674031	0.2607653	0.54603649	0.94019016	0.63397075	0.31429126	0.1763019
13	Cat10	1646	1436	1486	1218	1151	1278	1329	1942
14	Cat11	535	586	543	470	479	502	582	537
15	Cat12	13	19	15	18	11	19	18	10

IF

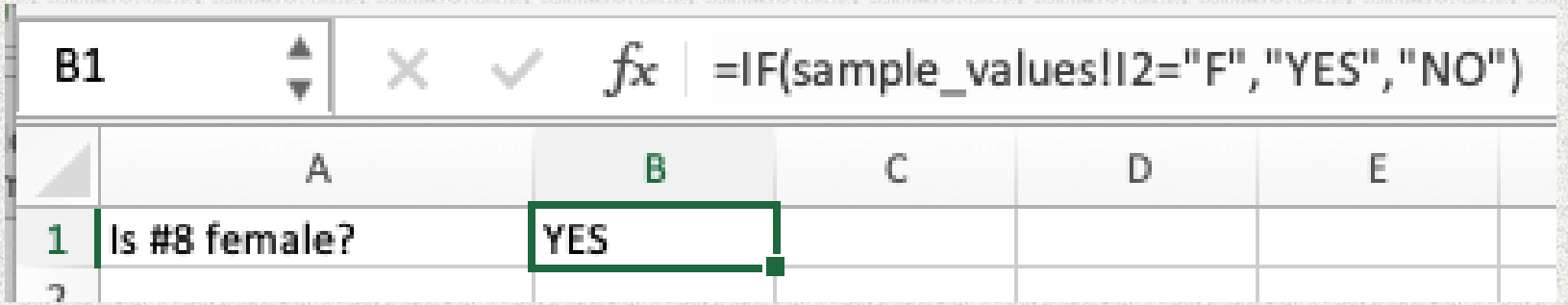


	A	B	C	D	E	F	G	H	I	
1	Sample_ID	1	2	3	4	5	6	7	8	
2	Sex	F	F	M	F	M	M	M	F	F
3	Status	N	Y	Y	N	N	N	Y	Y	N
4	Cat1	3	7	9	3	1	4	7	10	
5	Cat2	0.39980403	0.131167	0.83797435	0.99121998	0.37915433	0.4932649	0.31730995	0.78559809	0.2

- If my test is true – that is, if column I2 is “F” (in other words, individual #8 is female…
  - “YES”,
  - If it’s not true, I want Excel to say “NO”
- Don’t forget to close the parentheses so Excel knows you’re done.

IF

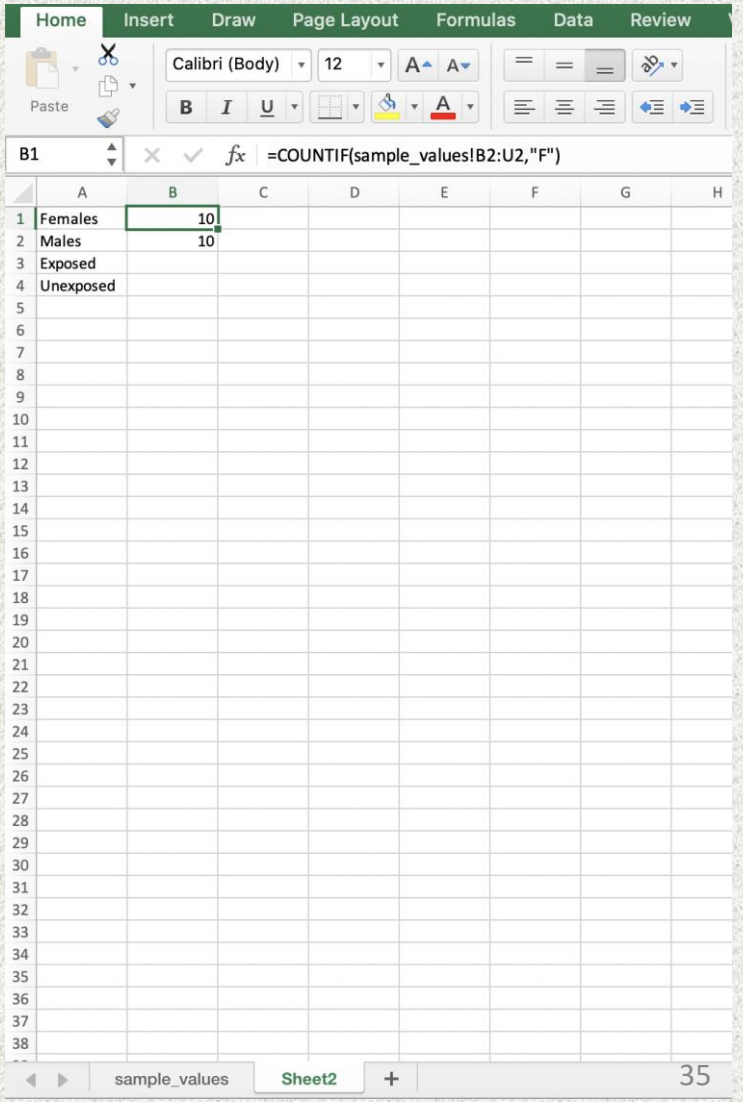
- When I add the close parenthesis (“)”) and hit Enter, I can go back to my original sheet (Sheet1, in this case) and see my answer.



How many are in certain categories…?

COUNTIF function counts how many cells in a given range have a certain value

=COUNTIF(Where do you want to look?, What do you want to look for?)

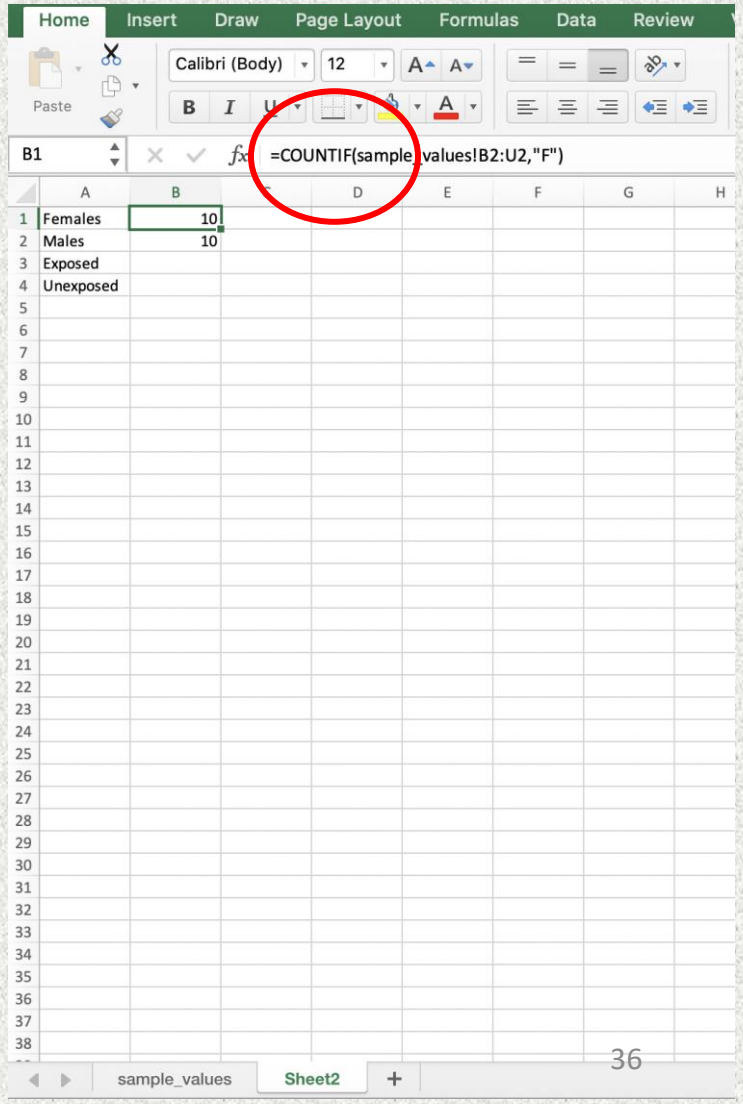


How many are in certain categories…?

COUNTIF function counts how many cells in a given range have a certain value

=COUNTIF
(the equals sign tells Excel you want it to carry out some function)

Then, open parenthesis to tell Excel exactly what you want it to search ("cells in a given range") and what you want it to search for ("a certain value")



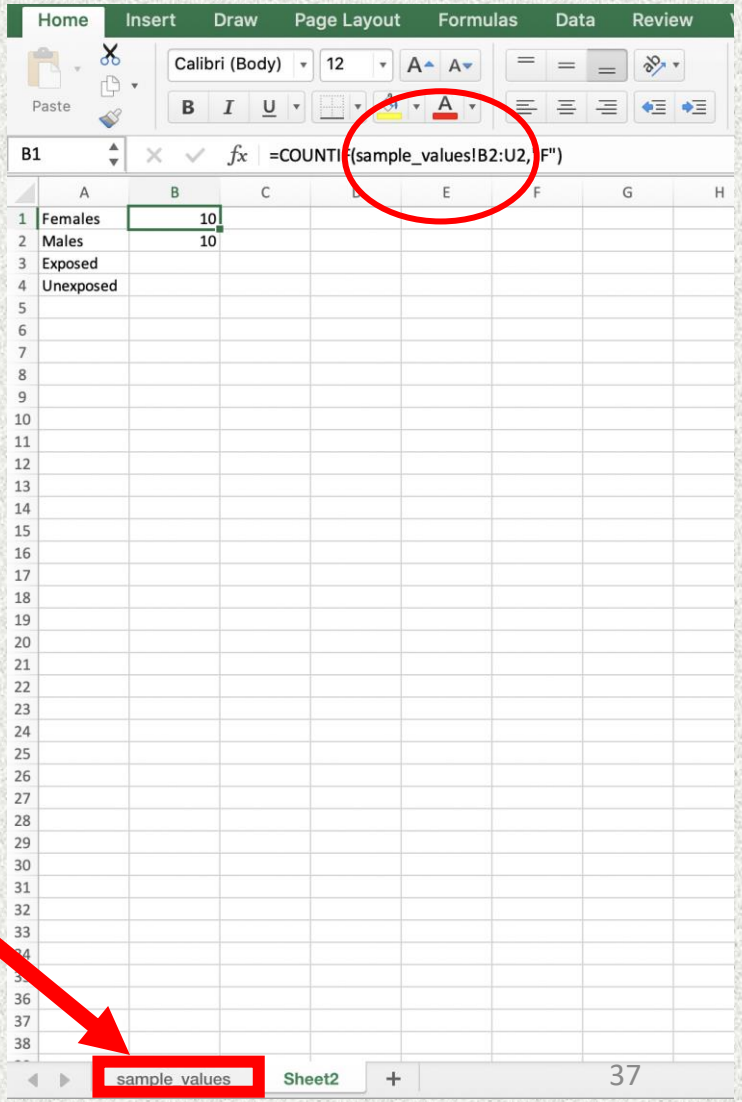
How many females are in this data set?

COUNTIF function counts how many ***cells in a given range*** have a certain value

Where is the "range" of data that contains what we want to count?

Here, the "range" is on a different sheet (sample_values). This is OK.

The exclamation point (!) tells Excel to look on a different sheet: sample_values!



How many females are in this data set?

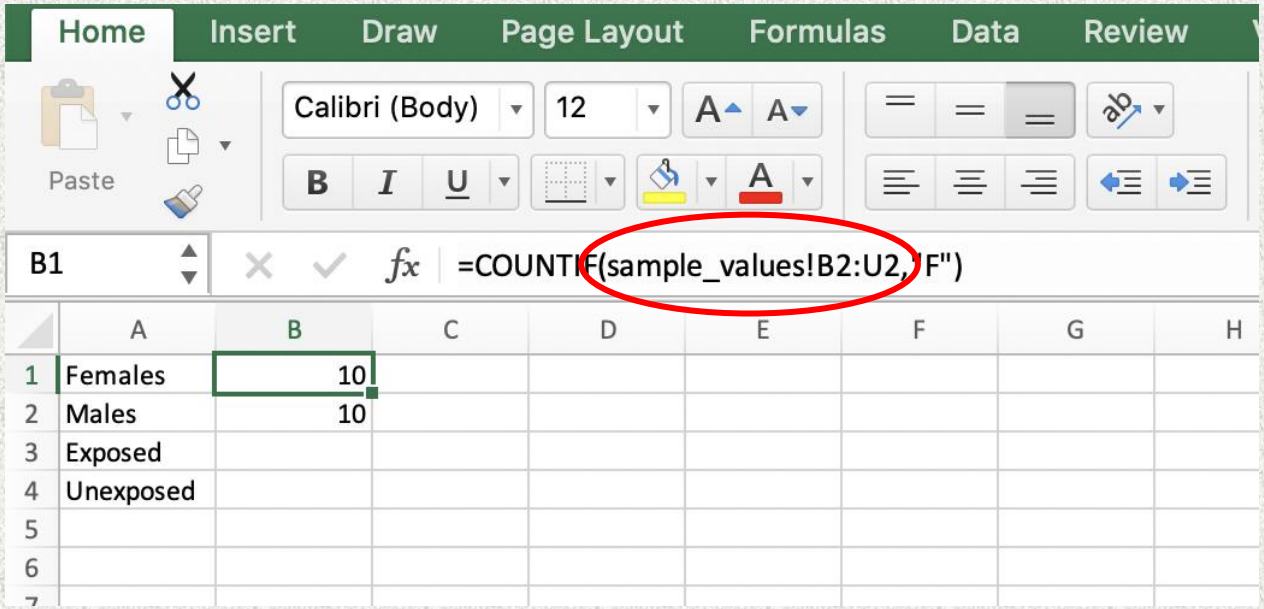
We want to count the number of females.

Range here is the second row of the sheet sample_values:

B2:U2 (cells B2 through U2)

or 2:2 (select entire row)

The colon (:) means "through"



	A	B	C	D	E	F	G	H
1	Females	10						
2	Males	10						
3	Exposed							
4	Unexposed							
5								
6								

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Sample_ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2	Sex	F	F	M	F	M	M	M	F	F	M	F	M	M	F	M	F	M	F	M	F
3	Status	N	Y	Y	N	N	N	Y	Y	N	Y	Y	N	Y	N	N	Y	Y	Y	N	N
4	Cat1	3	7	9	3	1	4	7	10	6	9	1	8	8	1	1	2	5	2	8	1
5	Cat2	0.39980403	0.131167	0.83797435	0.99121998	0.37915433	0.4932649	0.31730995	0.78559809	0.29231847	0.22800014	0.12970948	0.20895121	0.09570753	0.57328551	0.74719656	0.97775537	0.90570917	0.23738616	0.63079488	0.03630628
6	Cat3	0.01765164	0.74075219	0.4243295	0.84309817	0.41803328	0.36549674	0.08057165	0.73402944	0.9952279	0.6285995	0.66614321	0.15886006	0.29619534	0.522676	0.49873951	0.499717	0.57727393	0.35131784	0.24743548	0.36314088
7	Cat4	76	75	65	53	89	66	88	78	98	67	75	96	90	65	70	89	73	84	85	71
8	Cat5	3	10	3	1	7	5	5	3	3	3	8	6	7	5	8	7	4	2	9	9
9	Cat6	0.27817695	0.07110137	0.37747939	0.13183529	0.31935773	0.1925463	0.63984189	0.64802318	0.26027957	0.93419475	0.52918677	0.2509355	0.39102529	0.77463581	0.59241452	0.48270423	0.83625265	0.41941355	0.83425559	0.93651799
10	Cat7	0.47111146	0.89977883	0.42504482	0.09046157	0.43938449	0.85181114	0.7906725	0.03804672	0.98191601	0.06929245	0.17350784	0.00365075	0.62290958	0.10161224	0.129711	0.87430617	0.63898328	0.72340885	0.87721221	0.74542847
11	Cat8	59	79	81	65	54	91	50	92	96	51	80	63	78	70	91	60	80	59	81	92
12	Cat9	0.15067873	0.69674031	0.2607653	0.54603649	0.94019016	0.63397075	0.31429126	0.1763019	0.87520766	0.0101267	0.8489138	0.81448124	0.14730753	0.47184386	0.97195415	0.10610593	0.55886697	0.34184183	0.38806705	0.24976963
13	Cat10	1646	1436	1486	1218	1151	1278	1329	1942	1085	1662	1748	1922	1061	1850	1160	1166	1049	1419	1275	1207
14	Cat11	535	586	543	470	479	502	582	537	412	514	462	487	543	405	504	478	466	419	467	591
15	Cat12	13	19	15	18	11	19	18	10	16	11	15	19	20	18	13	17	16	14	14	14

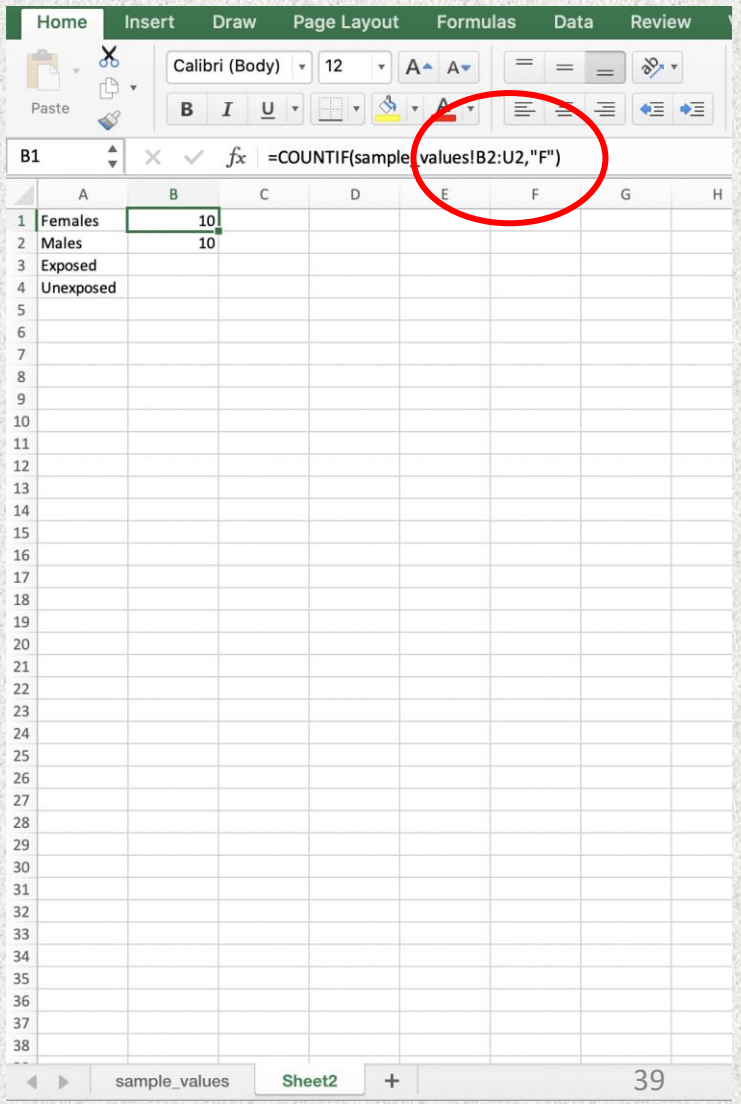
How many females are in this data set?

COUNTIF function counts how many cells in a given range ***have a certain value***

The value we want to count is number of females.

After we specify the range, add a comma (,) and then the value we want to count.

"F"



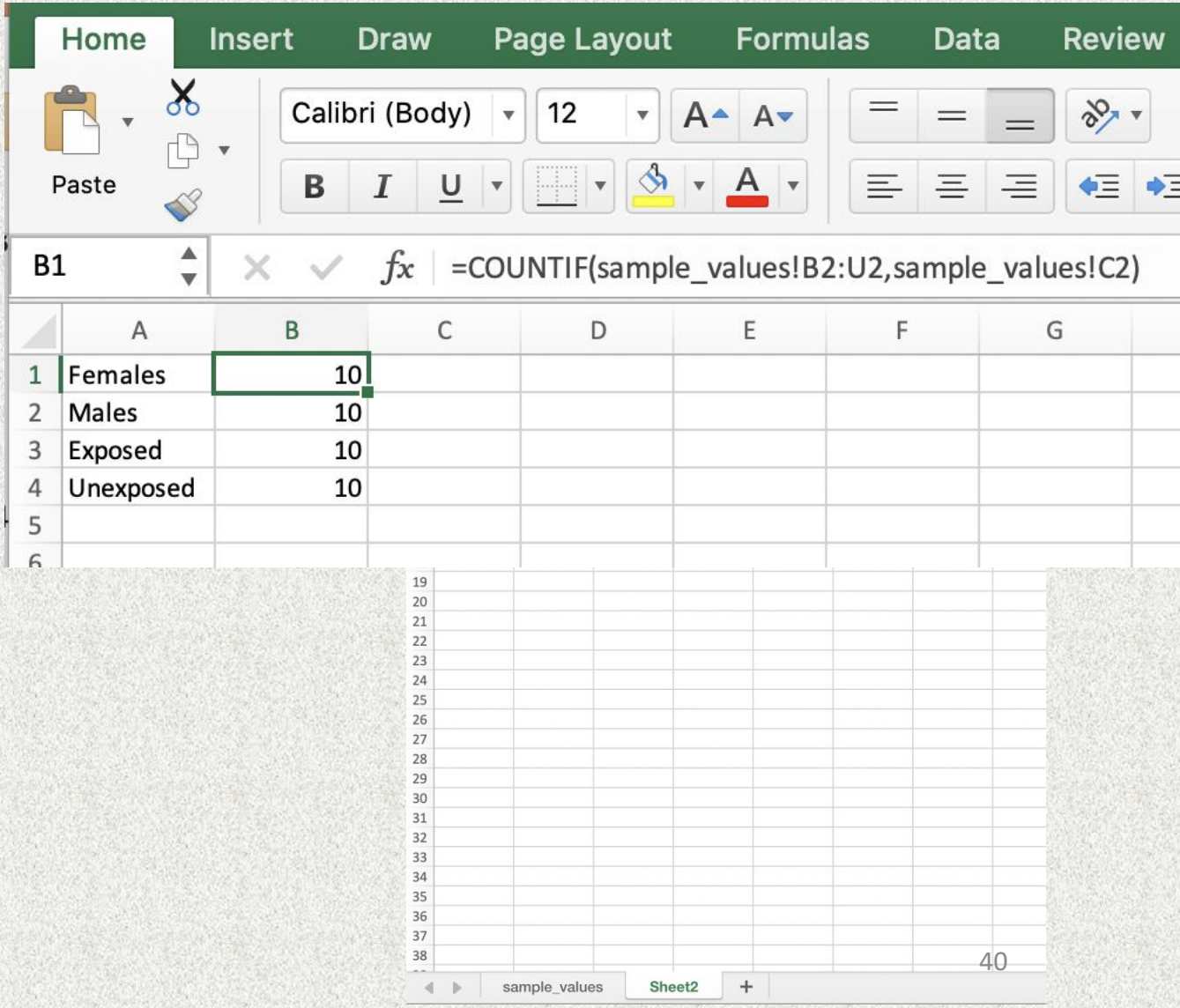
How many females are in this data set?

COUNTIF function counts how many cells in a given range ***have a certain value***

Another option: after we specify the range, add a comma (,) and then a cell containing the value we want to count.

sample_values!C2

("C2" will tell Excel to count cells containing C2 rather than the contents of cell C2)



QUESTION 1…

- Which formula should I use if I want to know how many individuals were exposed (status is Y)?

A: =COUNTIF(B3:U3,”Y”)

B: =COUNTIF(sample_valuesB3:U3,Y)

C: COUNTIF(sample_values!B3:U3,”Y”)

D: =COUNTIF(sample_values!B3:U3,C3)

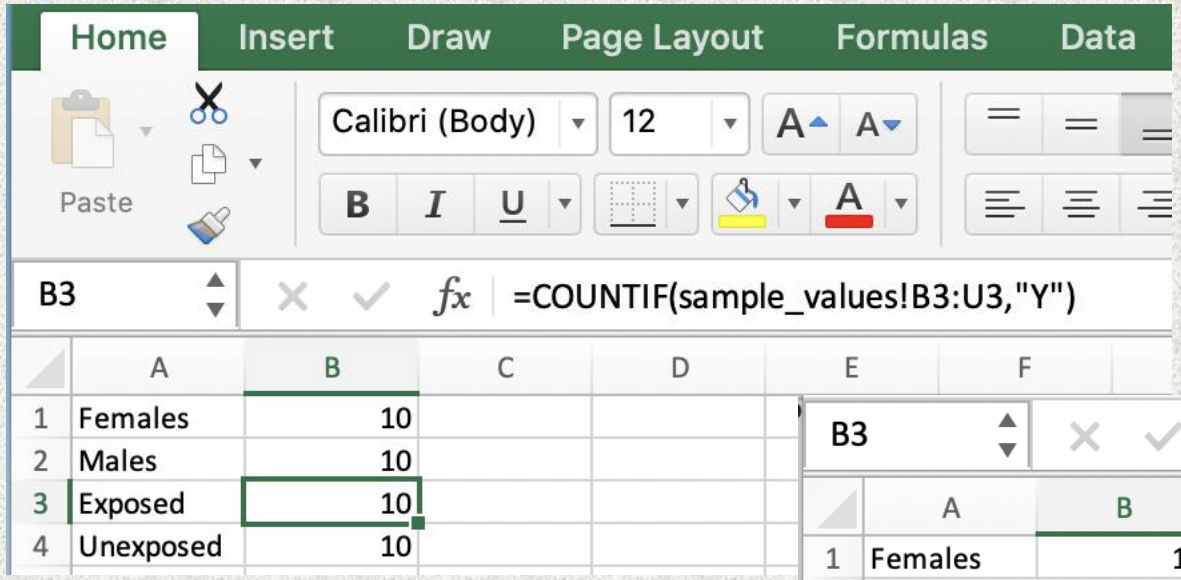
E: =COUNTIF(sample_values!B3:U3,”Y”)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Sample_ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2	Sex	F	F	M	F	M	M	M	F	F	M	F	M	M	F	M	F	M	F	M	F
3	Status	N	Y	Y	N	N	N	Y	Y	N	Y	Y	N	Y	N	N	Y	Y	Y	N	N
4	Cat1	3	7	9	3	1	4	7	10	6	9	1	8	8	1	1	2	5	2	8	1
5	Cat2	0.39980403	0.131167	0.83797435	0.99121998	0.37915433	0.4932649	0.31730995	0.78559809	0.29231847	0.22800014	0.12970948	0.20895121	0.09570753	0.57328551	0.74719656	0.97775537	0.90570917	0.23738616	0.63079488	0.03630628
6	Cat3	0.01765164	0.74075219	0.4243295	0.84309817	0.41803328	0.36549674	0.08057165	0.73402944	0.9952279	0.6285995	0.66614321	0.15886006	0.29619534	0.522676	0.49873951	0.499717	0.57727393	0.35131784	0.24743548	0.36314088
7	Cat4	76	75	65	53	89	66	88	78	98	67	75	96	90	65	70	89	73	84	85	71
8	Cat5	3	10	3	1	7	5	5	3	3	3	8	6	7	5	8	7	4	2	9	9
9	Cat6	0.27817695	0.07110137	0.37747939	0.13183529	0.31935773	0.1925463	0.63984189	0.64802318	0.26027957	0.93419475	0.52918677	0.2509355	0.39102529	0.77463581	0.59241452	0.48270423	0.83625265	0.41941355	0.83425559	0.93651799
10	Cat7	0.47111146	0.89977883	0.42504482	0.09046157	0.43938449	0.85181114	0.7906725	0.03804672	0.98191601	0.06929245	0.17350784	0.00365075	0.62290958	0.10161224	0.129711	0.87430617	0.63898328	0.72340885	0.87721221	0.74542847
11	Cat8	59	79	81	65	54	91	50	92	96	51	80	63	78	70	91	60	80	59	81	92
12	Cat9	0.15067873	0.69674031	0.2607653	0.54603649	0.94019016	0.63397075	0.31429126	0.1763019	0.87520766	0.0101267	0.8489138	0.81448124	0.14730753	0.47184386	0.97195415	0.10610593	0.55886697	0.34184183	0.38806705	0.24976963
13	Cat10	1646	1436	1486	1218	1151	1278	1329	1942	1085	1662	1748	1922	1061	1850	1160	1166	1049	1419	1275	1207
14	Cat11	535	586	543	470	479	502	582	537	412	514	462	487	543	405	504	478	466	419	467	591
15	Cat12	13	19	15	18	11	19	18	10	16	11	15	19	20	18	13	17	16	14	14	14

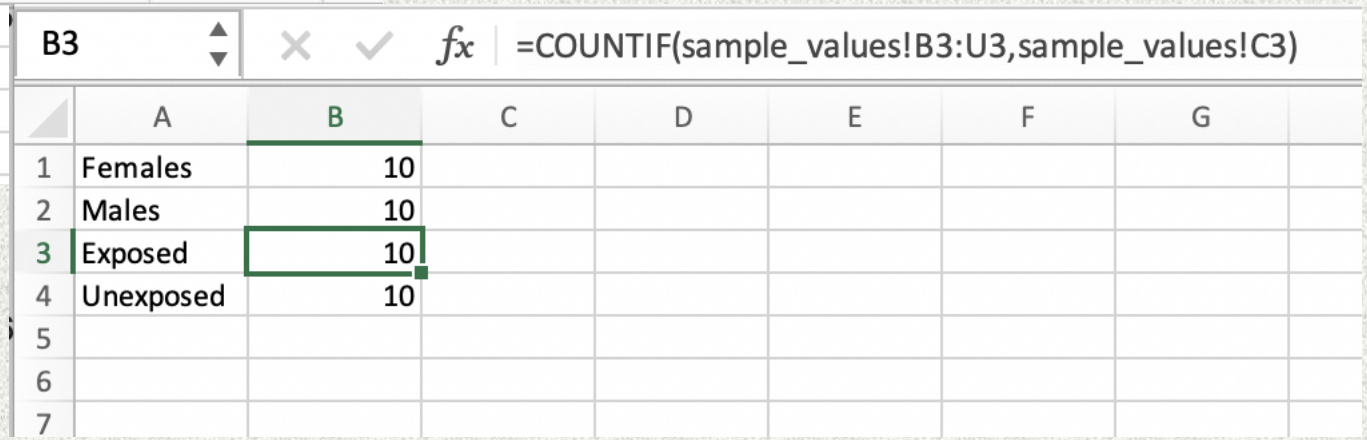
How to count how many were exposed

A: =COUNTIF(B3:U3,”Y”) ← did not specify the correct sheet where the data was located

B: =COUNTIF(sample_valuesB3:U3,Y)
C: COUNTIF(sample_values!B3:U3,”Y”)
D: =COUNTIF(sample_values!B3:U3,C3)
E: =COUNTIF(sample_values!B3:U3,”Y”)



	A	B	C	D	E	F
1	Females	10				
2	Males	10				
3	Exposed	10				
4	Unexposed	10				



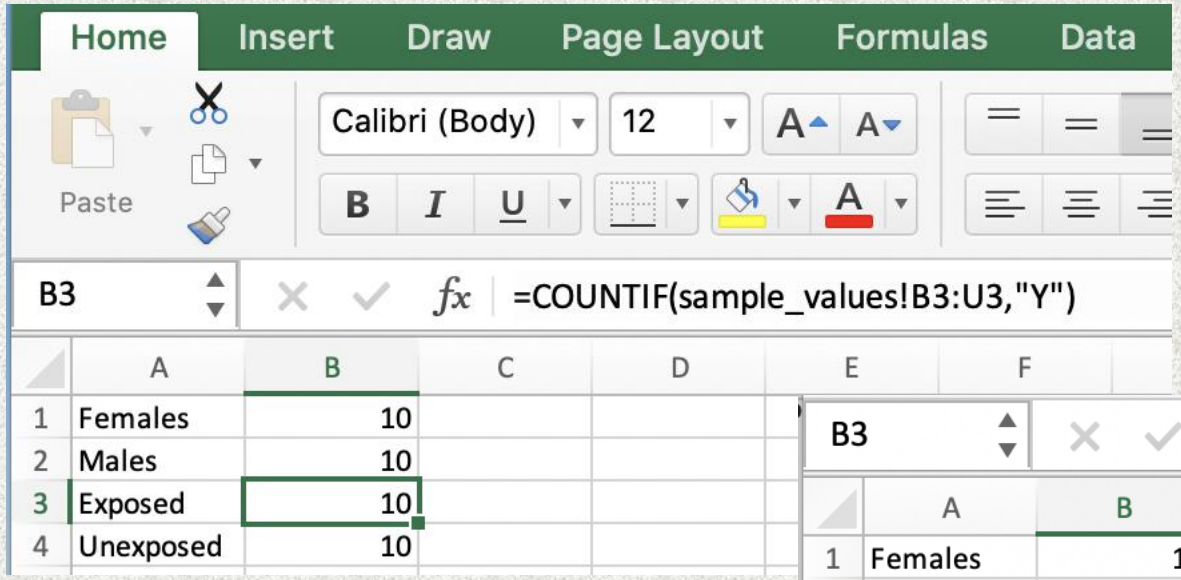
	A	B	C	D	E	F	G
1	Females	10					
2	Males	10					
3	Exposed	10					
4	Unexposed	10					
5							
6							
7							

How to count how many were exposed

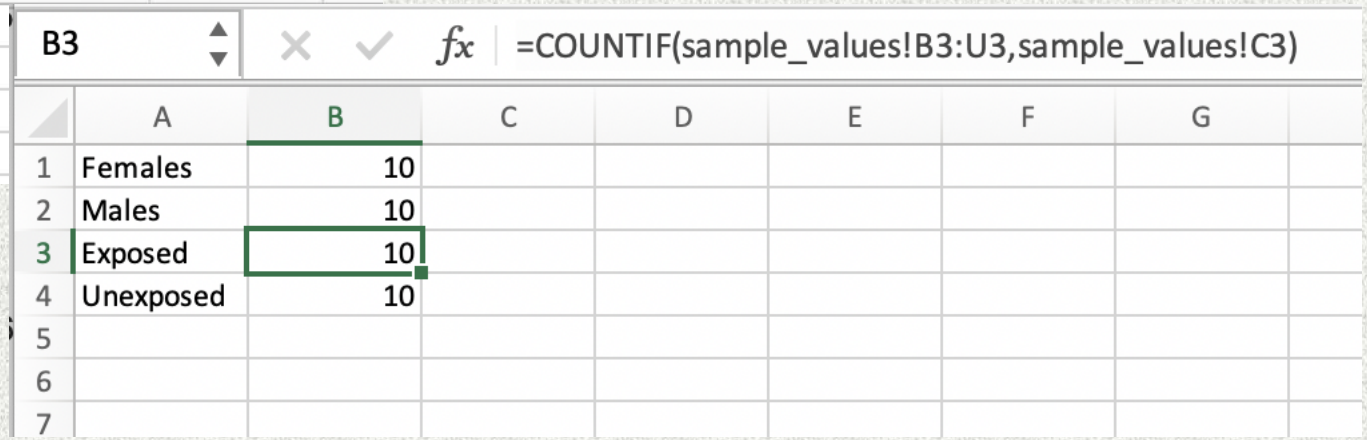
A: =COUNTIF(B3:U3,”Y”)

B: =COUNTIF(sample_valuesB3:U3,Y) ← there is no exclamation mark after the sheet name

C: COUNTIF(sample_values!B3:U3,”Y”)
D: =COUNTIF(sample_values!B3:U3,C3)
E: =COUNTIF(sample_values!B3:U3,”Y”)



	A	B	C	D	E	F
1	Females	10				
2	Males	10				
3	Exposed	10				
4	Unexposed	10				



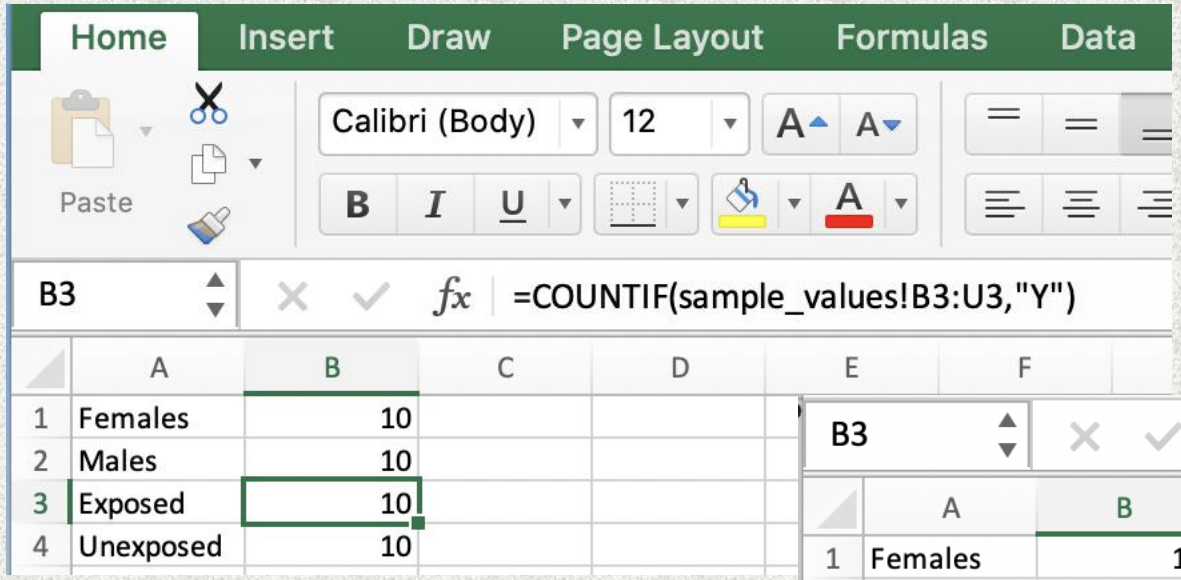
	A	B	C	D	E	F	G
1	Females	10					
2	Males	10					
3	Exposed	10					
4	Unexposed	10					
5							
6							
7							

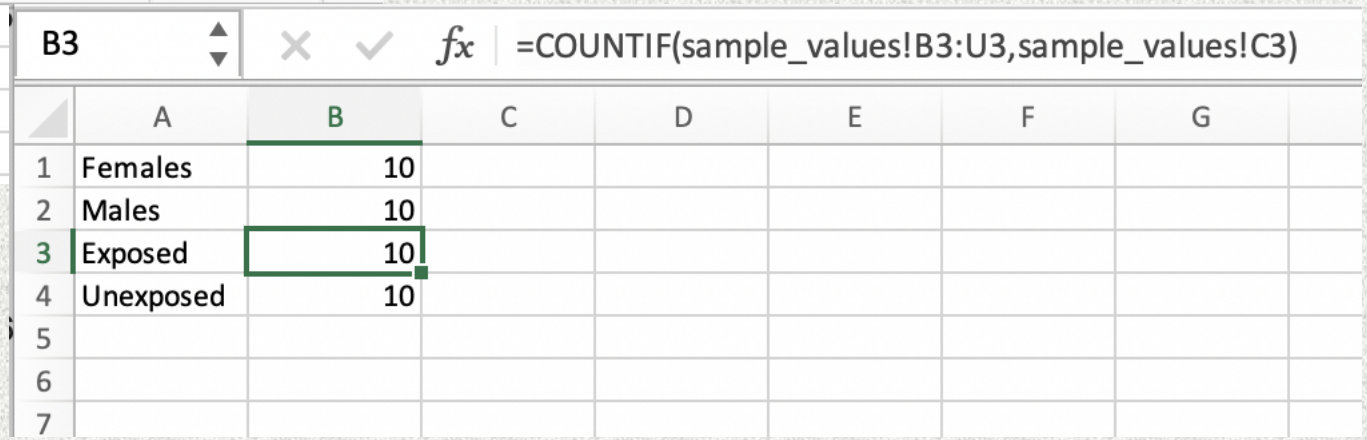
How to count how many were exposed

A: =COUNTIF(B3:U3,”Y”)
B: =COUNTIF(sample_valuesB3:U3,Y)

C: COUNTIF(sample_values!B3:U3,”Y”) ← there is no equals sign before COUNTIF

D: =COUNTIF(sample_values!B3:U3,C3)
E: =COUNTIF(sample_values!B3:U3,”Y”)



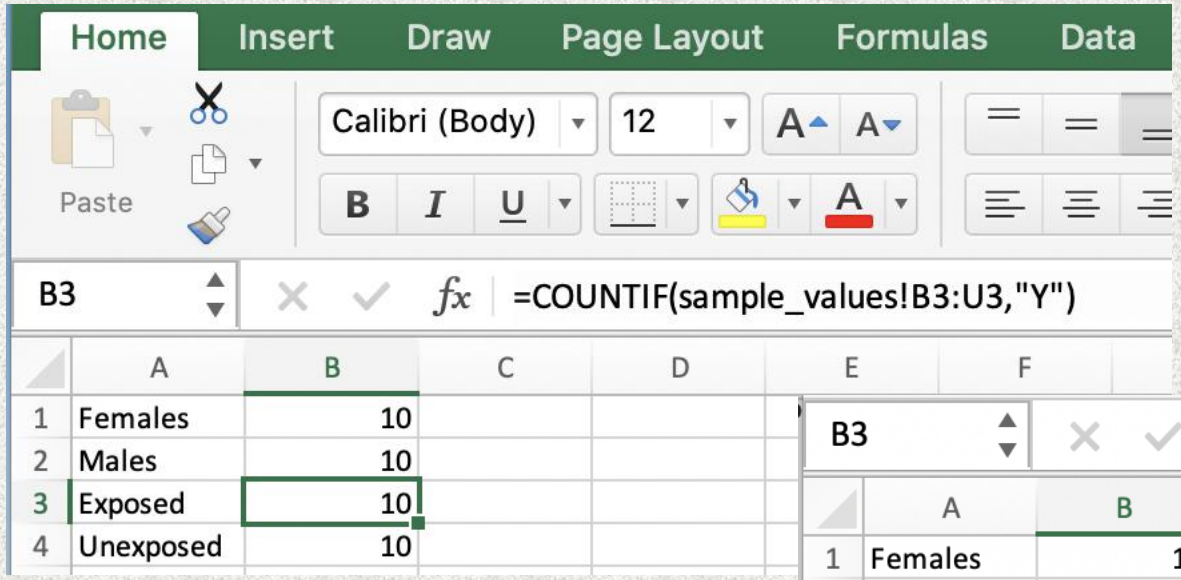


How to count how many were exposed

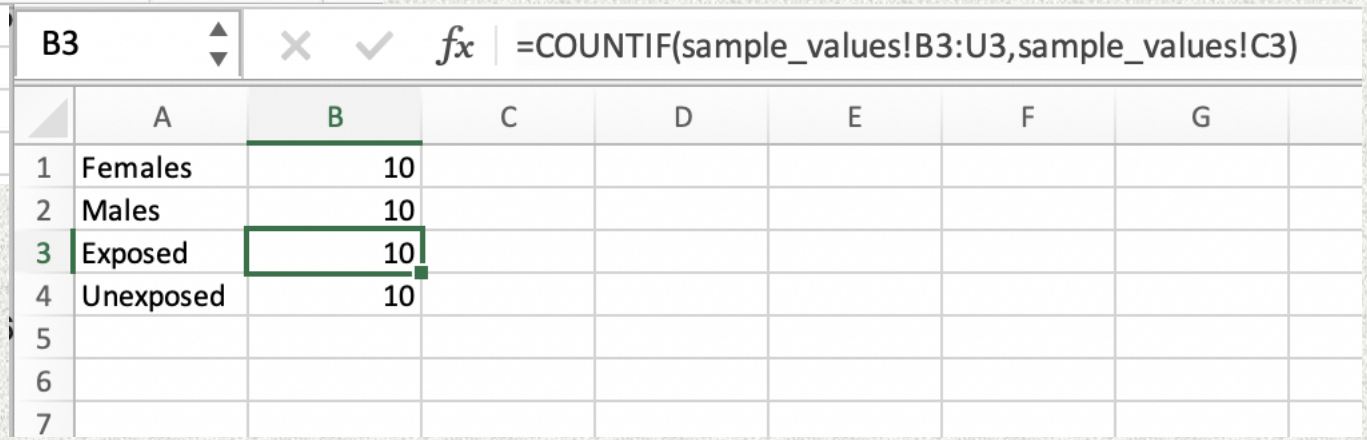
A: =COUNTIF(B3:U3,”Y”)
B: =COUNTIF(sample_valuesB3:U3,Y)
C: COUNTIF(sample_values!B3:U3,”Y”)

D: =COUNTIF(sample_values!B3:U3,C3) → The value for cell C3 will be searched for based on whatever sheet you’re on, rather than the sheet with the data

E: =COUNTIF(sample_values!B3:U3,”Y”)



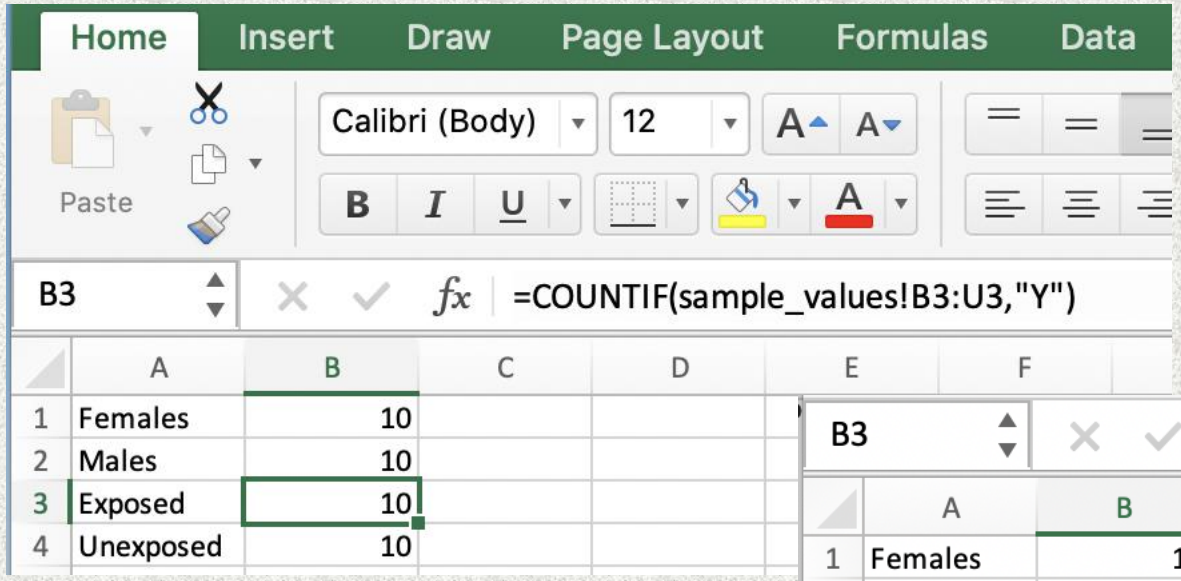
	A	B	C	D	E	F
1	Females	10				
2	Males	10				
3	Exposed	10				
4	Unexposed	10				



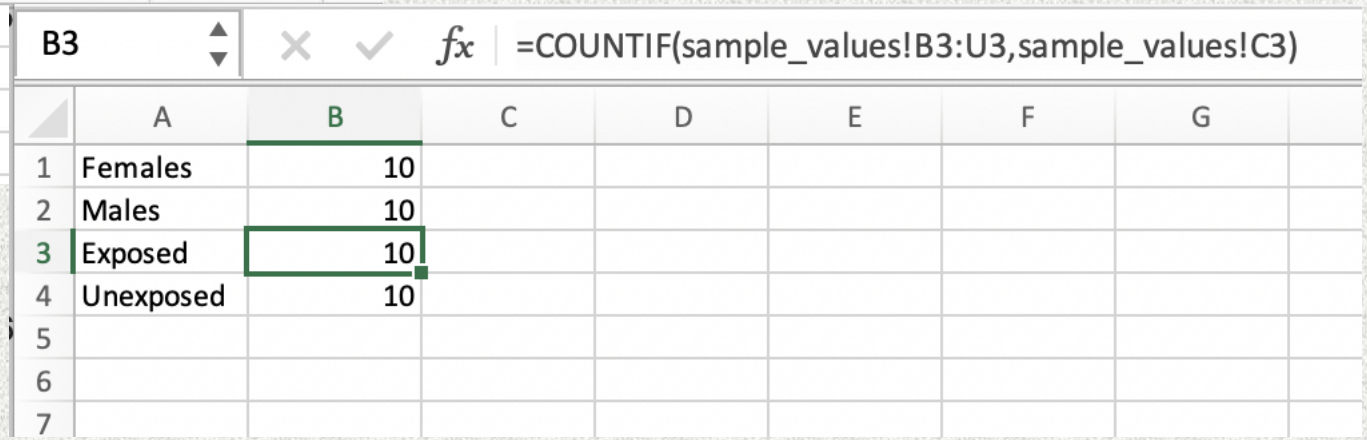
	A	B	C	D	E	F	G
1	Females	10					
2	Males	10					
3	Exposed	10					
4	Unexposed	10					
5							
6							
7							

How to count how many were exposed

A: =COUNTIF(B3:U3,”Y”)
B: =COUNTIF(sample_valuesB3:U3,Y)
C: COUNTIF(sample_values!B3:U3,”Y”)
D: =COUNTIF(sample_values!B3:U3,C3)
E: =COUNTIF(sample_values!B3:U3,”Y”) ← correct!
This will count how many individuals have “Y” in the Status row.
Could also search for the value in cell C3 of the sample_values sheet, which is also Y.



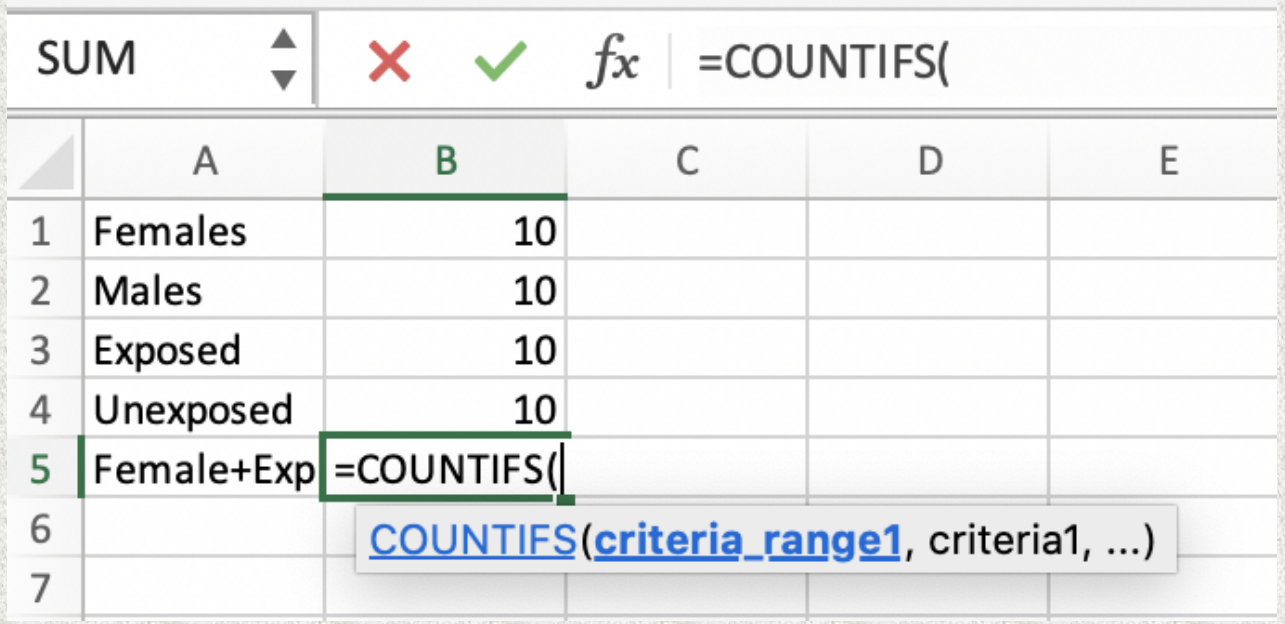
	A	B	C	D	E	F
1	Females	10				
2	Males	10				
3	Exposed	10				
4	Unexposed	10				



	A	B	C	D	E	F	G
1	Females	10					
2	Males	10					
3	Exposed	10					
4	Unexposed	10					
5							
6							
7							

What if I want to know *how many females were exposed*?

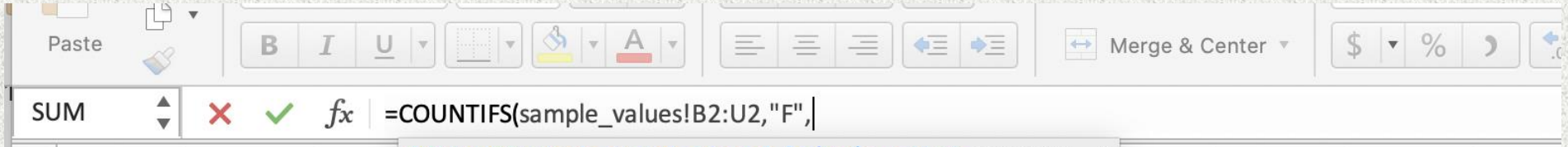
- You can "nest" multiple COUNTIF functions by using **COUNTIFS**



You still need to follow the same basic procedures as before.

How many females were exposed?

- You can "nest" multiple COUNTIF functions by using COUNTIFS



	A	B	C	D	E	F	G	H	I	J	K	L
1	Sample_ID	1	2	3	4	5	6	7	8	9	10	
2	Sex	F	F	M	F	M	M	M	F	F	M	F
3	Status	N	Y	Y	N	N	N	Y	Y	N	Y	Y
4	Cat1	3	7	9	3	1	4	7	10	6	9	
5	Cat2	=COUNTIFS(sample_values!B2:U2,"F",				0.37915433	0.4932649	0.31730995	0.78559809	0.29231847	0.22800014	0.12970
6	Cat3	0.01765164	0.74075219	0.4243295	0.84309817	0.41803328	0.36549674	0.08057165	0.73402944	0.9952279	0.6285995	0.66614
7	Cat4	76	75	65	53	89	66	88	78	98	67	
8	Cat5	3	10	3	1	7	5	5	3	3	3	
9	Cat6	0.27817695	0.07110137	0.37747939	0.13183529	0.31935773	0.1925463	0.63984189	0.64802318	0.26027957	0.93419475	0.52918
10	Cat7	0.47111146	0.89977883	0.42504482	0.09046157	0.43938449	0.85181114	0.7906725	0.03804672	0.98191601	0.06929245	0.17350
11	Cat8	59	79	81	65	54	91	50	92	96	51	
12	Cat9	0.15067873	0.69674031	0.2607653	0.54603649	0.94019016	0.63397075	0.31429126	0.1763019	0.87520766	0.0101267	0.8489
13	Cat10	1646	1436	1486	1218	1151	1278	1329	1942	1085	1662	1
14	Cat11	535	586	543	470	479	502	582	537	412	514	
15	Cat12	13	19	15	18	11	19	18	10	16	11	
16												
17												
18												
19												

Criteria_range1
= where to look
(first criteria)

Criteria1
= what to look for
(first criteria

How many females were exposed?

SUM | =COUNTIFS(sample_values!B2:U2,"F",sample_values!B3:U3,"Y")

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Sample_ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2	Sex	F	F	M	F	M	M	M	F	F	M	F	M	M	F	M	F	M	F	M	F
3	Status	N	Y	Y	N	N	N	Y	Y	N	Y	Y	N	Y	N	N	Y	Y	Y	N	N
4	Cat1	3	7	9	3	1	4	7	10	6	9	1	8	8	1	1	2	5	2	8	1
5	Cat2	=COUNTIFS(sample_values!B2:U2,"F",sample_values!B3:U3,"Y")						0.31730995	0.78559803	0.29231847	0.22800014	0.12970948	0.20895121	0.09570753	0.57328551	0.74719656	0.97775537	0.90570917	0.23738616	0.63079488	0.03630628
6	Cat3	0.01765164	0.74075219	0.4263295	0.84309817	0.41803328	0.36549674	0.08057165	0.73402944	0.9952279	0.6285995	0.66614321	0.15886006	0.29619534	0.522676	0.49873951	0.499717	0.57727393	0.35131784	0.24743548	0.36314088
7	Cat4	76	75	65	53	89	66	88	78	98	67	75	96	90	65	70	89	73	84	85	71
8	Cat5	3	10	3	1	7	5	5	3	3	3	8	6	7	5	8	7	4	2	9	9
9	Cat6	0.27817695	0.07110137	0.87747939	0.13183529	0.31935773	0.1925463	0.63984189	0.64802318	0.26027957	0.93419475	0.52918677	0.2503355	0.39102529	0.77463581	0.59241452	0.48270423	0.83625265	0.41941355	0.83425559	0.93651799
10	Cat7	0.47111146	0.89977883	0.42504482	0.09046157	0.43938449	0.85181114	0.7906725	0.03804672	0.98191601	0.06329245	0.17350784	0.00365075	0.62290958	0.10161224	0.129711	0.87430617	0.63898328	0.72340885	0.87721221	0.74542847
11	Cat8	59	79	81	65	54	91	50	92	96	51	80	63	78	70	91	60	80	59	81	92
12	Cat9	0.15067873	0.69674034	0.2607653	0.54603649	0.94019016	0.63397075	0.31429126	0.1763019	0.87520766	0.0101267	0.8489138	0.81448124	0.14730753	0.47184386	0.97195415	0.10610593	0.55886697	0.34184183	0.38806705	0.24976963
13	Cat10	1646	1486	1486	1218	1151	1278	1329	1942	1085	1662	1748	1922	1061	1850	1160	1166	1049	1419	1275	1207
14	Cat11	535	586	543	470	479	502	582	537	412	514	462	487	543	405	504	478	466	419	467	591
15	Cat12	13	19	15	18	11	19	18	10	16	11	13	19	20	18	13	17	16	14	14	14
16																					

(Criteria range1 = where to look, Criteria1 = what to look for, Criteria range2 = where to look, Criteria2 = what to look for)

How many females were exposed?

- Our answer, found using COUNTIFS!

SUM =COUNTIFS(sample_values!B2:U2,"F",sample_values!B3:U3,"Y")

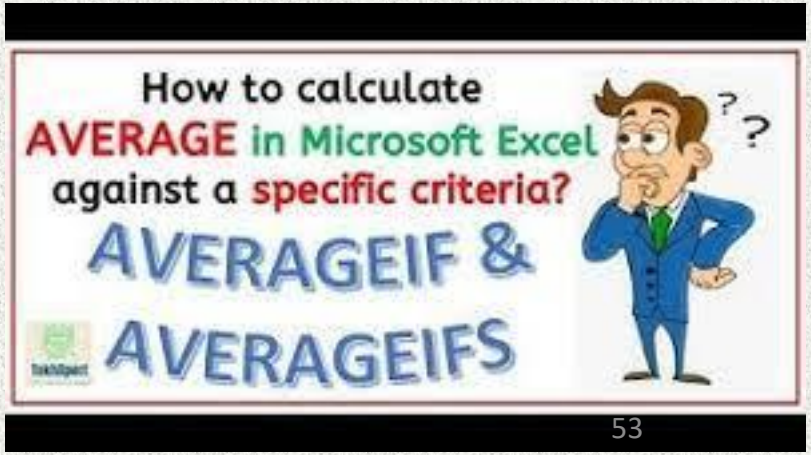
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Sample_ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2	Sex	F	F	M	F	M	M	M	F	F	M	F	M	M	F	M	F	M	F	M	F
3	Status	N	Y	Y	N	N	N	Y	Y	N	Y	Y	N	Y	N	N	Y	Y	Y	N	N
4	Cat1	3	7	9	3	1	4	7	10	6	9	1	8	8	1	1	2	5	2	8	1
5	Cat2	=COUNTIFS(sample_values!B2:U2,"F",sample_values!B3:U3,"Y")						0.31730995	0.78559809	0.29231847	0.22800014	0.12970948	0.20895121	0.09570753	0.57328551	0.74719656	0.97775537	0.90570917	0.23738616	0.63079488	0.03630628
6	Cat3	0.01765164	0.74075219	0.4243295	0.84309817	0.41803328	0.36549674	0.08057165	0.73402944	0.9952279	0.6285995	0.66614321	0.15886006	0.29619534	0.522676	0.49873951	0.499717	0.57727393	0.35131784	0.24743548	0.36314088
7	Cat4	76	75	65	53	89	66	88	78	98	67	75	96	90	65	70	89	73	84	85	71
8	Cat5	3	10	3	1	7	5	5	3	3	3	8	6	7	5	8	7	4	2	9	9
9	Cat6	0.27817695	0.07110137	0.37747939	0.13183529	0.31935773	0.1925463	0.63984189	0.64802318	0.26027957	0.93419475	0.52918677	0.2509355	0.39102529	0.77463581	0.59241452	0.48270423	0.83625265	0.41941355	0.83425559	0.93651799
10	Cat7	0.47111146	0.89977883	0.42504482	0.09046157	0.43938449	0.85181114	0.7906725	0.03804672	0.98191601	0.06929245	0.17350784	0.00365075	0.62290958	0.10161224	0.129711	0.87430617	0.63898328	0.72340885	0.87721221	0.74542847
11	Cat8	59	79	81	65	54	91	50	92	96	51	80	63	78	70	91	60	80	59	81	92
12	Cat9	0.15067873	0.69674031	0.2607653	0.54603649	0.94019016	0.63397075	0.31429126	0.1763019	0.87520766	0.0101267	0.8489138	0.81448124	0.14730753	0.47184386	0.97195415	0.10610593	0.55886697	0.34184183	0.38806705	0.24976963
13	Cat10	1646	1436	1486														1049	1419	1275	1207
14	Cat11	535	586	543														466	419	467	591
15	Cat12	13	19	15														16	14	14	14
16																					

B5 =COUNTIFS(sample_values!B2:U2,"F",sample_values!B3:U3,"Y")

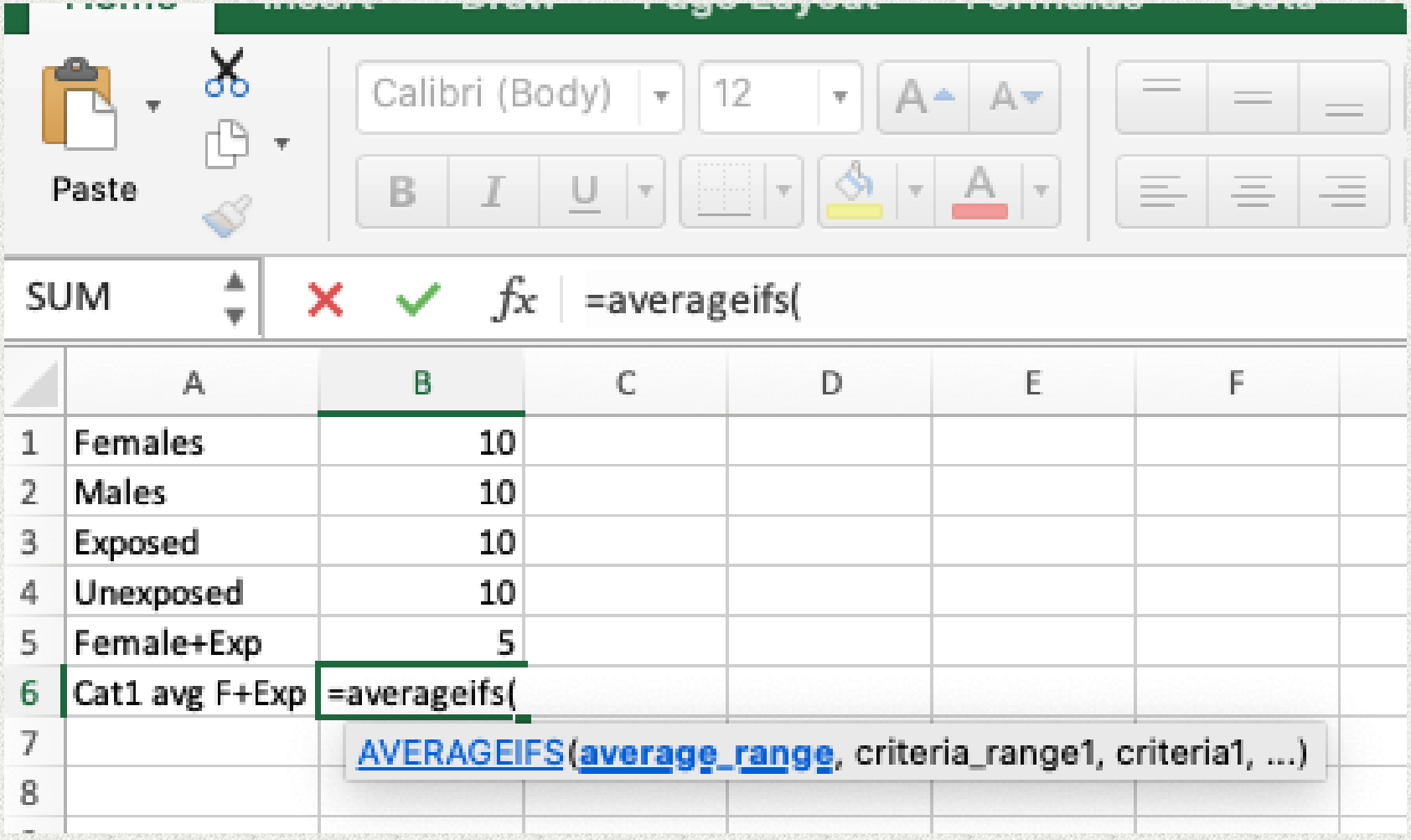
	A	B	C	D	E	F	G	H
1	Females	10						
2	Males	10						
3	Exposed	10						
4	Unexposed	10						
5	Female+Exp	5						
6								
7								

What if I want to know the average Cat1 value for all females that were exposed?

- AVERAGEIF: averages cells that meet one criteria
- AVERAGEIFS: averages cells that meet multiple criteria
- =AVERAGEIFS(range you want to average, first range of interest, what to look for in first range, second range of interest, what to look for in the second range….)
  - Range we want to average: **Cat1**
  - First range of interest: sex
    - Look for "**F**"
  - Second range of interest: exposure status
    - Look for "**Y**"



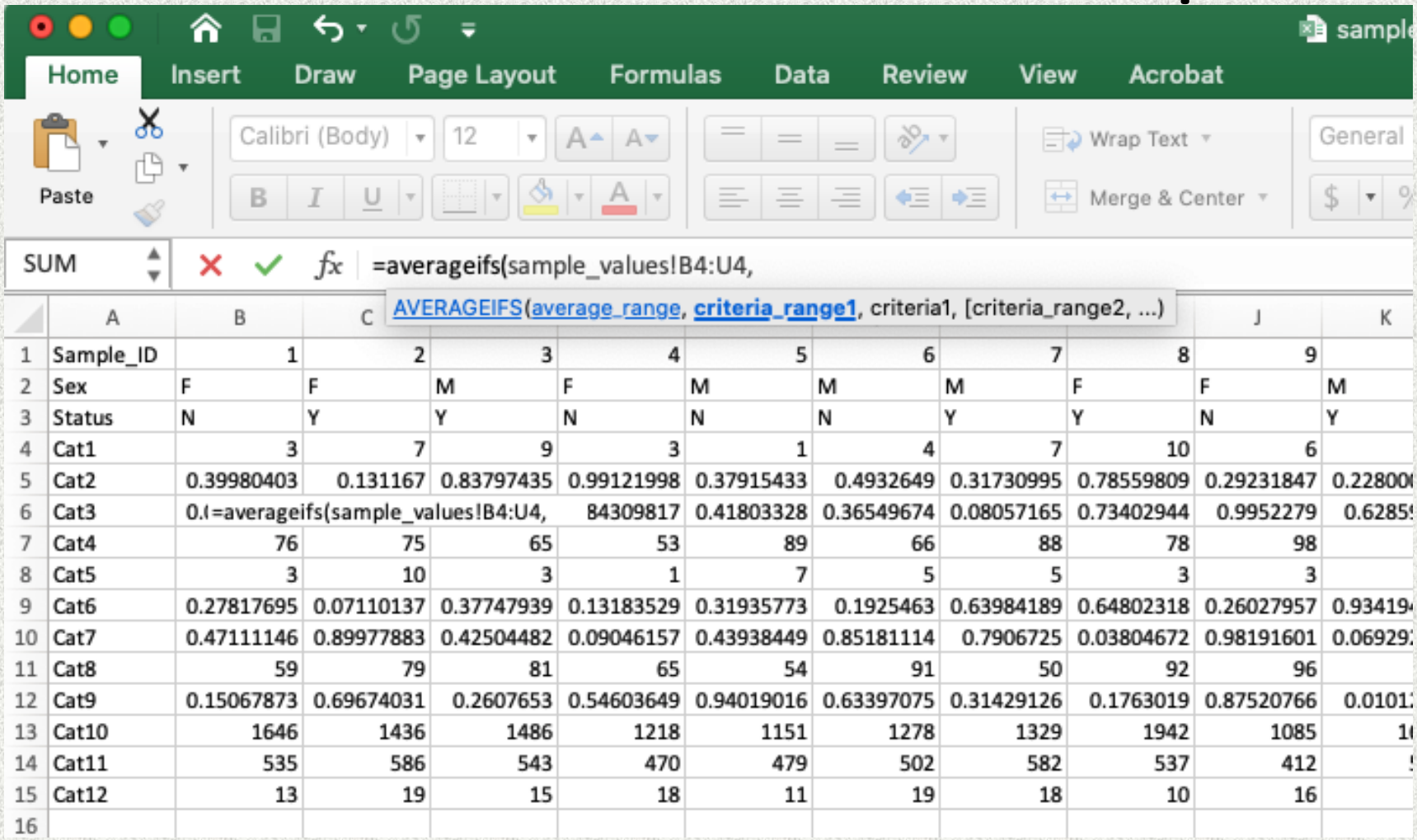
What if I want to know the average Cat1 value for all females that were exposed?



	A	B	C	D	E	F
1	Females	10				
2	Males	10				
3	Exposed	10				
4	Unexposed	10				
5	Female+Exp	5				
6	Cat1 avg F+Exp	=averageifs(				
7						
8						

On Sheet1, add a place for **AVERAGEIFS** calculation…

What if I want to know the average Cat1 value for all females that were exposed?

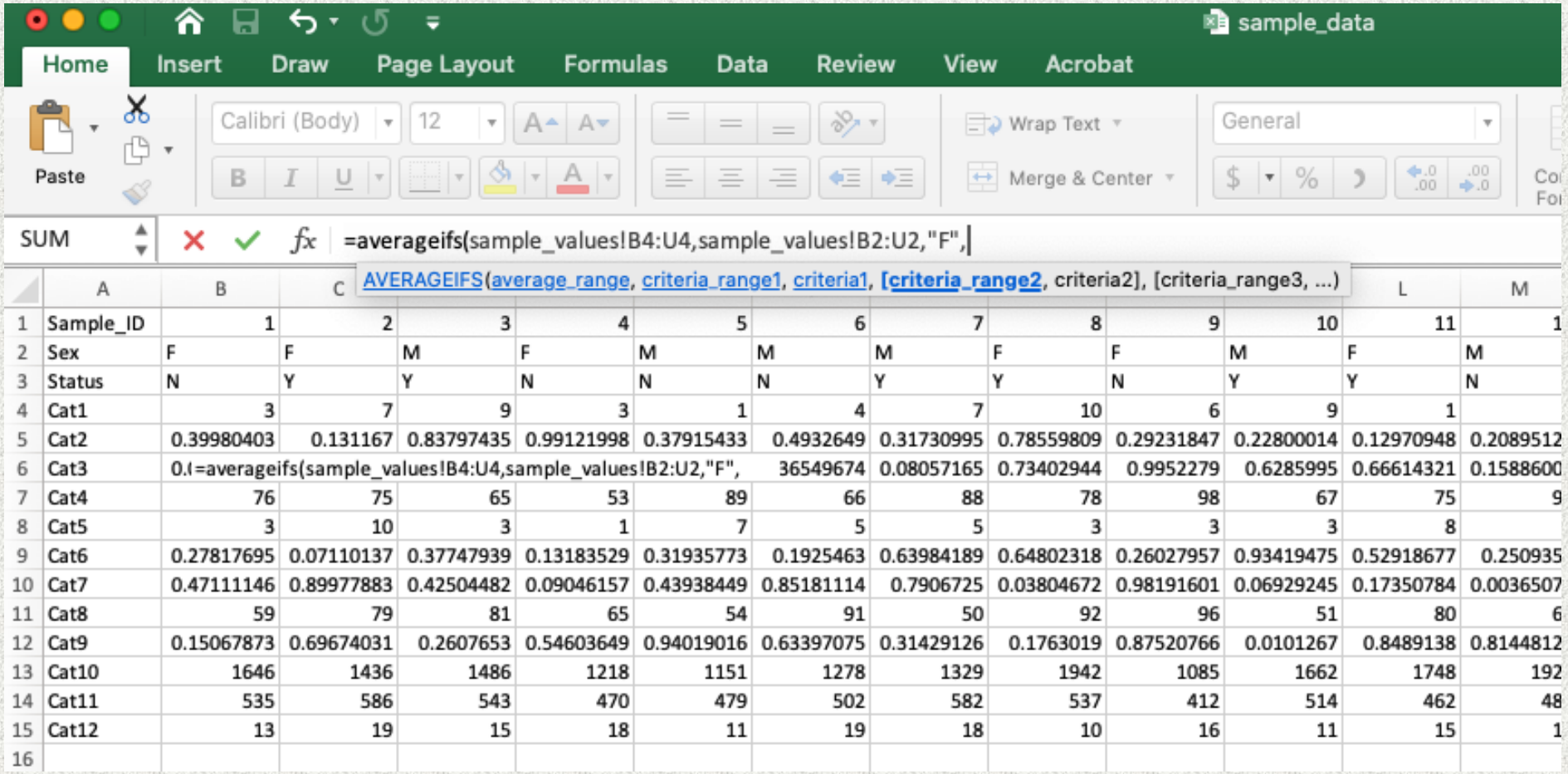


	A	B	C	D	E	F	G	H	I	J	K
1	Sample_ID	1	2	3	4	5	6	7	8	9	
2	Sex	F	F	M	F	M	M	M	F	F	M
3	Status	N	Y	Y	N	N	N	Y	Y	N	Y
4	Cat1	3	7	9	3	1	4	7	10	6	
5	Cat2	0.39980403	0.131167	0.83797435	0.99121998	0.37915433	0.4932649	0.31730995	0.78559809	0.29231847	0.22800
6	Cat3	0.(	=averageifs(sample_values!B4:U4,		84309817	0.41803328	0.36549674	0.08057165	0.73402944	0.9952279	0.6285
7	Cat4	76	75	65	53	89	66	88	78	98	
8	Cat5	3	10	3	1	7	5	5	3	3	
9	Cat6	0.27817695	0.07110137	0.37747939	0.13183529	0.31935773	0.1925463	0.63984189	0.64802318	0.26027957	0.93419
10	Cat7	0.47111146	0.89977883	0.42504482	0.09046157	0.43938449	0.85181114	0.7906725	0.03804672	0.98191601	0.06929
11	Cat8	59	79	81	65	54	91	50	92	96	
12	Cat9	0.15067873	0.69674031	0.2607653	0.54603649	0.94019016	0.63397075	0.31429126	0.1763019	0.87520766	0.0101
13	Cat10	1646	1436	1486	1218	1151	1278	1329	1942	1085	1
14	Cat11	535	586	543	470	479	502	582	537	412	
15	Cat12	13	19	15	18	11	19	18	10	16	
16											

Cat1 values are in the **sample_values** sheet, **row 4**.

We are telling Excel, “The numbers I want averaged are located here.”

What if I want to know the average Cat1 value for all females that were exposed?

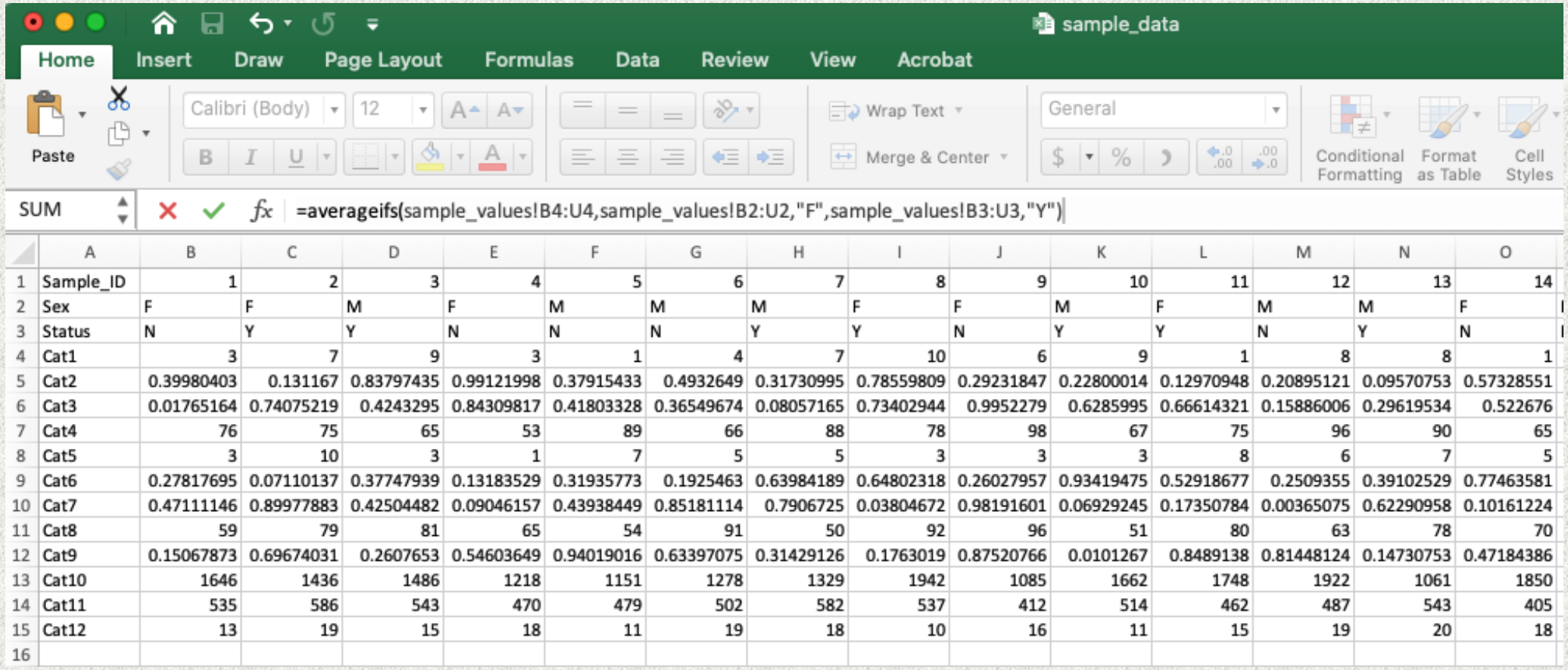


	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Sample_ID	1	2	3	4	5	6	7	8	9	10	11	1
2	Sex	F	F	M	F	M	M	M	F	F	M	F	M
3	Status	N	Y	Y	N	N	N	Y	Y	N	Y	Y	N
4	Cat1	3	7	9	3	1	4	7	10	6	9	1	
5	Cat2	0.39980403	0.131167	0.83797435	0.99121998	0.37915433	0.4932649	0.31730995	0.78559809	0.29231847	0.22800014	0.12970948	0.2089512
6	Cat3	0.(	=averageifs(sample_values!B4:U4,sample_values!B2:U2,"F",				36549674	0.08057165	0.73402944	0.9952279	0.6285995	0.66614321	0.1588600
7	Cat4	76	75	65	53	89	66	88	78	98	67	75	9
8	Cat5	3	10	3	1	7	5	5	3	3	3	8	
9	Cat6	0.27817695	0.07110137	0.37747939	0.13183529	0.31935773	0.1925463	0.63984189	0.64802318	0.26027957	0.93419475	0.52918677	0.250935
10	Cat7	0.47111146	0.89977883	0.42504482	0.09046157	0.43938449	0.85181114	0.7906725	0.03804672	0.98191601	0.06929245	0.17350784	0.0036507
11	Cat8	59	79	81	65	54	91	50	92	96	51	80	6
12	Cat9	0.15067873	0.69674031	0.2607653	0.54603649	0.94019016	0.63397075	0.31429126	0.1763019	0.87520766	0.0101267	0.8489138	0.8144812
13	Cat10	1646	1436	1486	1218	1151	1278	1329	1942	1085	1662	1748	192
14	Cat11	535	586	543	470	479	502	582	537	412	514	462	48
15	Cat12	13	19	15	18	11	19	18	10	16	11	15	1
16													

Values for **Sex** are in the **sample_values** sheet, **row 2**.

We are telling Excel, “When you calculate the average, only include values from females.”

What if I want to know the average Cat1 value for all females that were exposed?

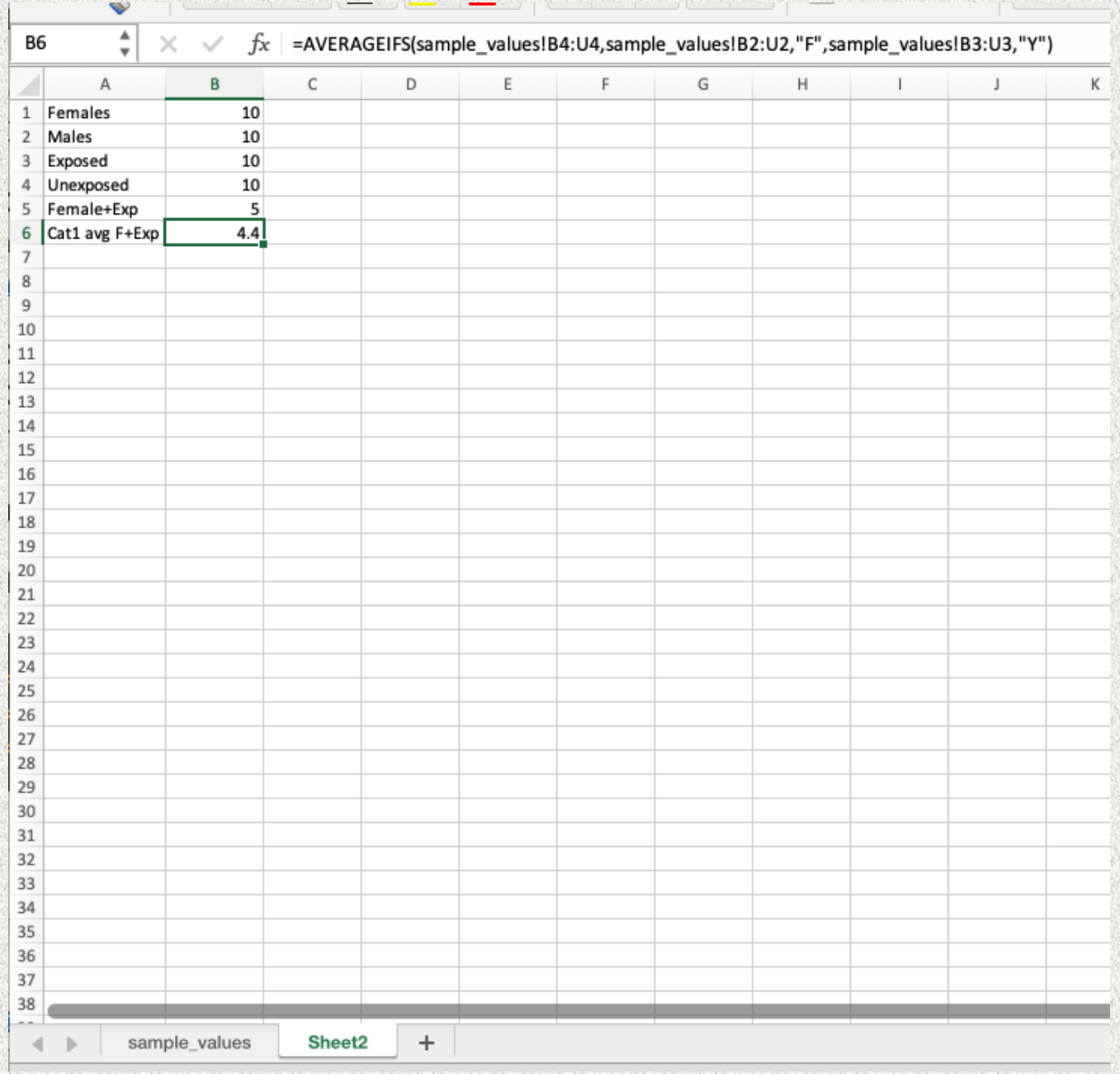


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Sample_ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14
2	Sex	F	F	M	F	M	M	M	F	F	M	F	M	M	F
3	Status	N	Y	Y	N	N	N	Y	Y	N	Y	Y	N	Y	N
4	Cat1	3	7	9	3	1	4	7	10	6	9	1	8	8	1
5	Cat2	0.39980403	0.131167	0.83797435	0.99121998	0.37915433	0.4932649	0.31730995	0.78559809	0.29231847	0.22800014	0.12970948	0.20895121	0.09570753	0.57328551
6	Cat3	0.01765164	0.74075219	0.4243295	0.84309817	0.41803328	0.36549674	0.08057165	0.73402944	0.9952279	0.6285995	0.66614321	0.15886006	0.29619534	0.522676
7	Cat4	76	75	65	53	89	66	88	78	98	67	75	96	90	65
8	Cat5	3	10	3	1	7	5	5	3	3	3	8	6	7	5
9	Cat6	0.27817695	0.07110137	0.37747939	0.13183529	0.31935773	0.1925463	0.63984189	0.64802318	0.26027957	0.93419475	0.52918677	0.2509355	0.39102529	0.77463581
10	Cat7	0.47111146	0.89977883	0.42504482	0.09046157	0.43938449	0.85181114	0.7906725	0.03804672	0.98191601	0.06929245	0.17350784	0.00365075	0.62290958	0.10161224
11	Cat8	59	79	81	65	54	91	50	92	96	51	80	63	78	70
12	Cat9	0.15067873	0.69674031	0.2607653	0.54603649	0.94019016	0.63397075	0.31429126	0.1763019	0.87520766	0.0101267	0.8489138	0.81448124	0.14730753	0.47184386
13	Cat10	1646	1436	1486	1218	1151	1278	1329	1942	1085	1662	1748	1922	1061	1850
14	Cat11	535	586	543	470	479	502	582	537	412	514	462	487	543	405
15	Cat12	13	19	15	18	11	19	18	10	16	11	15	19	20	18
16															

Exposure status values are in the **sample_values** sheet, **row 3**.

We are telling Excel, "When you calculate the average, only include values for exposed individuals."

We have an answer!



TRUE or FALSE:

Understanding the effect of exposure on average Cat1 values for females could involve an AVERAGEIFS calculation for *unexposed (control) females*.

Let's see…

- Start out the same as before:
  - We still want average Cat1 values
  - We still want data for females

=AVERAGEIFS(sample_values!B4:U4, sample_values!B2:U2, "F",

SUM

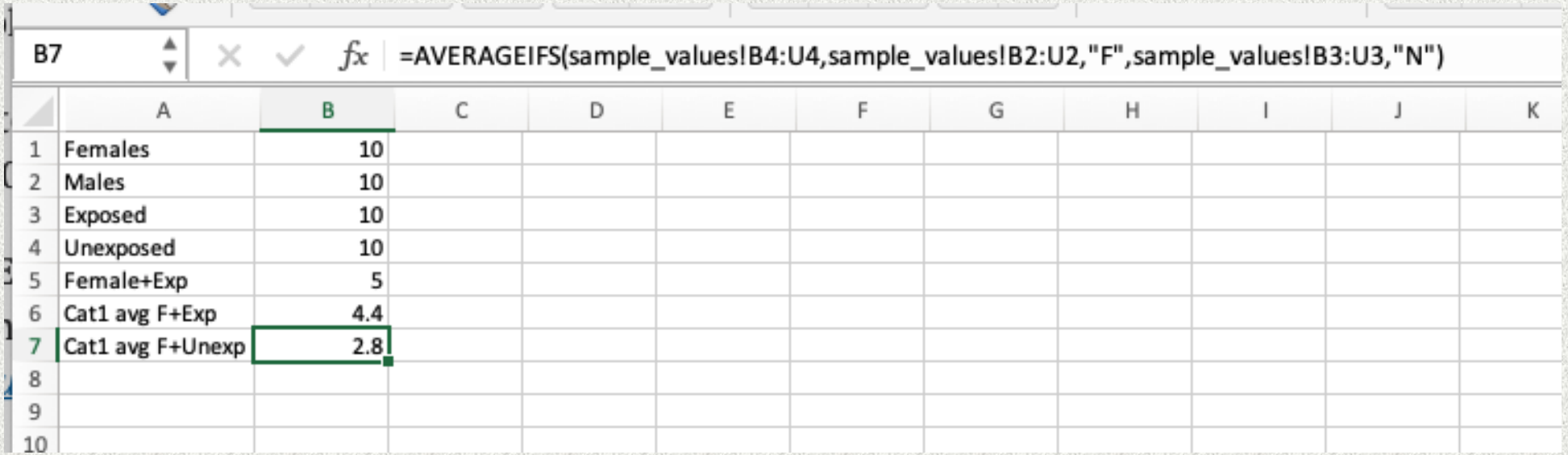
=averageifs(sample_values!B4:U4,sample_values!B2:U2,"F",sample_values!B3:U3,"N"

AVERAGEIFS(average_range, criteria_range1, criteria1, [criteria_range2, criteria2], [criteria_range3, criteria3], ...)

	A	B	C											
1	Sample_ID	1	2	3	4	5	6	7	8	9	10	11	12	
2	Sex	F	F	M	F	M	M	M	F	F	M	F	M	M
3	Status	N	Y	Y	N	N	N	Y	Y	N	Y	Y	N	Y
4	Cat1	3	7	9	3	1	4	7	10	6	9	1	8	
5	Cat2	0.39980403	0.131167	0.83797435	0.99121998	0.37915433	0.4932649	0.31730995	0.78559809	0.29231847	0.22800014	0.12970948	0.20895121	0.095
6	Cat3	0.01765164	0.74075219	0.4243295	0.84309817	0.41803328	0.36549674	0.08057165	0.73402944	0.9952279	0.6285995	0.66614321	0.15886006	0.296
7	Cat4	=averageifs(sample_values!B4:U4,sample_values!B2:U2,"F",sample_values!B3:U3,"N"							78	98	67	75	96	
8	Cat5	3	10	3	1	7	5	5	3	3	3	8	6	
9	Cat6	0.27817695	0.07110137	0.37747939	0.13183529	0.31935773	0.1925463	0.63984189	0.64802318	0.26027957	0.93419475	0.52918677	0.2509355	0.391
10	Cat7	0.47111146	0.89977883	0.42504482	0.09046157	0.43938449	0.85181114	0.7906725	0.03804672	0.98191601	0.06929245	0.17350784	0.00365075	0.622
11	Cat8	59	79	81	65	54	91	50	92	96	51	80	63	
12	Cat9	0.15067873	0.69674031	0.2607653	0.54603649	0.94019016	0.63397075	0.31429126	0.1763019	0.87520766	0.0101267	0.8489138	0.81448124	0.147
13	Cat10	1646	1436	1486	1218	1151	1278	1329	1942	1085	1662	1748	1922	
14	Cat11	535	586	543	470	479	502	582	537	412	514	462	487	
15	Cat12	13	19	15	18	11	19	18	10	16	11	15	19	

Let's see…

- Now we have calculated the average value of Cat1 for females that were NOT exposed. sample_values!B3:U3, "N")

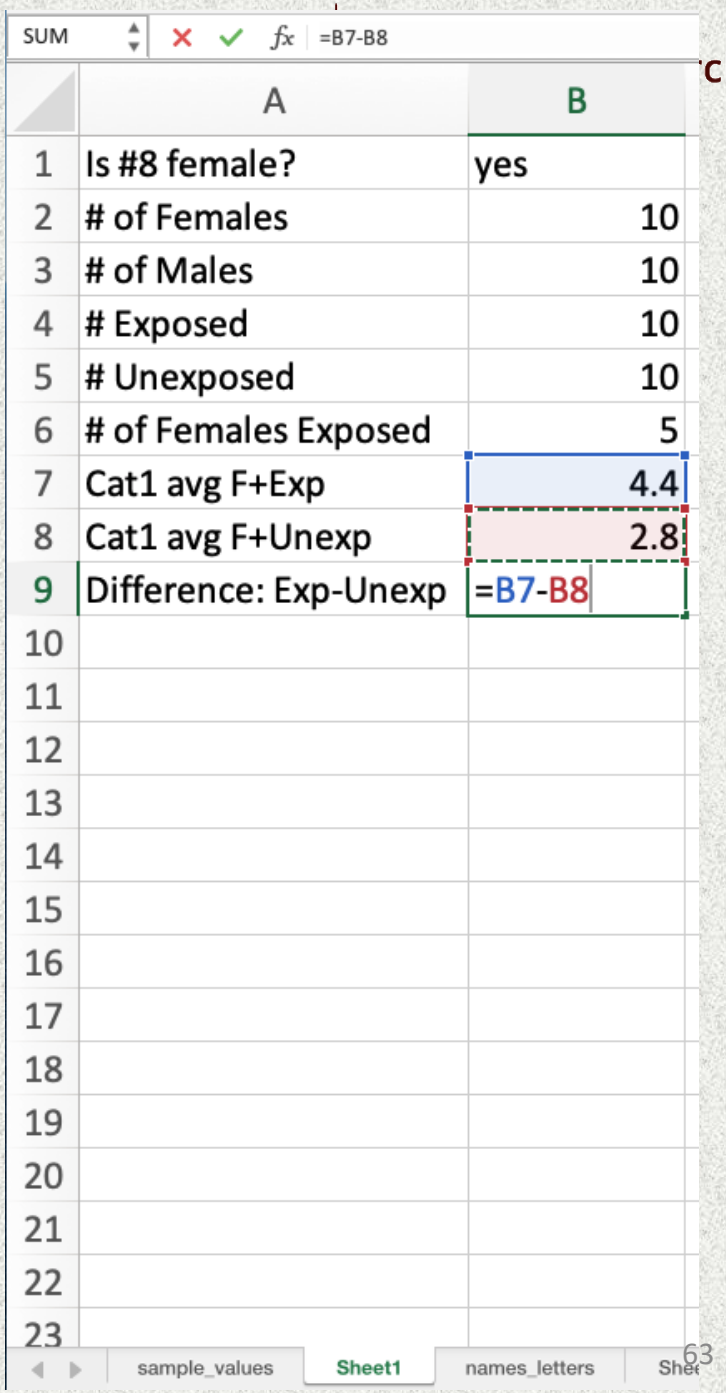


	A	B
1	Females	10
2	Males	10
3	Exposed	10
4	Unexposed	10
5	Female+Exp	5
6	Cat1 avg F+Exp	4.4
7	Cat1 avg F+Unexp	2.8
8		
9		
10		

Difference between exposed and unexposed (control)

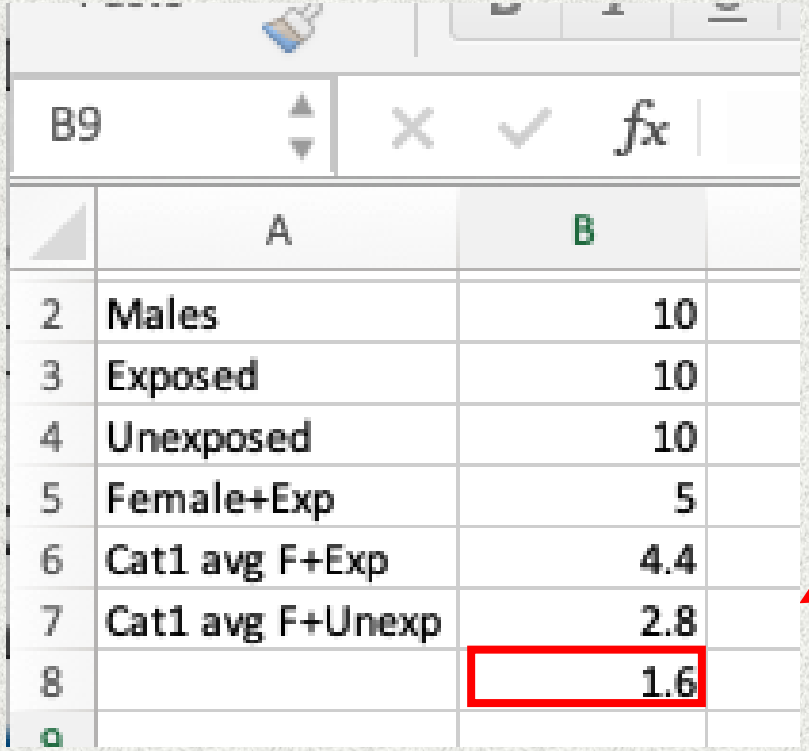
- Calculate the difference by subtraction.

- This isn't a formula, so no parentheses



TRUE!

- If we want to see how exposure affects the average Cat1 values for females, we could start by doing an AVERAGEIFS calculation for unexposed females.



Cat1 average value is *higher* in exposed females versus unexposed females.

How to find a specific value

- **LOOKUP** function
  - When you want to search a specific row or column for some value that corresponds to another row or column

- **=LOOKUP(**
  - the specific thing you already know [“lookup_value”]**,**
  - the row/column that contains the information you already know [“lookup_vector”]**,**
  - the row/column that contains the value you’re interested in [“results_vector”]**)**

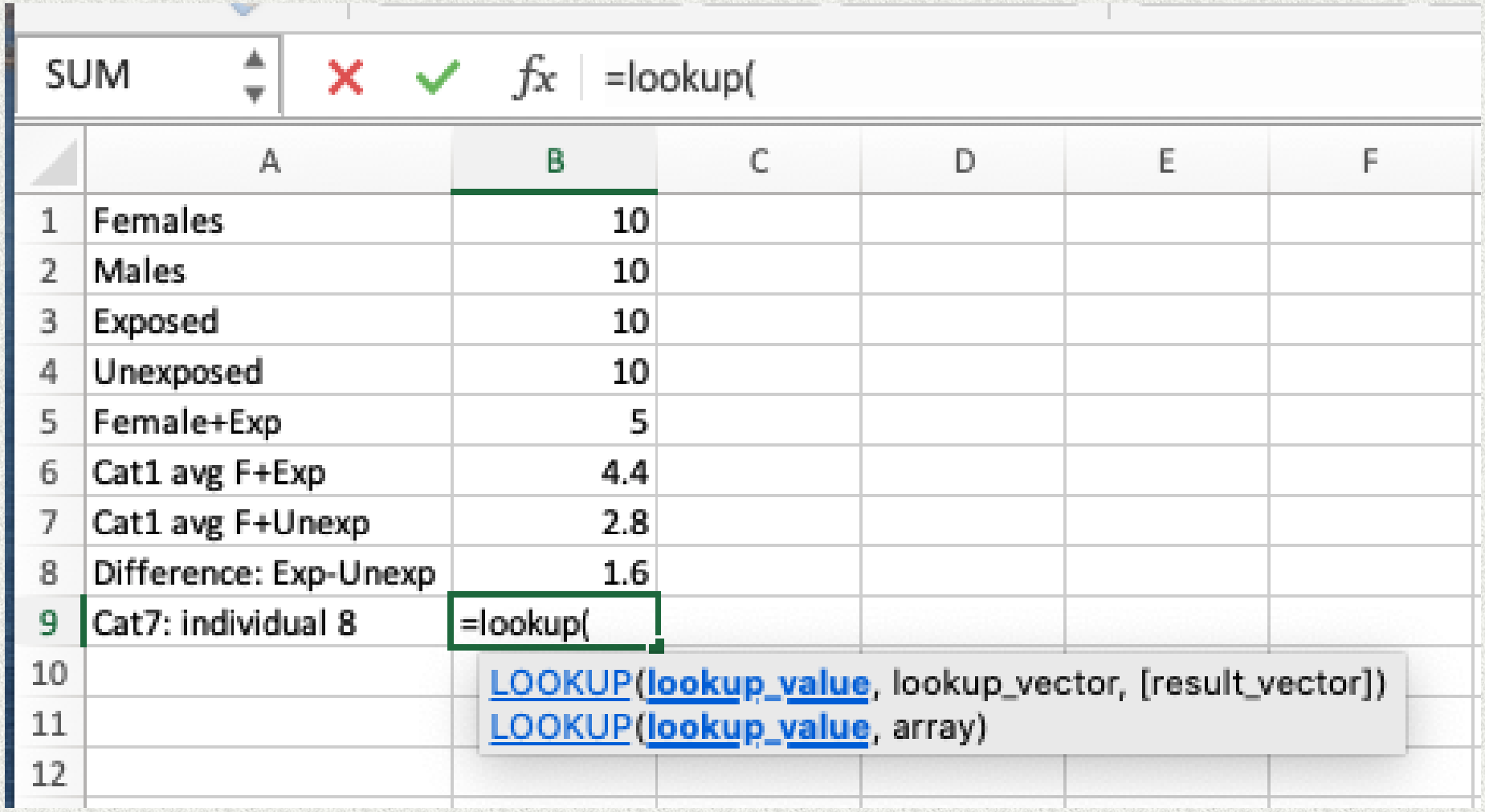
What is the Cat7 value for individual 8?

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Sample_ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2	Sex	F	F	M	F	M	M	M	F	F	M	F	M	M	F	M	F	M	F	M	F
3	Status	N	Y	Y	N	N	N	Y	Y	N	Y	Y	N	Y	N	N	Y	Y	Y	N	N
4	Cat1	3	7	9	3	1	4	7	10	6	9	1	8	8	1	1	2	5	2	8	1
5	Cat2	0.39980403	0.131167	0.83797435	0.99121998	0.37915433	0.4932649	0.31730995	0.78559809	0.29231847	0.22800014	0.12970948	0.20895121	0.09570753	0.57328551	0.74719656	0.97775537	0.90570917	0.23738616	0.63079488	0.03630628
6	Cat3	0.01765164	0.74075219	0.4243295	0.84309817	0.41803328	0.36549674	0.08057165	0.73402944	0.9952279	0.6285995	0.66614321	0.15886006	0.29619534	0.522676	0.49873951	0.499717	0.57727393	0.35131784	0.24743548	0.36314088
7	Cat4	76	75	65	53	89	66	88	78	98	67	75	96	90	65	70	89	73	84	85	71
8	Cat5	3	10	3	1	7	5	5	3	3	3	8	6	7	5	8	7	4	2	9	9
9	Cat6	0.27817695	0.07110137	0.37747939	0.13183529	0.31935773	0.1925463	0.63984189	0.64802318	0.26027957	0.93419475	0.52918677	0.2509355	0.39102529	0.77463581	0.59241452	0.48270423	0.83625265	0.41941355	0.83425559	0.93651799
10	Cat7	0.47111146	0.89977883	0.42504482	0.09046157	0.43938449	0.85181114	0.7906725	0.03804672	0.98191601	0.06929245	0.17350784	0.00365075	0.62290958	0.10161224	0.129711	0.87430617	0.63898328	0.72340885	0.87721221	0.74542847
11	Cat8	59	79	81	65	54	91	50	92	96	51	80	63	78	70	91	60	80	59	81	92
12	Cat9	0.15067873	0.69674031	0.2607653	0.54603649	0.94019016	0.63397075	0.31429126	0.1763019	0.87520766	0.0101267	0.8489138	0.81448124	0.14730753	0.47184386	0.97195415	0.10610593	0.55886697	0.34184183	0.38806705	0.24976963
13	Cat10	1646	1436	1486	1218	1151	1278	1329	1942	1085	1662	1748	1922	1061	1850	1160	1166	1049	1419	1275	1207
14	Cat11	535	586	543	470	479	502	582	537	412	514	462	487	543	405	504	478	466	419	467	591
15	Cat12	13	19	15	18	11	19	18	10	16	11	15	19	20	18	13	17	16	14	14	14
16																					

=LOOKUP(the specific thing you already know [“**lookup_value**”]**,** the row/column that contains the information you already know [“**lookup_vector**”]**,** the row/column that contains some value that information corresponds to [“**results_vector**”]**)**

- The **lookup_value** here is ID number 8
- The **lookup_vector** here is the row with all the ID numbers
  - **IMPORTANT: Lookup_vector must be in ascending order.** (Ours is already ordered numerically, left to right.)
- The **results_vector** here is the row with all the Cat7 measurement values.

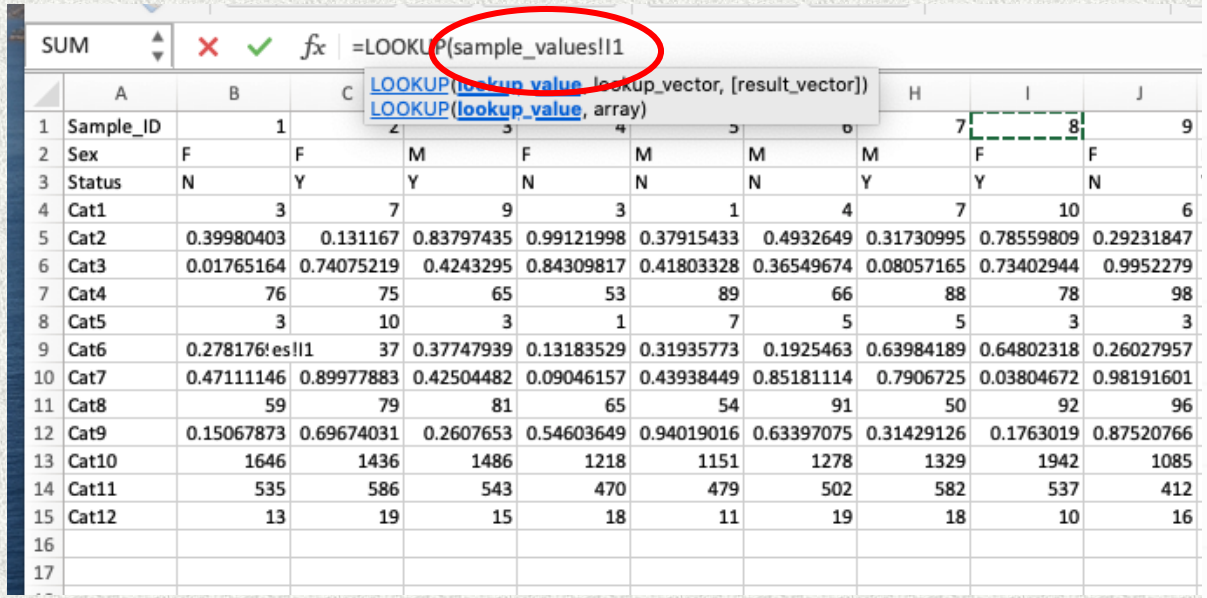
What is the Cat7 value for individual 8?



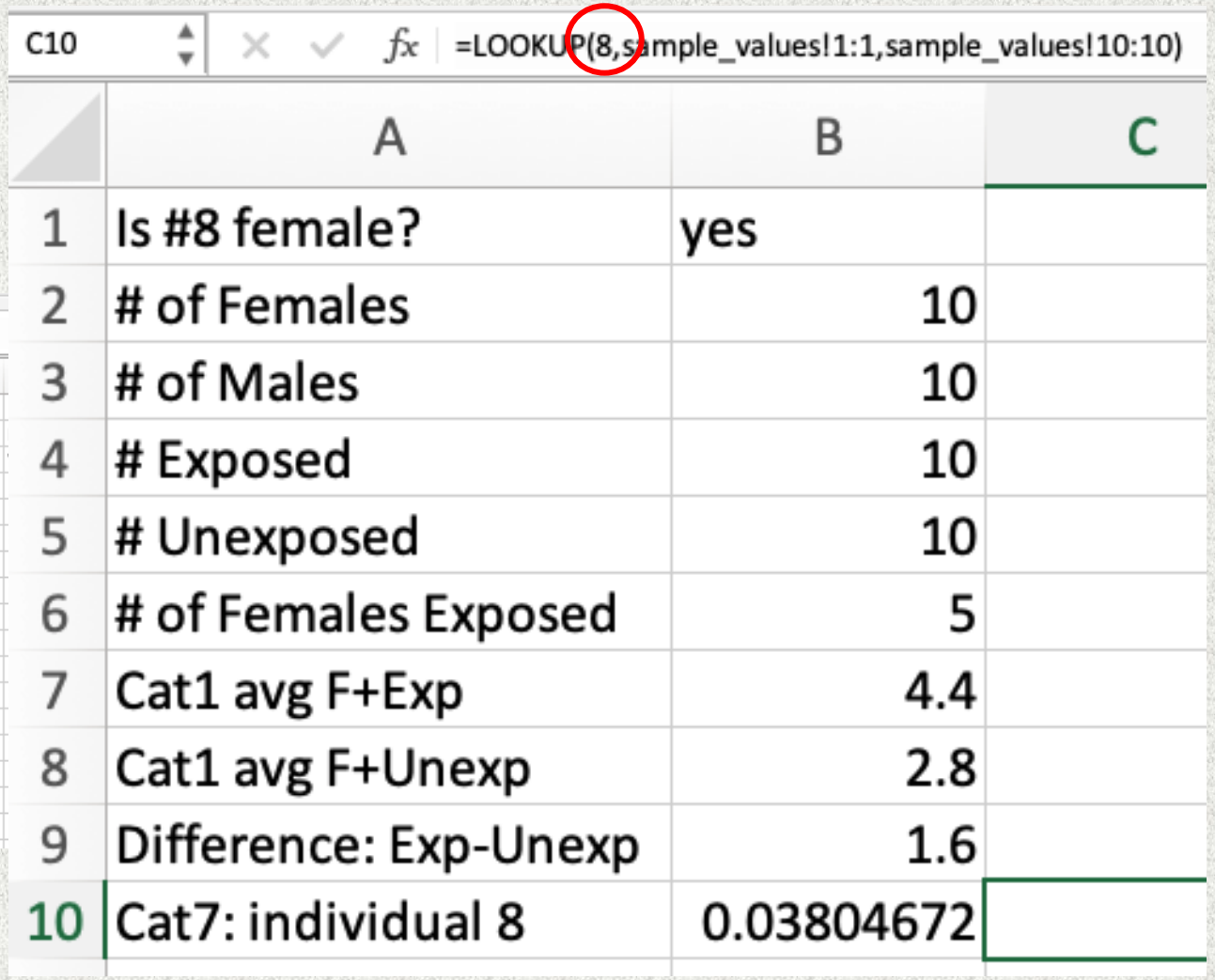
What is the Cat7 value for individual 8?
Lookup value: individual 8

Lookup value = 8

I could also select cell containing the sample ID for individual 8.



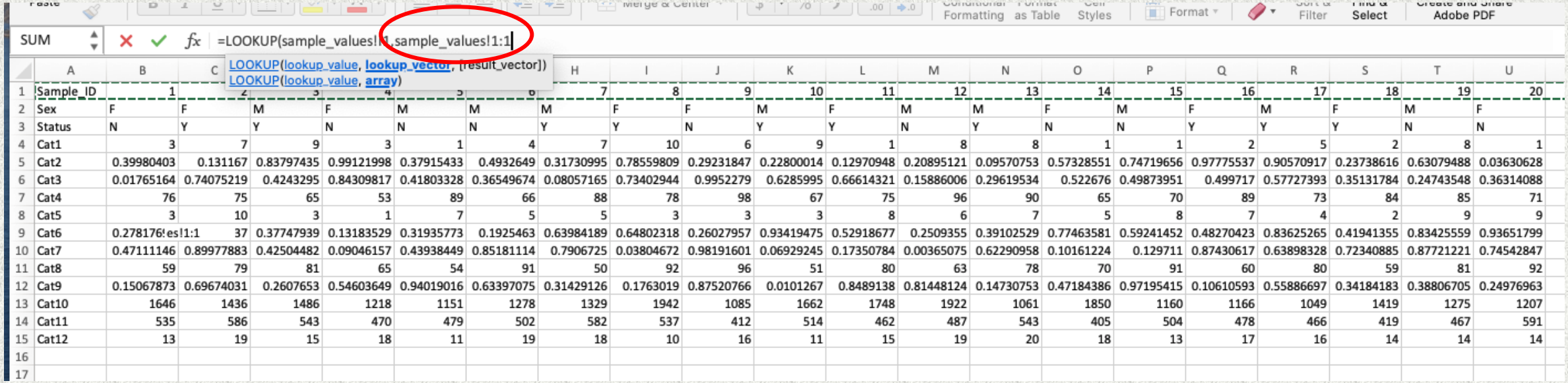
	A	B	C	D	E	F	G	H	I	J
1	Sample_ID	1	2	3	4	5	6	7	8	9
2	Sex	F	F	M	F	M	M	M	F	F
3	Status	N	Y	Y	N	N	N	Y	Y	N
4	Cat1	3	7	9	3	1	4	7	10	6
5	Cat2	0.39980403	0.131167	0.83797435	0.99121998	0.37915433	0.4932649	0.31730995	0.78559809	0.29231847
6	Cat3	0.01765164	0.74075219	0.4243295	0.84309817	0.41803328	0.36549674	0.08057165	0.73402944	0.9952279
7	Cat4	76	75	65	53	89	66	88	78	98
8	Cat5	3	10	3	1	7	5	5	3	3
9	Cat6	0.278176	es!I1	37	0.37747939	0.13183529	0.31935773	0.1925463	0.63984189	0.64802318	0.26027957
10	Cat7	0.47111146	0.89977883	0.42504482	0.09046157	0.43938449	0.85181114	0.7906725	0.03804672	0.98191601
11	Cat8	59	79	81	65	54	91	50	92	96
12	Cat9	0.15067873	0.69674031	0.2607653	0.54603649	0.94019016	0.63397075	0.31429126	0.1763019	0.87520766
13	Cat10	1646	1436	1486	1218	1151	1278	1329	1942	1085
14	Cat11	535	586	543	470	479	502	582	537	412
15	Cat12	13	19	15	18	11	19	18	10	16
16										
17										



	A	B	C
1	Is #8 female?	yes	
2	# of Females	10	
3	# of Males	10	
4	# Exposed	10	
5	# Unexposed	10	
6	# of Females Exposed	5	
7	Cat1 avg F+Exp	4.4	
8	Cat1 avg F+Unexp	2.8	
9	Difference: Exp-Unexp	1.6	
10	Cat7: individual 8	0.03804672	

What is the Cat7 value for individual 8?
Lookup vector: all the sample IDs

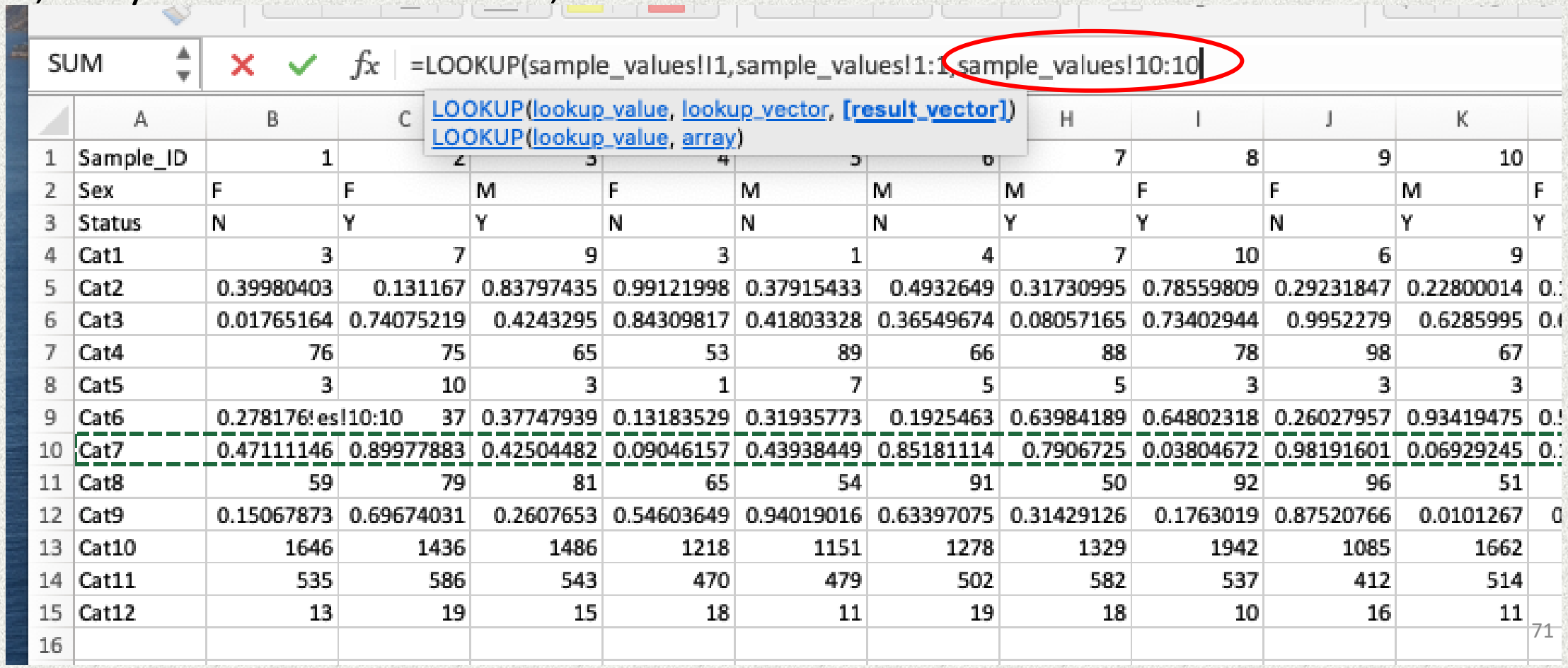
I select the entire Sample ID row for sheet sample_values by clicking on the row number 1. (If my data were in columns, I could click the column letter to select the entire column.) This also tells Excel what sheet to look at.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Sample_ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2	Sex	F	F	M	F	M	M	M	F	F	M	F	M	M	F	M	F	M	F	M	F
3	Status	N	Y	Y	N	N	N	Y	Y	N	Y	Y	N	Y	N	N	Y	Y	Y	N	N
4	Cat1	3	7	9	3	1	4	7	10	6	9	1	8	8	1	1	2	5	2	8	1
5	Cat2	0.39980403	0.131167	0.83797435	0.99121998	0.37915433	0.4932649	0.31730995	0.78559809	0.29231847	0.22800014	0.12970948	0.20895121	0.09570753	0.57328551	0.74719656	0.97775537	0.90570917	0.23738616	0.63079488	0.03630628
6	Cat3	0.01765164	0.74075219	0.4243295	0.84309817	0.41803328	0.36549674	0.08057165	0.73402944	0.9952279	0.6285995	0.66614321	0.15886006	0.29619534	0.522676	0.49873951	0.499717	0.57727393	0.35131784	0.24743548	0.36314088
7	Cat4	76	75	65	53	89	66	88	78	98	67	75	96	90	65	70	89	73	84	85	71
8	Cat5	3	10	3	1	7	5	5	3	3	3	8	6	7	5	8	7	4	2	9	9
9	Cat6	0.278176	es!1:1	37	0.37747939	0.13183529	0.31935773	0.1925463	0.63984189	0.64802318	0.26027957	0.93419475	0.52918677	0.2509355	0.39102529	0.77463581	0.59241452	0.48270423	0.83625265	0.41941355	0.83425559	0.93651799
10	Cat7	0.47111146	0.89977883	0.42504482	0.09046157	0.43938449	0.85181114	0.7906725	0.03804672	0.98191601	0.06929245	0.17350784	0.00365075	0.62290958	0.10161224	0.129711	0.87430617	0.63898328	0.72340885	0.87721221	0.74542847
11	Cat8	59	79	81	65	54	91	50	92	96	51	80	63	78	70	91	60	80	59	81	92
12	Cat9	0.15067873	0.69674031	0.2607653	0.54603649	0.94019016	0.63397075	0.31429126	0.1763019	0.87520766	0.0101267	0.8489138	0.81448124	0.14730753	0.47184386	0.97195415	0.10610593	0.55886697	0.34184183	0.38806705	0.24976963
13	Cat10	1646	1436	1486	1218	1151	1278	1329	1942	1085	1662	1748	1922	1061	1850	1160	1166	1049	1419	1275	1207
14	Cat11	535	586	543	470	479	502	582	537	412	514	462	487	543	405	504	478	466	419	467	591
15	Cat12	13	19	15	18	11	19	18	10	16	11	15	19	20	18	13	17	16	14	14	14
16																					
17																					

What is the Cat7 value for individual 8?
Result vector: all the Cat7 values

I select the entire Cat7 row by clicking on the row number 10 in the sheet sample_values.
(Again, if my data were in columns, I could click the column letter to select the entire column.)



	A	B	C	H	I	J	K
1	Sample_ID	1	2	7	8	9	10
2	Sex	F	F	M	F	F	M
3	Status	N	Y	Y	N	Y	Y
4	Cat1	3	7	9	3	1	4
5	Cat2	0.39980403	0.131167	0.83797435	0.99121998	0.37915433	0.4932649
6	Cat3	0.01765164	0.74075219	0.4243295	0.84309817	0.41803328	0.36549674
7	Cat4	76	75	65	53	89	66
8	Cat5	3	10	3	1	7	5
9	Cat6	0.278176	es!10:10	37	0.37747939	0.13183529	0.31935773
10	Cat7	0.47111146	0.89977883	0.42504482	0.09046157	0.43938449	0.85181114
11	Cat8	59	79	81	65	54	91
12	Cat9	0.15067873	0.69674031	0.2607653	0.54603649	0.94019016	0.63397075
13	Cat10	1646	1436	1486	1218	1151	1278
14	Cat11	535	586	543	470	479	502
15	Cat12	13	19	15	18	11	19
16							

We got the correct answer!

	A	B
1	Is #8 female?	yes
2	# of Females	10
3	# of Males	10
4	# Exposed	10
5	# Unexposed	10
6	# of Females Exposed	5
7	Cat1 avg F+Exp	4.4
8	Cat1 avg F+Unexp	2.8
9	Difference: Exp-Unexp	1.6
10	Cat7: individual 8	0.03804672
11		
12		
13		
14		
15		
16		
17		
18		
19		
20		
21		
22		
23		

sample_values | Sheet1 | names_letters | She

I10 fx 0.0380467181436256

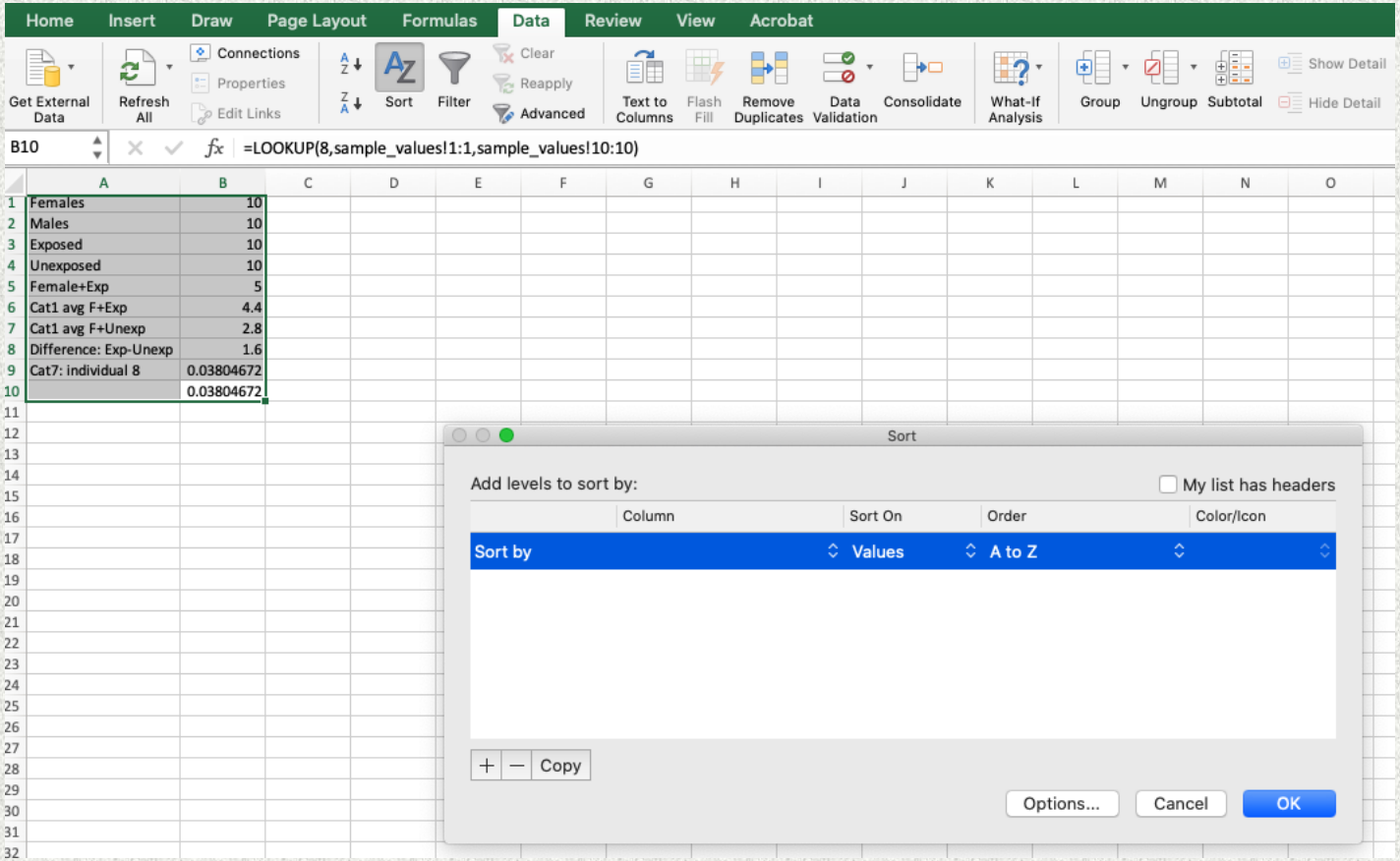
	A	B	C	D	E	F	G	H	I	J	K	L
1	Sample_ID	1	2	3	4	5	6	7	8	9	10	
2	Sex	F	F	M	F	M	M	M	F	F	M	F
3	Status	N	Y	Y	N	N	N	Y	Y	N	Y	Y
4	Cat1	3	7	9	3	1	4	7	10	6	9	
5	Cat2	0.39980403	0.131167	0.83797435	0.99121998	0.37915433	0.4932649	0.31730995	0.78559809	0.29231847	0.22800014	0.1297
6	Cat3	0.01765164	0.74075219	0.4243295	0.84309817	0.41803328	0.36549674	0.08057165	0.73402944	0.9952279	0.6285995	0.6661
7	Cat4	76	75	65	53	89	66	88	78	98	67	
8	Cat5	3	10	3	1	7	5	5	3	3	3	
9	Cat6	0.27817695	0.07110137	0.37747939	0.13183529	0.31935773	0.1925463	0.63984189	0.64802318	0.26027957	0.93419475	0.5291
10	Cat7	0.47111146	0.89977883	0.42504482	0.09046157	0.43938449	0.85181114	0.7906725	0.03804672	0.98191601	0.06929245	0.1735
11	Cat8	59	79	81	65	54	91	50	92	96	51	
12	Cat9	0.15067873	0.69674031	0.2607653	0.54603649	0.94019016	0.63397075	0.31429126	0.1763019	0.87520766	0.0101267	0.848
13	Cat10	1646	1436	1486	1218	1151	1278	1329	1942	1085	1662	
14	Cat11	535	586	543	470	479	502	582	537	412	514	
15	Cat12	13	19	15	18	11	19	18	10	16	11	
16												
17												
18												
19												
20												
21												

When to use LOOKUP

- LOOKUP can be very useful when you have a large dataset to search.
- LOOKUP is used **when the value you want to match is in the first row/column**.
- If your lookup_vector isn't in **ascending order**, it is possible you will get a result... But it may not be the correct result!

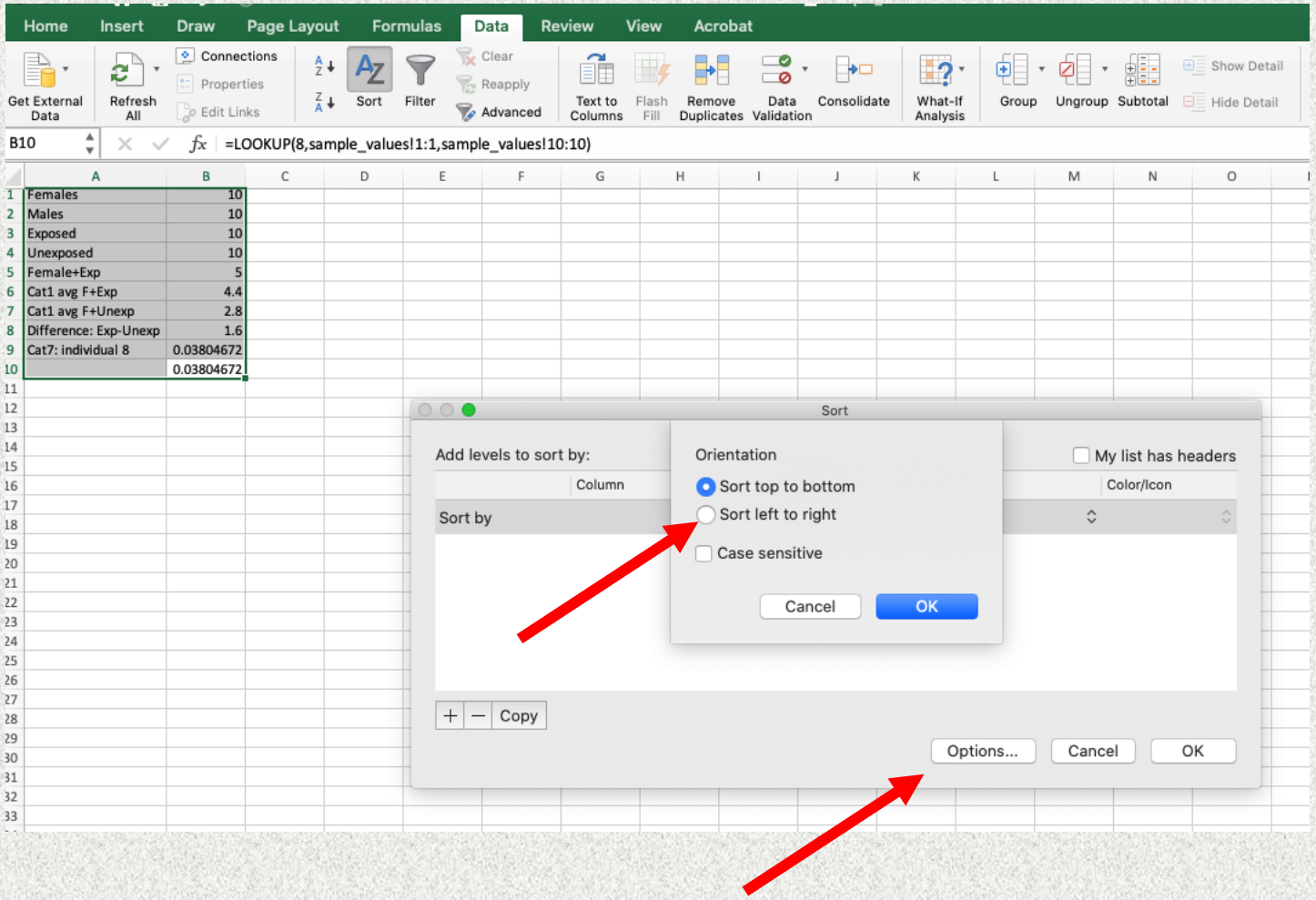
Ordering data: columns

- Make sure your lookup_vector data is in ascending order by SORTING the lookup_vector row/column before you try to use LOOKUP.
- Data → Sort → choose which column you want to sort, and how you want it sorted (e.g. ascending order [A to Z], descending order [Z to A], etc.)



Ordering data: rows

- If you need to sort a row instead of a column, click "Options…" and select "Sort left to right"



VLOOKUP

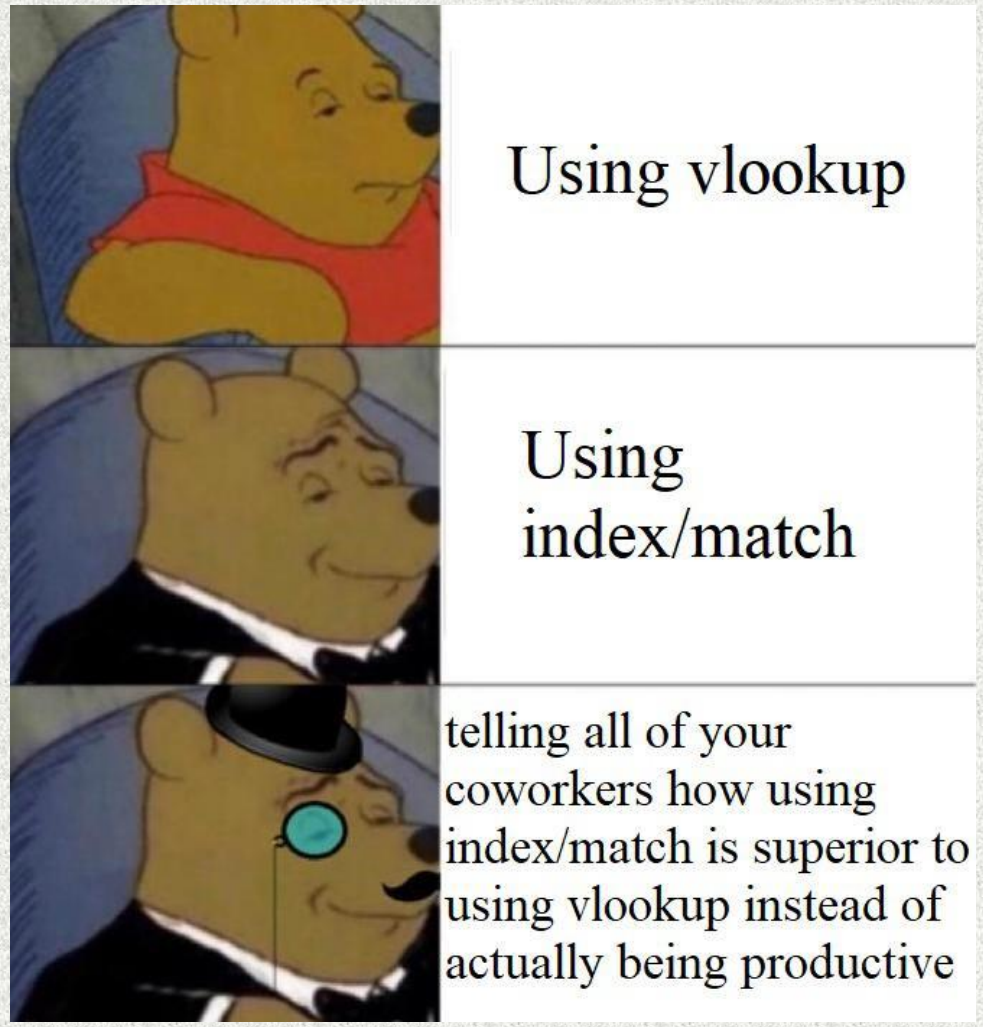
- =VLOOKUP(lookup_value,table_array,col_index_num,[range_lookup])

- =VLOOKUP(
  - Lookup_value, What you want to look up
  - Table_array, Where you want to look for it
  - Col_index_num, The column number in the range containing the value to return
  - Range_lookup) Return an Approximate (TRUE) or Exact (FALSE) match

VLOOKUP

- VLOOKUP doesn't require anything be sorted by ascending order!
- *BUT, the value you look up needs to be to the left of the return value you want to find.*

- **XLOOKUP** can look in any direction, but it's not available on all versions of Excel.
- **INDEX** and **MATCH**, used independently or nested, may be superior to VLOOKUP for some circumstances.
  - **https://exceljet.net/index-and-match**



VLOOKUP: What is the Cat7 value for individual 8?

- Repeating the search we did before, but this time pretending nothing is in order

B11 =VLOOKUP(sample_values!A10,sample_values!A1:U15,9,FALSE)

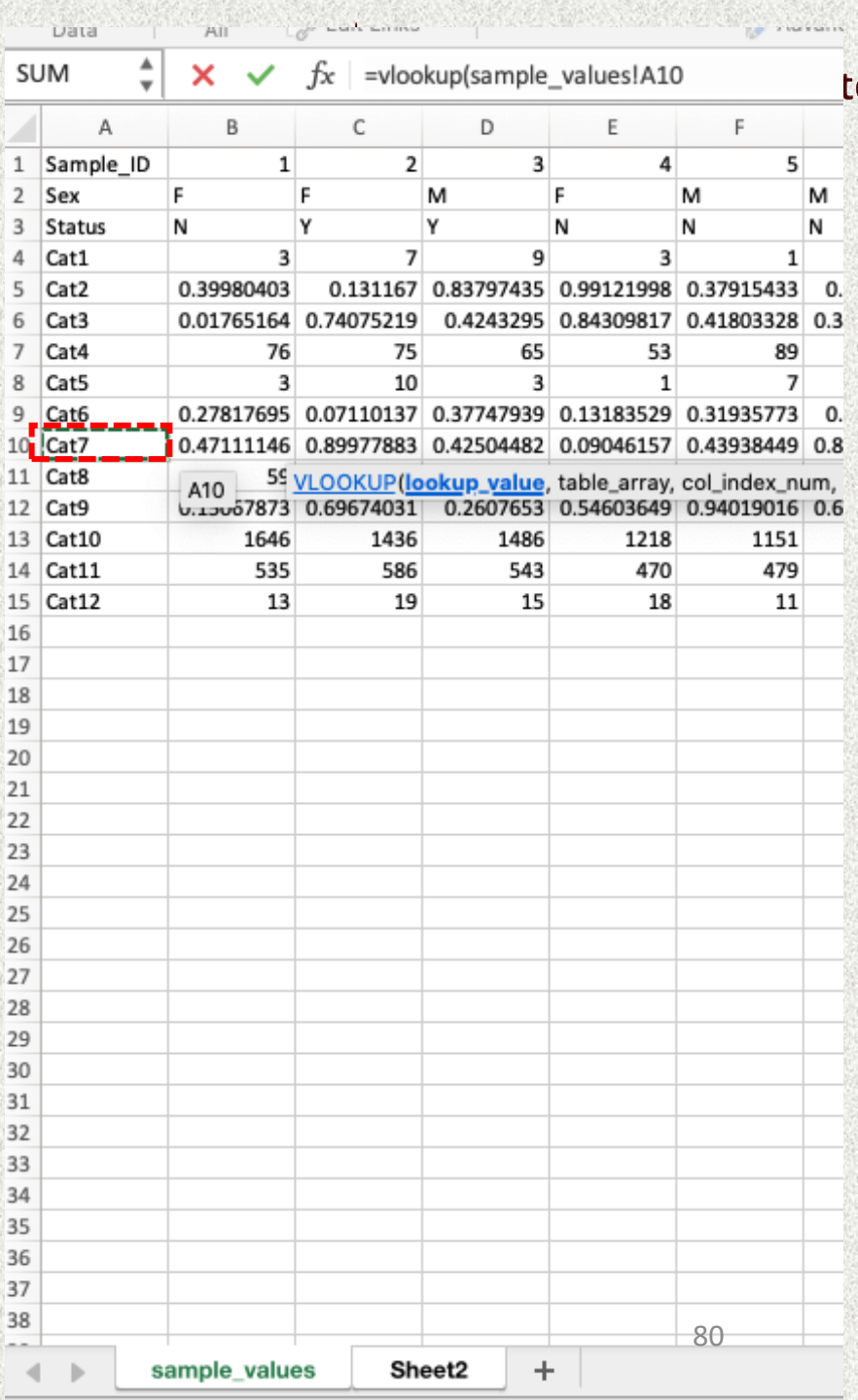
	A	B	C
1	Is #8 female?	yes	
2	# of Females	10	
3	# of Males	10	
4	# Exposed	10	
5	# Unexposed	10	
6	# of Females Exposed	5	
7	Cat1 avg F+Exp	4.4	
8	Cat1 avg F+Unexp	2.8	
9	Difference: Exp-Unexp	1.6	
10	Cat7: individual 8	0.03804672	
11	Cat7: individual 8 (v2)	0.03804672	

VLOOKUP

- We want to know the Cat7 value… so that's my lookup_value
  - Don't highlight the whole row/column, just select the category name

- =VLOOKUP(lookup_value,table_array,col_index_num,[range_lookup])

- =VLOOKUP(
  - Lookup_value, — What you want to look up
  - Table_array, — Where you want to look for it
  - Col_index_num, — The column number in the range containing the value to return
  - Range_lookup) — Return an Approximate (TRUE) or Exact (FALSE) match

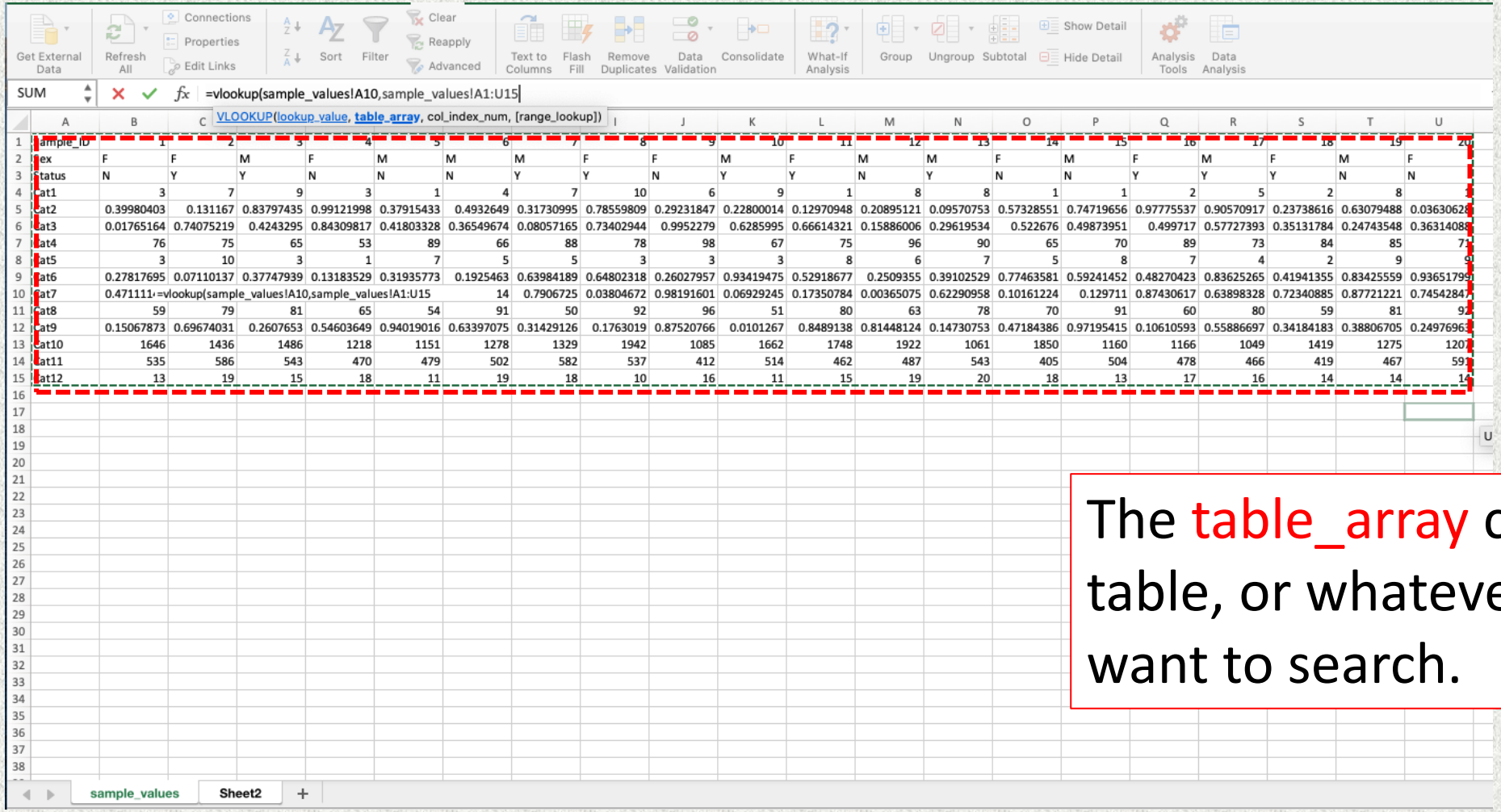


	A	B	C	D	E	F	
1	Sample_ID	1	2	3	4	5	
2	Sex	F	F	M	F	M	M
3	Status	N	Y	Y	N	N	N
4	Cat1	3	7	9	3	1	
5	Cat2	0.39980403	0.131167	0.83797435	0.99121998	0.37915433	0.
6	Cat3	0.01765164	0.74075219	0.4243295	0.84309817	0.41803328	0.3
7	Cat4	76	75	65	53	89	
8	Cat5	3	10	3	1	7	
9	Cat6	0.27817695	0.07110137	0.37747939	0.13183529	0.31935773	0.
10	Cat7	0.47111146	0.89977883	0.42504482	0.09046157	0.43938449	0.8
11	Cat8	59					
12	Cat9	0.13067873	0.69674031	0.2607653	0.54603649	0.94019016	0.6
13	Cat10	1646	1436	1486	1218	1151	
14	Cat11	535	586	543	470	479	
15	Cat12	13	19	15	18	11	

VLOOKUP

- =VLOOKUP(lookup_value,table_array,col_index_num,[range_lookup])

- =VLOOKUP(
  - Lookup_value, — What you want to look up
  - Table_array, — Where you want to look for it
  - Col_index_num, — The column number in the range containing the value to return
  - Range_lookup) — Return an Approximate (TRUE) or Exact (FALSE) match

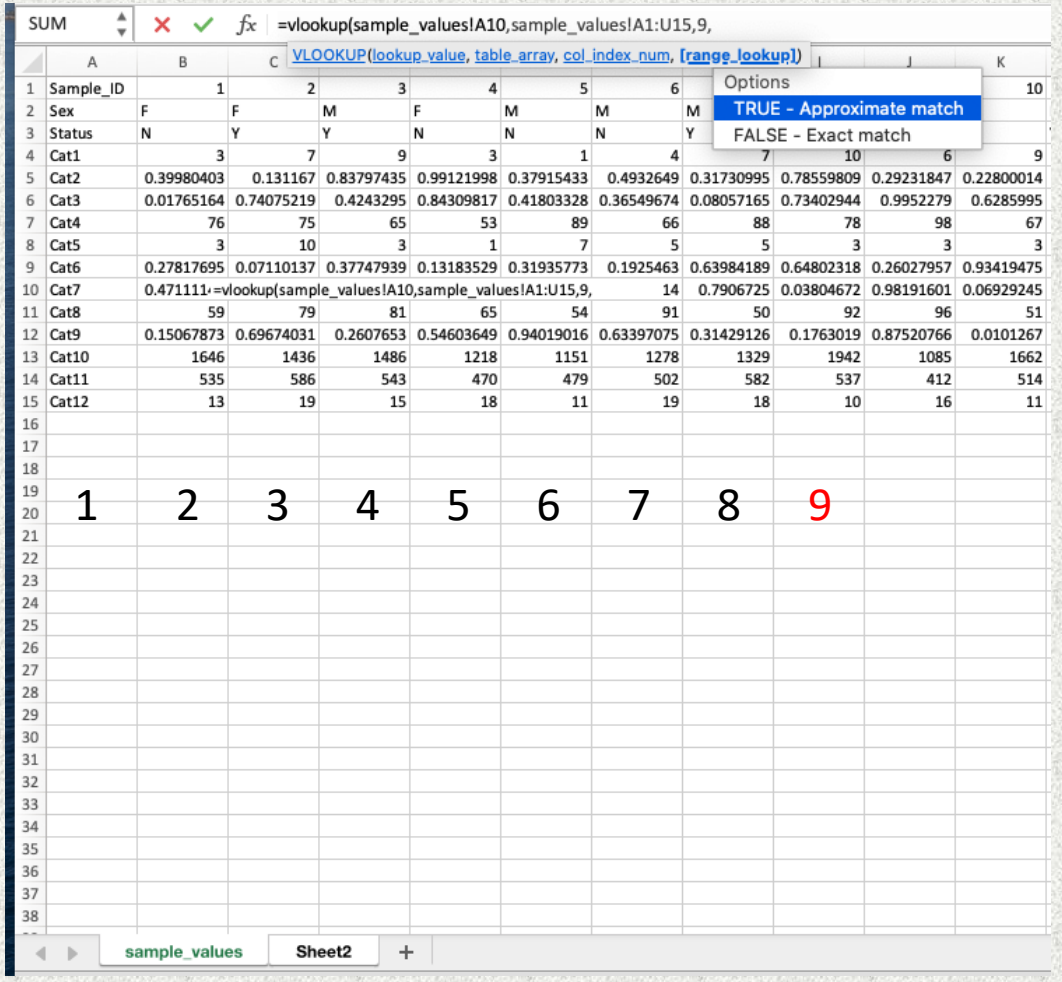


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	sample_ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2	sex	F	F	M	F	M	M	M	F	F	M	F	M	M	F	M	F	M	F	M	F
3	status	N	Y	Y	N	N	N	Y	Y	N	Y	Y	N	Y	N	N	Y	Y	Y	N	N
4	cat1	3	7	9	3	1	4	7	10	6	9	1	8	8	1	1	2	5	2	8	1
5	cat2	0.39980403	0.131167	0.83797435	0.99121998	0.37915433	0.4932649	0.31730995	0.78559809	0.29231847	0.22800014	0.12970948	0.20895121	0.09570753	0.57328551	0.74719656	0.97775537	0.90570917	0.23738616	0.63079488	0.03630628
6	cat3	0.01765164	0.74075219	0.4243295	0.84309817	0.41803328	0.36549674	0.08057165	0.73402944	0.9952279	0.6285995	0.66614321	0.15886006	0.29619534	0.522676	0.49873951	0.499717	0.57727393	0.35131784	0.24743548	0.36314088
7	cat4	76	75	65	53	89	66	88	78	98	67	75	96	90	65	70	89	73	84	85	71
8	cat5	3	10	3	1	7	5	5	3	3	3	8	6	7	5	8	7	4	2	9	9
9	cat6	0.27817695	0.07110137	0.37747939	0.13183529	0.31935773	0.1925463	0.63984189	0.64802318	0.26027957	0.93419475	0.52918677	0.2509355	0.39102529	0.77463581	0.59241452	0.48270423	0.83625265	0.41941355	0.83425559	0.93651799
10	cat7	0.471111	=vlookup(sample_values!A10,sample_values!A1:U15				14	0.7906725	0.03804672	0.98191601	0.06929245	0.17350784	0.00365075	0.62290958	0.10161224	0.129711	0.87430617	0.63898328	0.72340885	0.87721221	0.74542847
11	cat8	59	79	81	65	54	91	50	92	96	51	80	63	78	70	91	60	80	59	81	92
12	cat9	0.15067873	0.69674031	0.2607653	0.54603649	0.94019016	0.63397075	0.31429126	0.1763019	0.87520766	0.0101267	0.8489138	0.81448124	0.14730753	0.47184386	0.97195415	0.10610593	0.55886697	0.34184183	0.38806705	0.24976963
13	cat10	1646	1436	1486	1218	1151	1278	1329	1942	1085	1662	1748	1922	1061	1850	1160	1166	1049	1419	1275	1207
14	cat11	535	586	543	470	479	502	582	537	412	514	462	487	543	405	504	478	466	419	467	591
15	cat12	13	19	15	18	11	19	18	10	16	11	15	19	20	18	13	17	16	14	14	14

The table_array can be the entire table, or whatever section you want to search.

VLOOKUP

- Individual #8, the one I'm interested in, is 9 columns to the right of my lookup_value (Cat7).
- So here, 9 is my col_index_num



	A	B	C	D	E	F	G	H	I	J	K
1	Sample_ID	1	2	3	4	5	6				10
2	Sex	F	F	M	F	M	M	M			
3	Status	N	Y	Y	N	N	N	Y			
4	Cat1	3	7	9	3	1	4	7	10	6	9
5	Cat2	0.39980403	0.131167	0.83797435	0.99121998	0.37915433	0.4932649	0.31730995	0.78559809	0.29231847	0.22800014
6	Cat3	0.01765164	0.74075219	0.4243295	0.84309817	0.41803328	0.36549674	0.08057165	0.73402944	0.9952279	0.6285995
7	Cat4	76	75	65	53	89	66	88	78	98	67
8	Cat5	3	10	3	1	7	5	5	3	3	3
9	Cat6	0.27817695	0.07110137	0.37747939	0.13183529	0.31935773	0.1925463	0.63984189	0.64802318	0.26027957	0.93419475
10	Cat7	0.471111	=vlookup(sample_values!A10,sample_values!A1:U15,9,				14	0.7906725	0.03804672	0.98191601	0.06929245
11	Cat8	59	79	81	65	54	91	50	92	96	51
12	Cat9	0.15067873	0.69674031	0.2607653	0.54603649	0.94019016	0.63397075	0.31429126	0.1763019	0.87520766	0.0101267
13	Cat10	1646	1436	1486	1218	1151	1278	1329	1942	1085	1662
14	Cat11	535	586	543	470	479	502	582	537	412	514
15	Cat12	13	19	15	18	11	19	18	10	16	11

- =VLOOKUP(lookup_value,table_array,col_index_num,[range_lookup])

- =VLOOKUP(
  - Lookup_value, — What you want to look up
  - Table_array, — Where you want to look for it
  - Col_index_num, — The column number in the range containing the value to return
  - Range_lookup) — Return an Approximate (TRUE) or Exact (FALSE) match

VLOOKUP

- Now I put **FALSE** (or 0) because I want an *exact* match.

- If I'd put TRUE, or 1, I might get an *approximate* match.

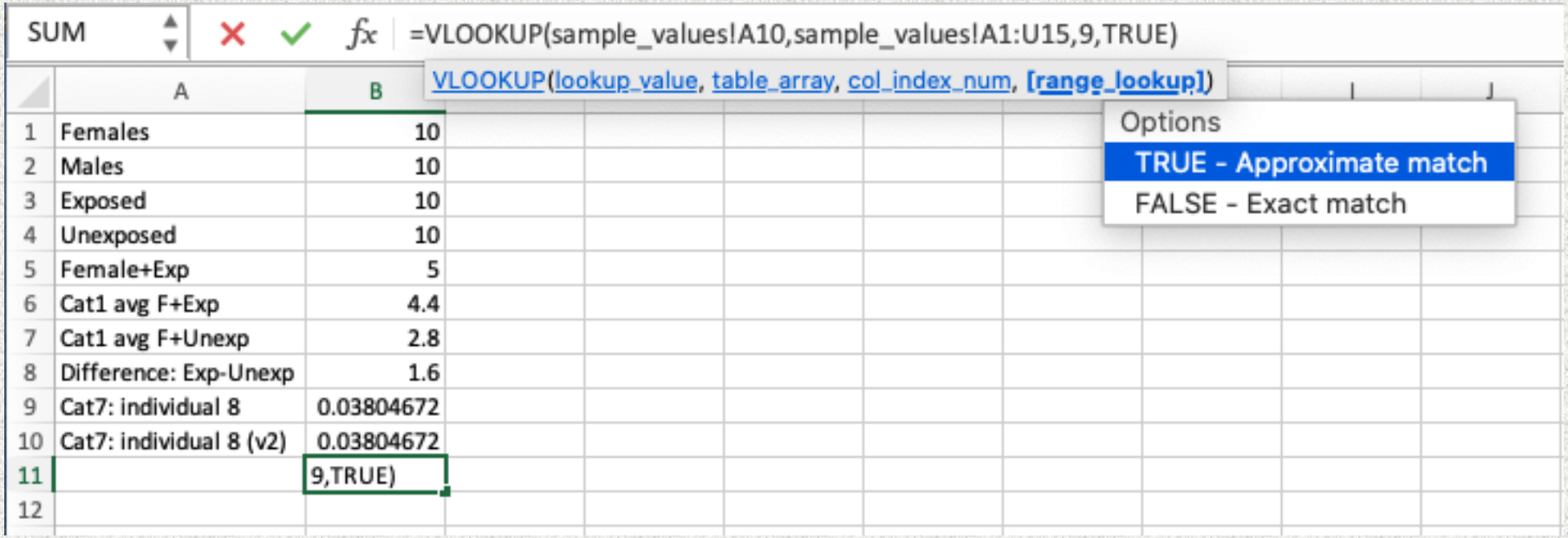
B10 | =VLOOKUP(sample_values!A10,sample_values!A1:U15,9,FALSE)

	A	B	C	D	E	F	G	H
1	Females	10						
2	Males	10						
3	Exposed	10						
4	Unexposed	10						
5	Female+Exp	5						
6	Cat1 avg F+Exp	4.4						
7	Cat1 avg F+Unexp	2.8						
8	Difference: Exp-Unexp	1.6						
9	Cat7: individual 8	0.03804672						
10	Cat7: individual 8 (v2)	0.03804672						
11								
12								
13								

- =VLOOKUP(lookup_value,table_array,col_index_num,[range_lookup])

- =VLOOKUP(
  - Lookup_value, — What you want to look up
  - Table_array, — Where you want to look for it
  - Col_index_num, — The column number in the range containing the value to return
  - Range_lookup) — Return an Approximate (TRUE) or Exact (FALSE) match

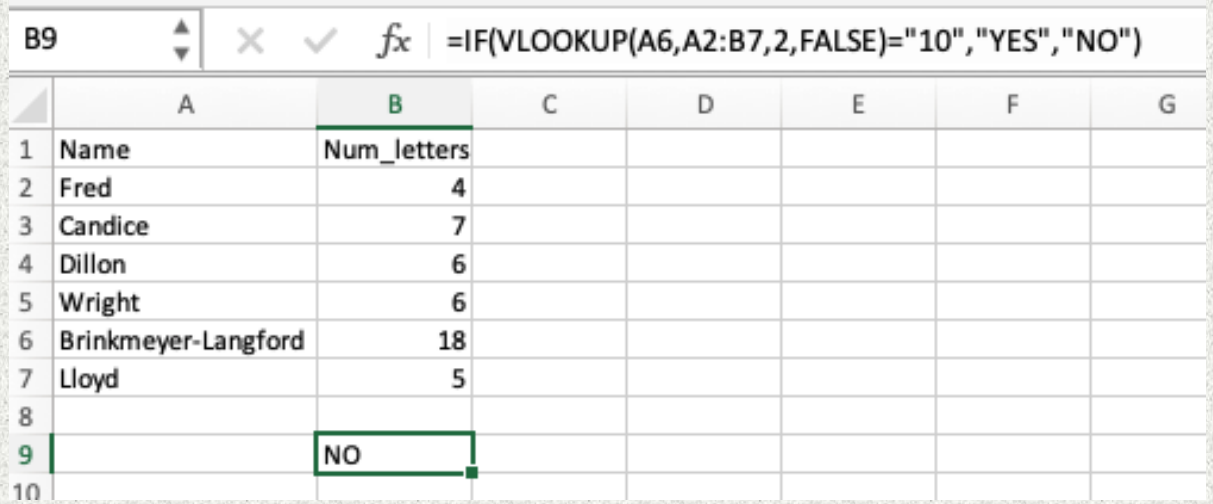
VLOOKUP



(For fun, I tried putting “TRUE” and I got the same result in this situation.)

Nested VLOOKUP

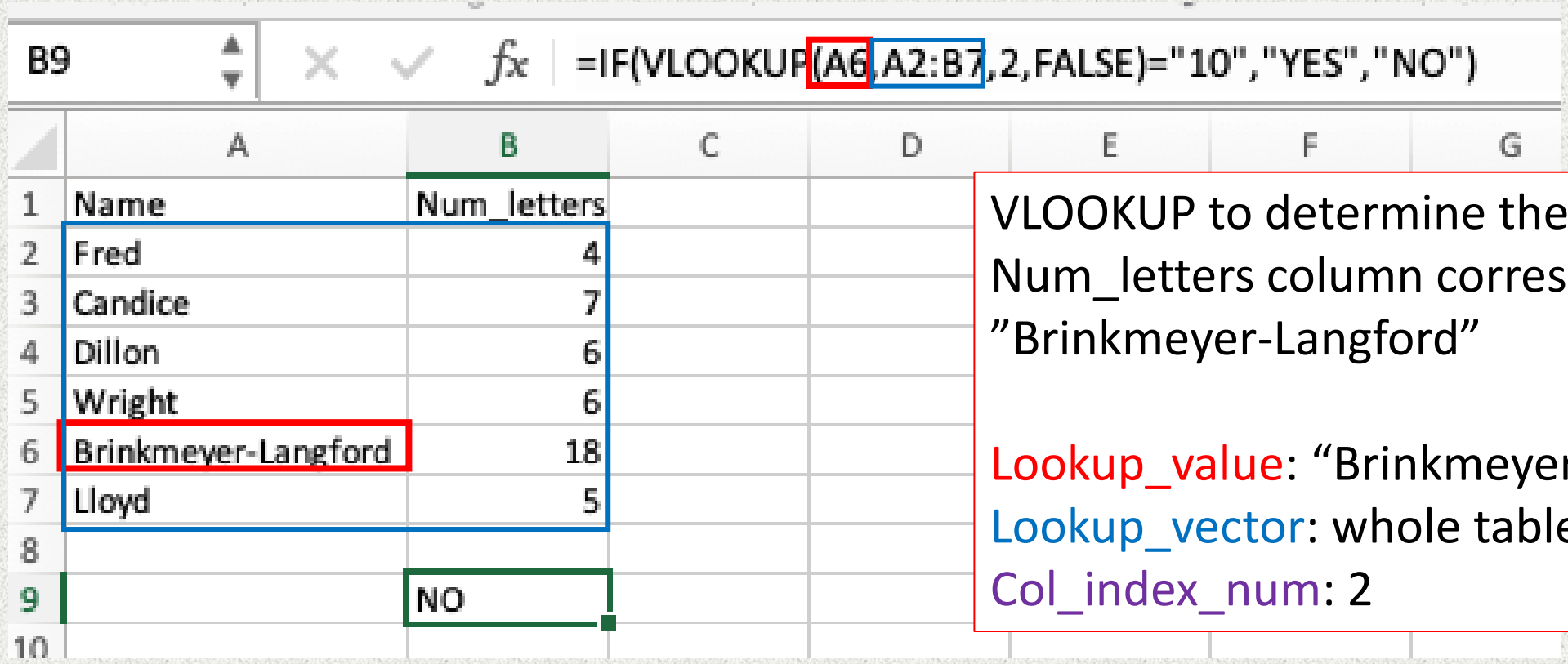
- What if I want to know whether a certain result could be found for a certain individual?
  - Here: is the value “10” found for “Brinkmeyer-Langford”?
- We can nest the VLOOKUP function inside another function, in this example an IF.
  - We just want a true/false answer here, so we start with the IF function.



	A	B	C	D	E	F	G
1	Name	Num_letters					
2	Fred	4					
3	Candice	7					
4	Dillon	6					
5	Wright	6					
6	Brinkmeyer-Langford	18					
7	Lloyd	5					
8							
9		NO					
10							

NOTE: Excel can also count or look for values that are greater or less than something, using “<“ and “>”.

Nested VLOOKUP



VLOOKUP to determine the value in the Num_letters column corresponding to ”Brinkmeyer-Langford”

Lookup_value: “Brinkmeyer-Langford” or A6
Lookup_vector: whole table
Col_index_num: 2

Nested VLOOKUP

B9 =IF(VLOOKUP(A6,A2:B7,2,FALSE)="10","YES","NO")

	A	B	C	D	E	F	G
1	Name	Num_letters					
2	Fred	4					
3	Candice	7					
4	Dillon	6					
5	Wright	6					
6	Brinkmeyer-Langford	18					
7	Lloyd	5					
8							
9		NO					
10							

=IF(VLOOKUP(A6,A2:B7,2,FALSE)=”10”,”YES”,”NO”)

TIP: KEEP AN EYE ON THE PARENTHESES!

The **VLOOKUP** part is all together, *within* the **IF** part. Excel color-codes them for you automatically.

Excel sees you want to do an **IF** (true/false) test, but will first complete the **VLOOKUP** part, then use the result value it gets to finish working out the **IF** part. If parentheses aren’t in the proper place, you’ll get an error message.

QUESTION 3:

- Which of these formulas will calculate the total number of letters in "Candice Brinkmeyer-Langford"?

A
=(VLOOKUP("Candice",A2:B7,2,FALSE))+(VLOOKUP("Brinkmeyer-Langford",A2:B7,2,FALSE))

B
=LOOKUP(A3,B2:B7)+LOOKUP(A6,B2:B7)

C
=(VLOOKUP(A3,A2:B7,2,FALSE))+(VLOOKUP(A6,A2:B7,2,FALSE))

D
=VLOOKUP((A3,A2:B7,2,FALSE)+(A6,A2:B7,2,FALSE))

	A	B
1	Name	Num_letters
2	Fred	4
3	Candice	7
4	Dillon	6
5	Wright	6
6	Brinkmeyer-Langford	18
7	Lloyd	5

QUESTION 3:

- Which of these formulas will calculate the total number of letters in "Candice Brinkmeyer-Langford"?

A

=(VLOOKUP("Candice",A2:B7,2,FALSE))+(VLOOKUP("Brinkmeyer-Langford",A2:B7,2,FALSE))

~~B~~

~~=LOOKUP(A3,B2:B7)+LOOKUP(A6,B2:B7)~~

C

=(VLOOKUP(A3,A2:B7,2,FALSE))+(VLOOKUP(A6,A2:B7,2,FALSE))

~~D~~

~~=VLOOKUP((A3,A2:B7,2,FALSE)+(A6,A2:B7,2,FALSE))~~

	A	B
1	Name	Num_letters
2	Fred	4
3	Candice	7
4	Dillon	6
5	Wright	6
6	Brinkmeyer-Langford	18
7	Lloyd	5

=VLOOKUP(lookup_value,table_array,col_index_num,[range_lookup])

Another example of a nested VLOOKUP

- How can I determine whether Brinkmeyer-Langford is the longest name?
- And if it is longest, how much longer is “Brinkmeyer-Langford” than the average name?

	A	B
1	Name	Num_letters
2	Fred	4
3	Candice	7
4	Dillon	6
5	Wright	6
6	Brinkmeyer-Langford	18
7	Lloyd	5

Another example of a nested VLOOKUP

- How can I determine whether Brinkmeyer-Langford is the longest name?
- And if it is longest, how much longer is "Brinkmeyer-Langford" than the average name?

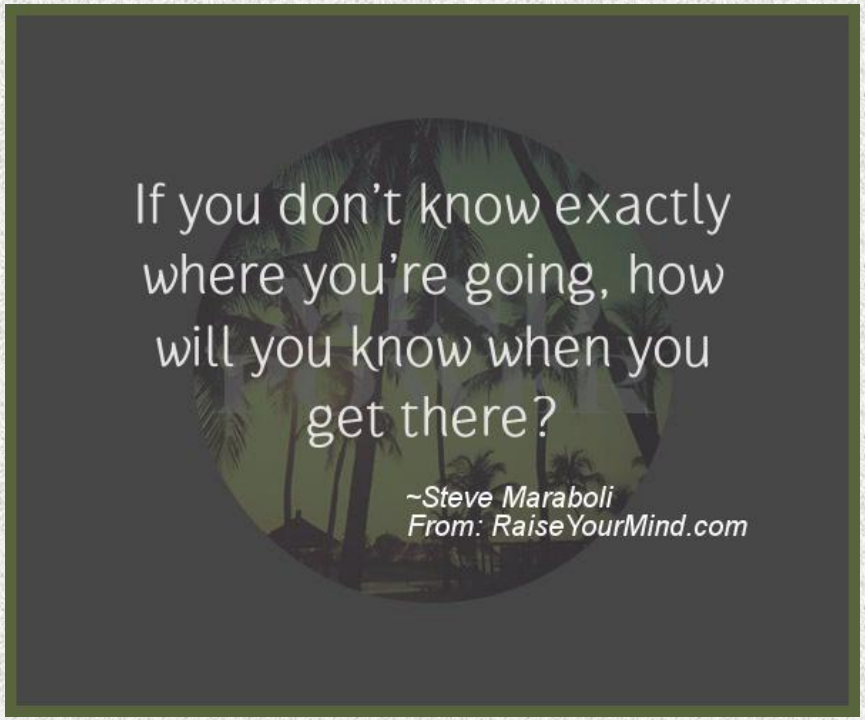
	A	B
1	Name	Num_letters
2	Fred	4
3	Candice	7
4	Dillon	6
5	Wright	6
6	Brinkmeyer-Langford	18
7	Lloyd	5

=IF((VLOOKUP(A6,A2:B7,2,FALSE))=(MAX(B2:B7)),(VLOOKUP(A6,A2:B7,2,FALSE))-(AVERAGE(B2:B7)),0)

If the value for # of letters in Brinkmeyer-Langford is the same as the maximum value in the Num_letters column, then how many more letters does Brinkmeyer-Langford have than the average value of the Num_letters column? Or if the # of letters in Brinkmeyer-Langford is NOT equal to the greatest value for Num_letters, return "0".

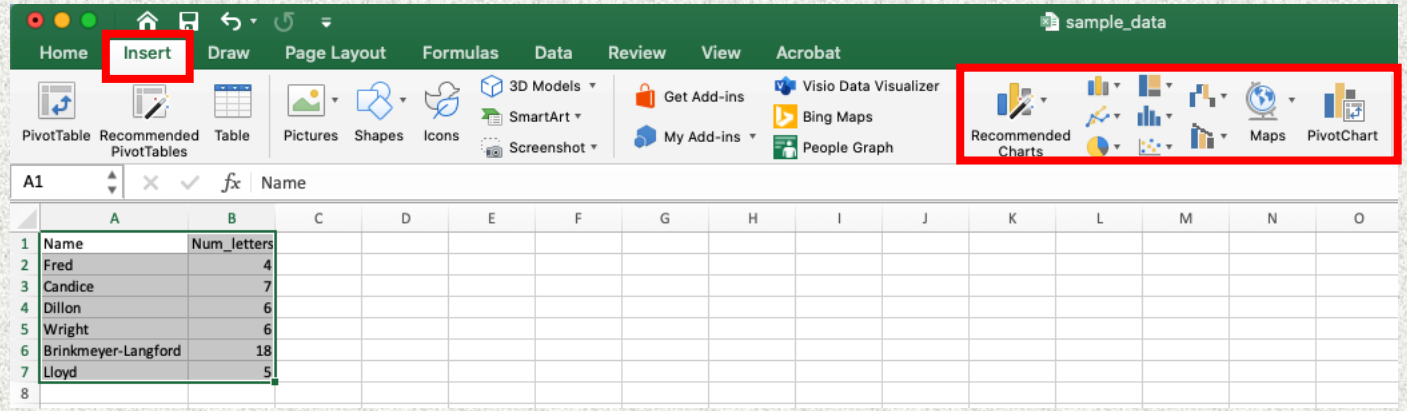
Charting data

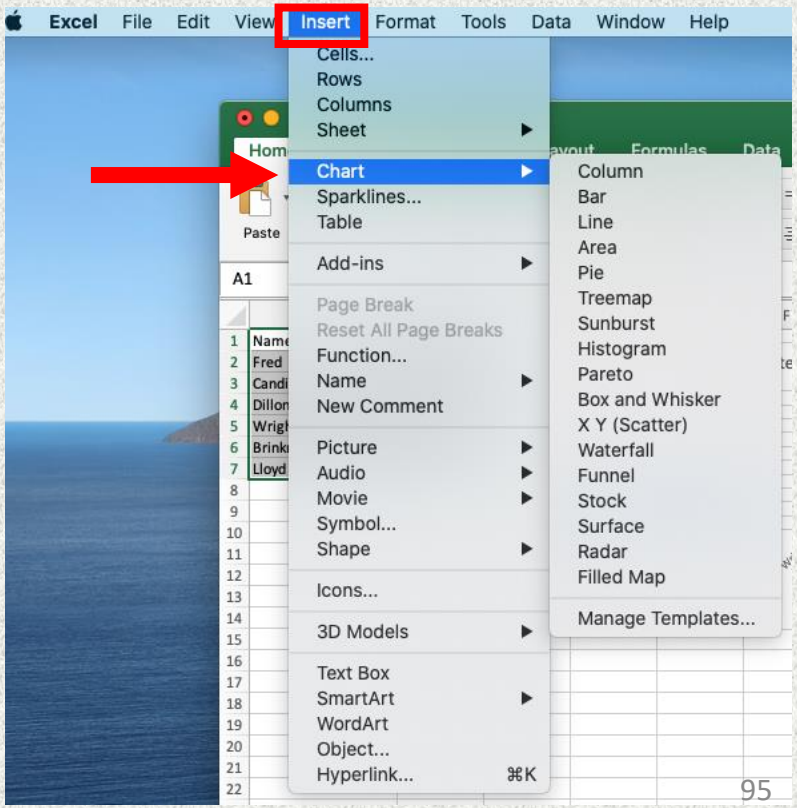
- A good hypothesis will take you a long way in knowing:
  - What to chart
  - What type of chart to use
  - What data to represent on X and Y axes
  - If a secondary axis could be useful for telling the story



Bar charts

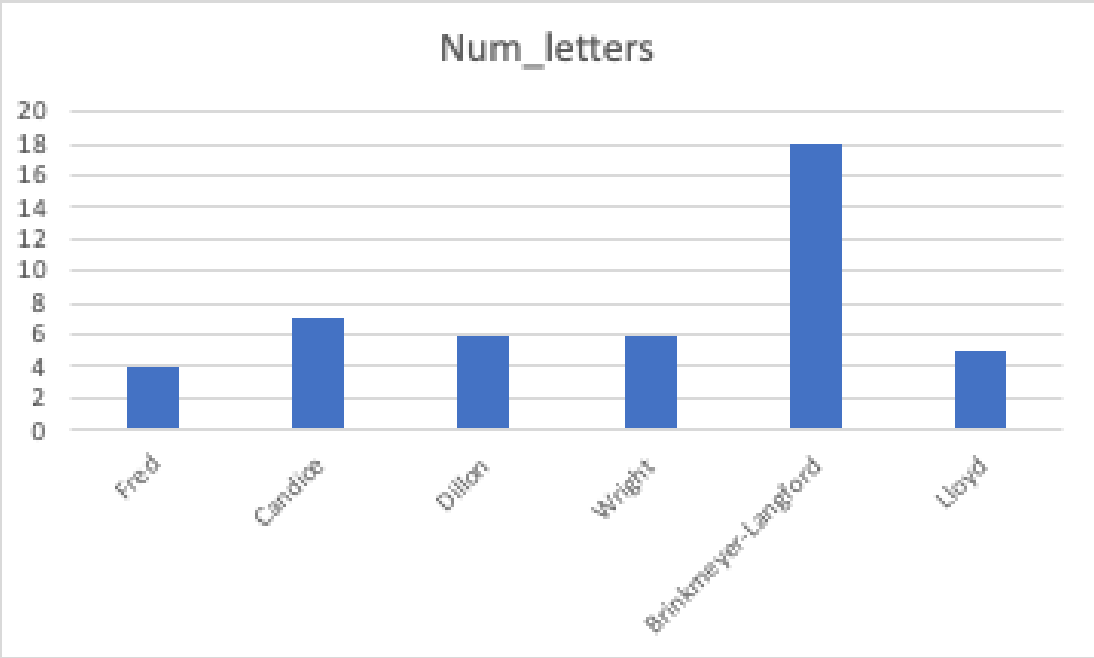
- Start by selecting the data you wish to graph.
- Then, **Insert** (either Insert menu gets you to Chart)

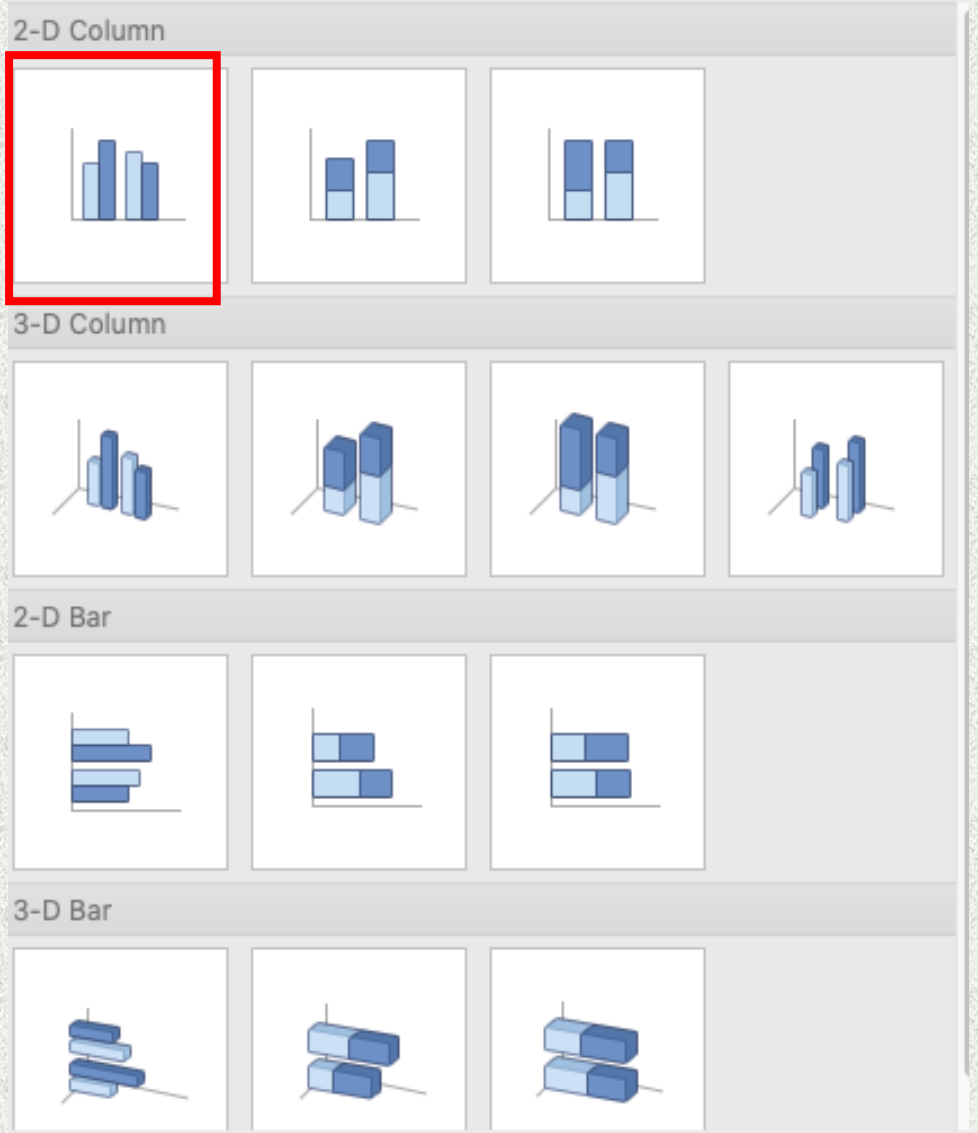




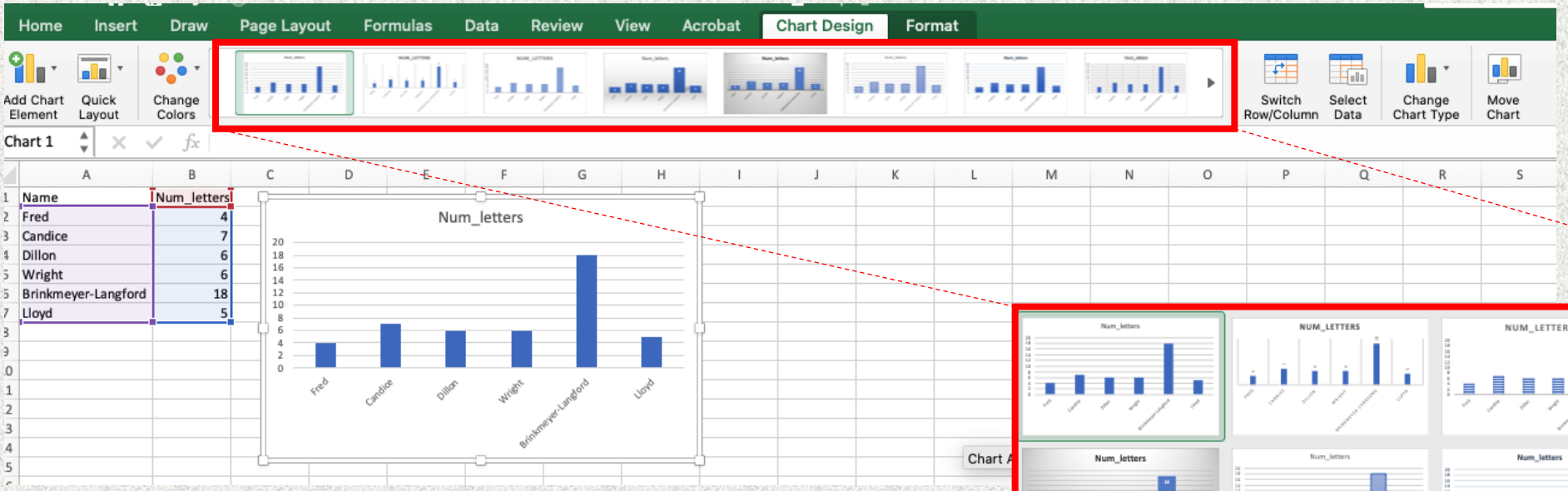
Bar charts

- Excel has multiple options for bar charts.
- Simplest: **2-D Column**





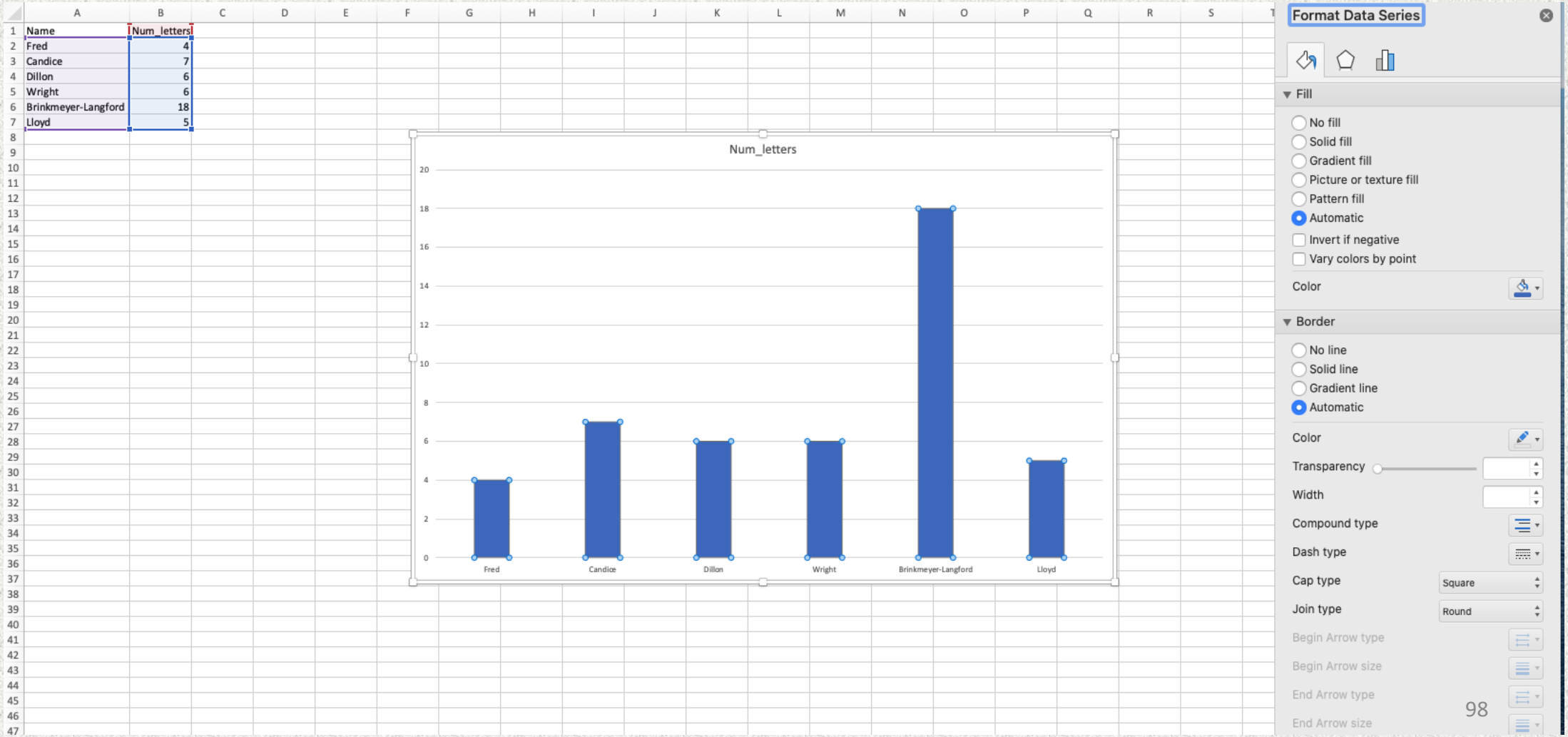
Bar charts



- Customize your look from the menu, or…

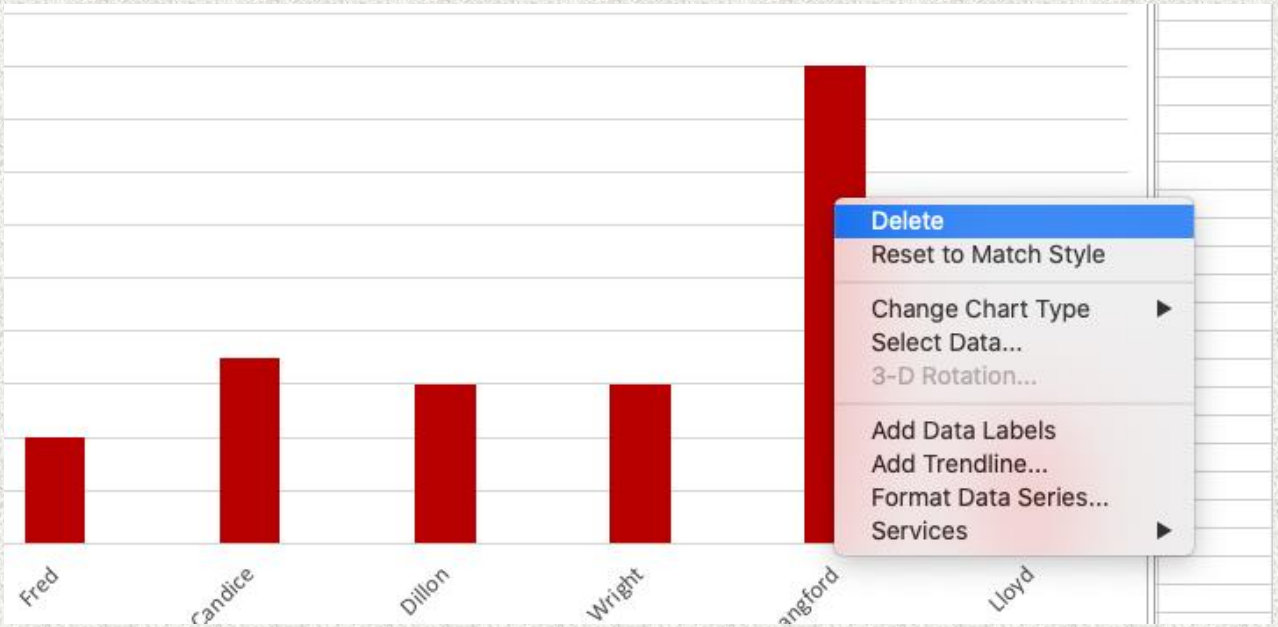
Bar charts

- Change colors, font, etc. by selecting the item.



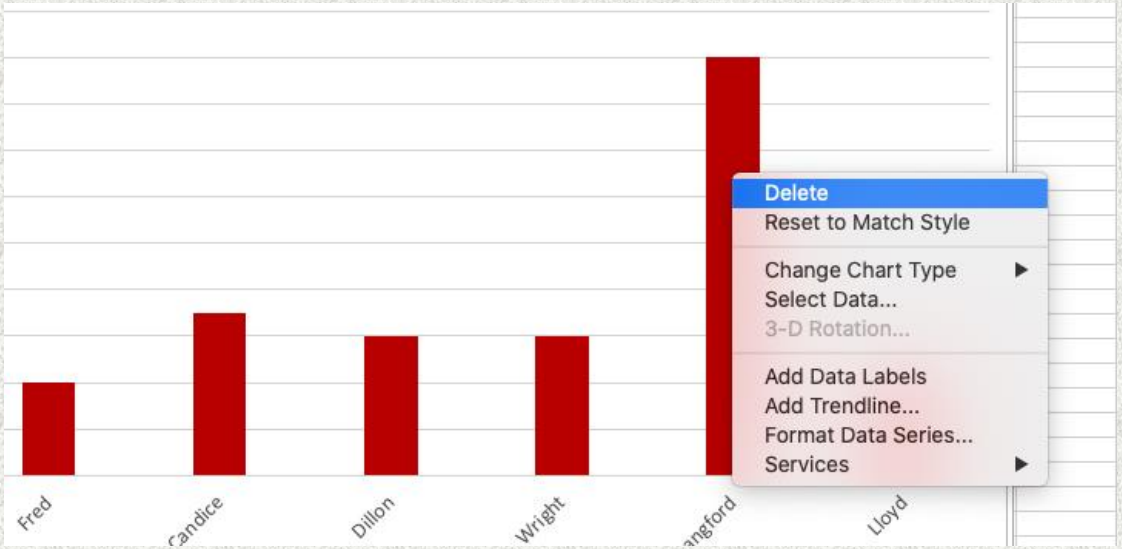
Bar charts

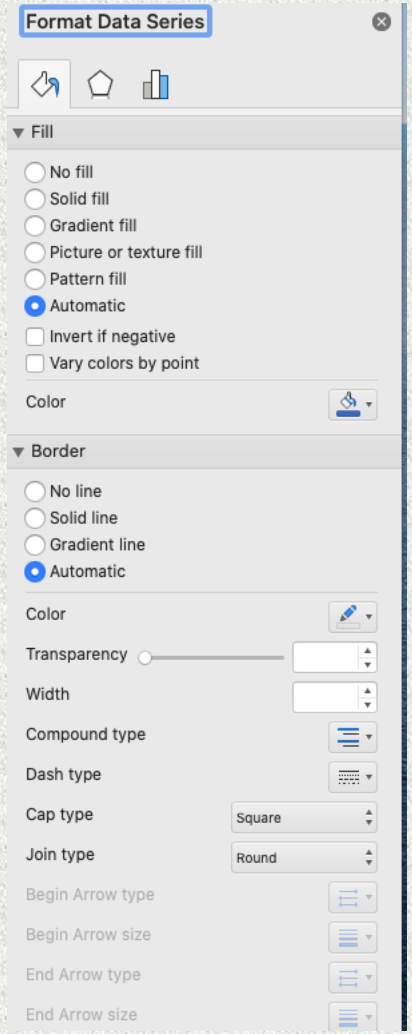
- Change colors, font, etc. by selecting the item.
- Right click to get a menu.
- Choose "**Format [whatever]**"



Bar charts

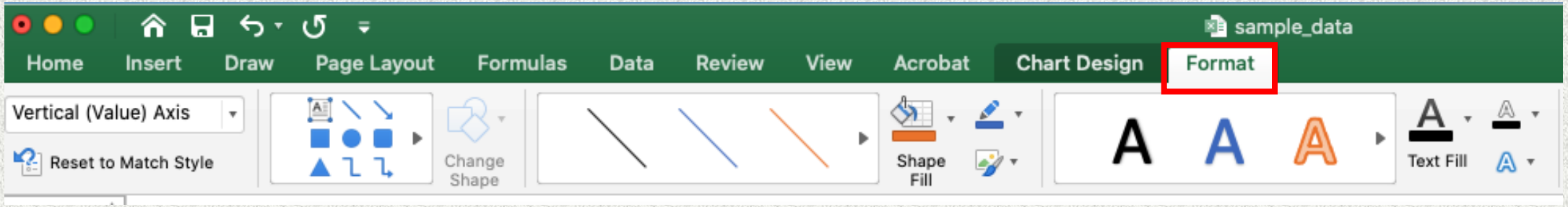
- Change colors, font, etc. by selecting the item.
- Right click to get a menu.
- Here, I right clicked on one of the bars and chose "**Format Data Series**" to change the bar colors.
- A menu appears to the right.



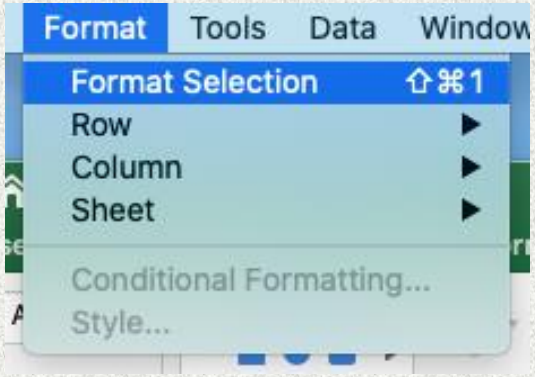


Bar charts

- Another option: select the Format menu at the top of the spreadsheet window.

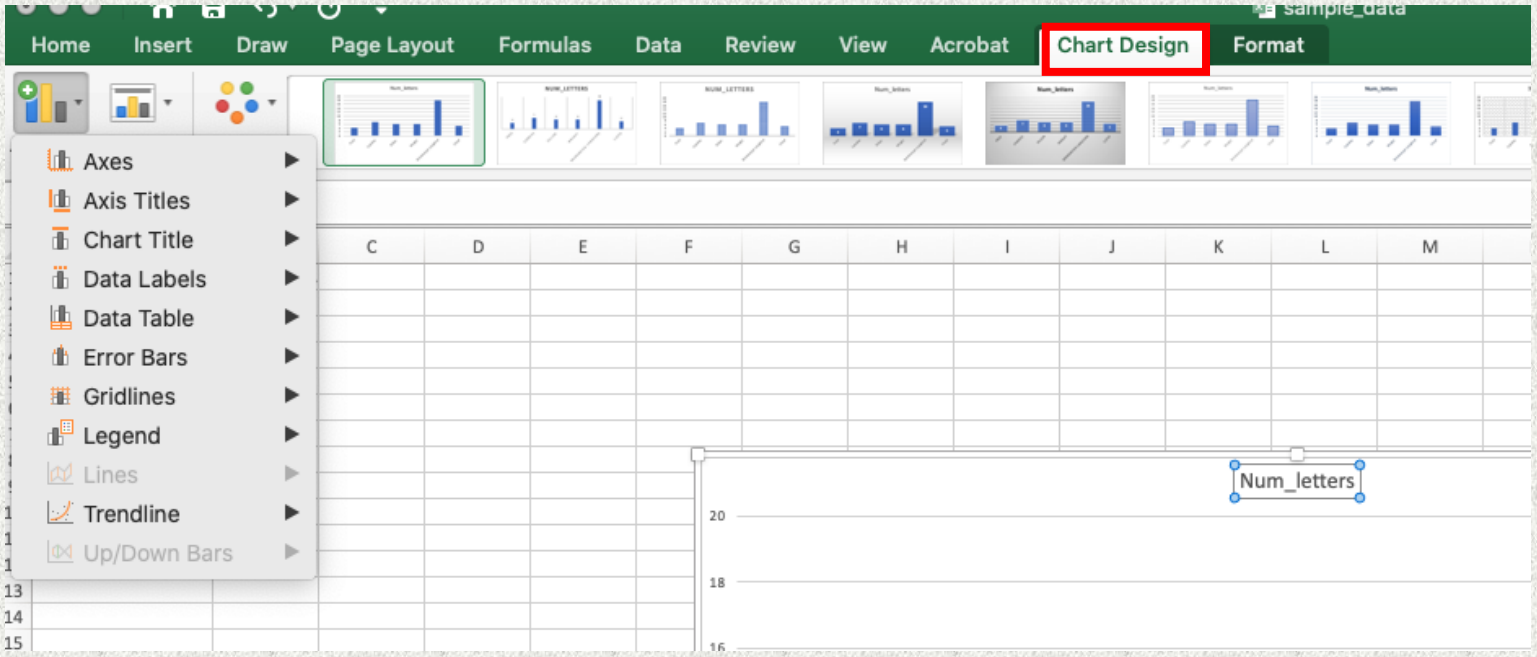


- You can also use the Format menu at the top of the Excel window.



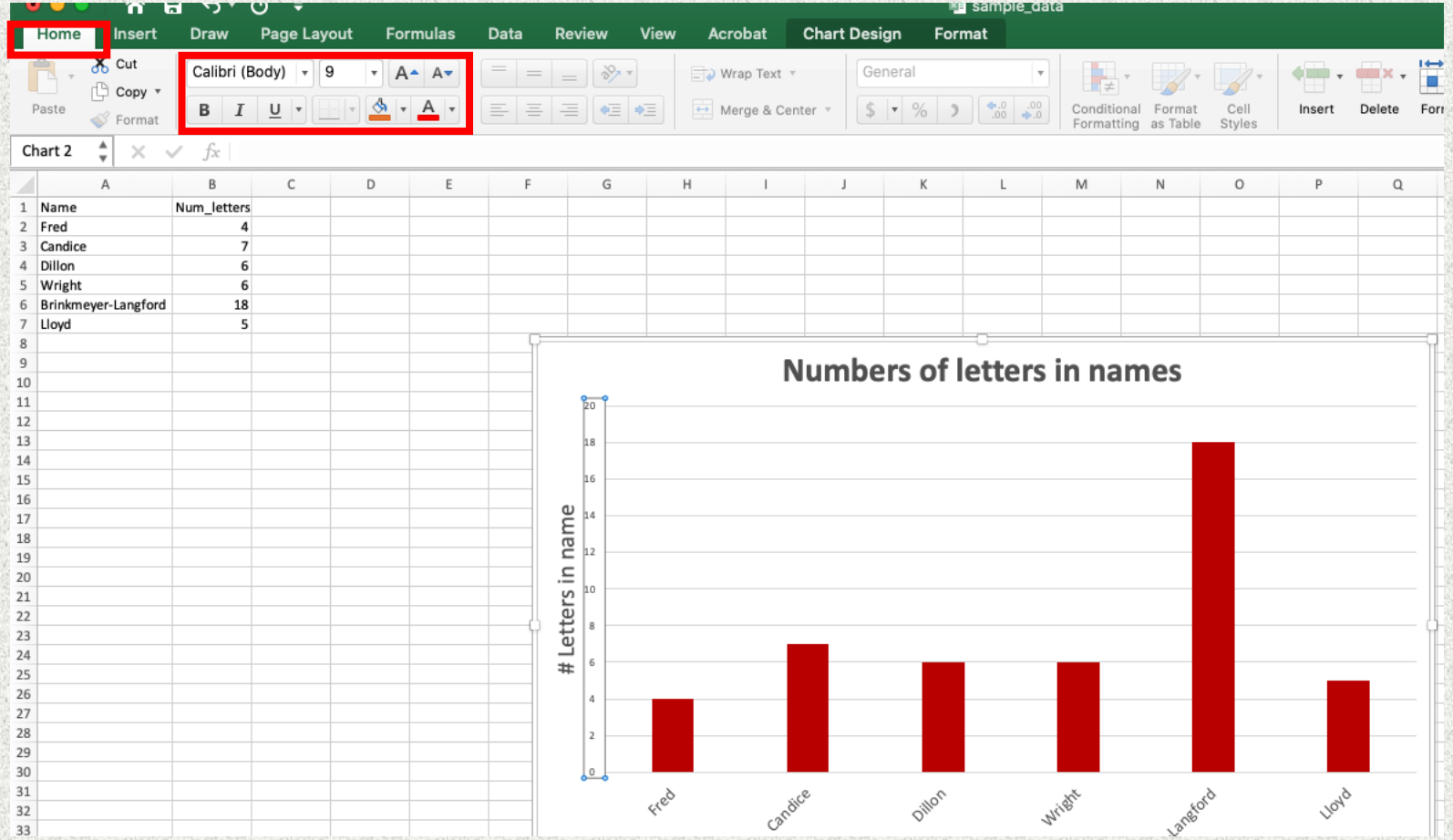
Bar charts

- Add additional items like axes titles, error bars, etc. using **Chart Design** menu at the top.



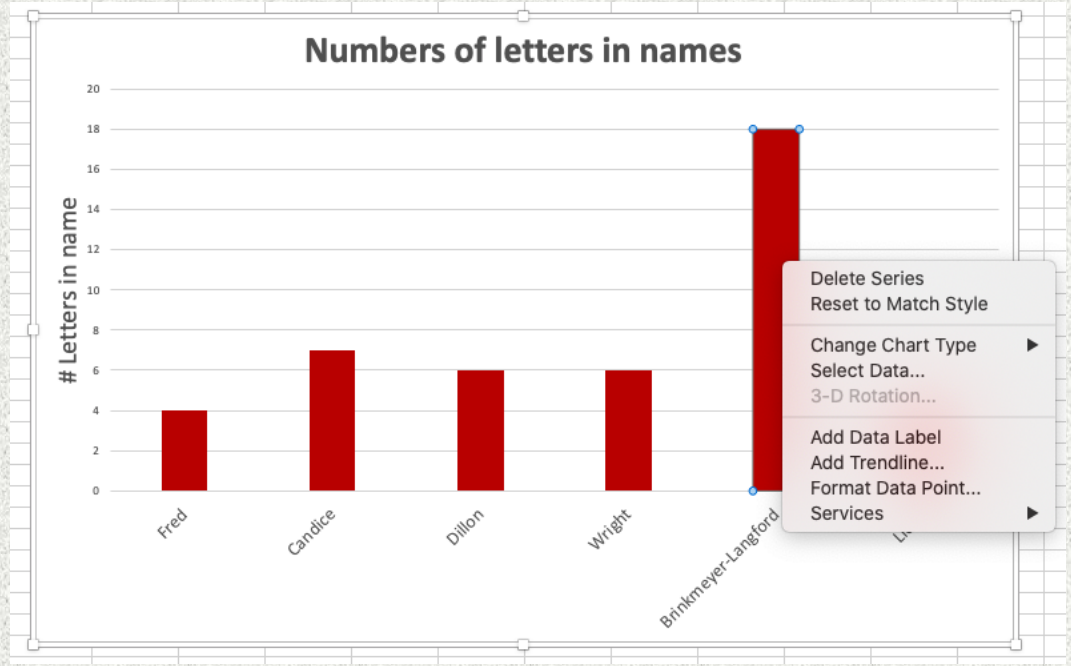
Bar charts

- To change font options for any part of the graph, select what you want to changes, then use use the **Home** menu.



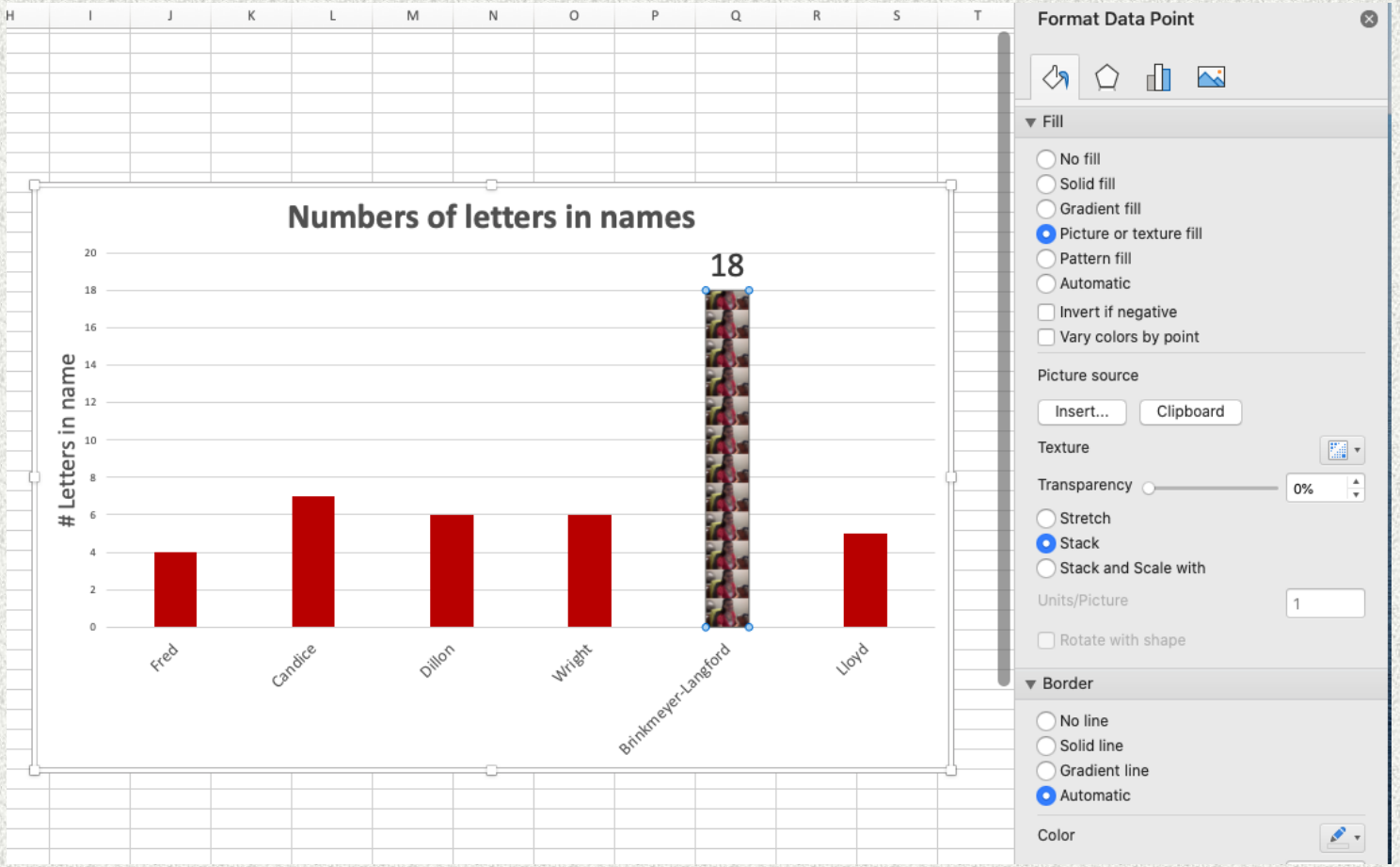
Bar charts

- Here, I have selected one data point (Brinkmeyer-Langford).
  - Format Data Point lets me change color, pattern, etc. just for that one bar.
  - Add Data Label shows the values for each selected data point (or for everything, if you choose).



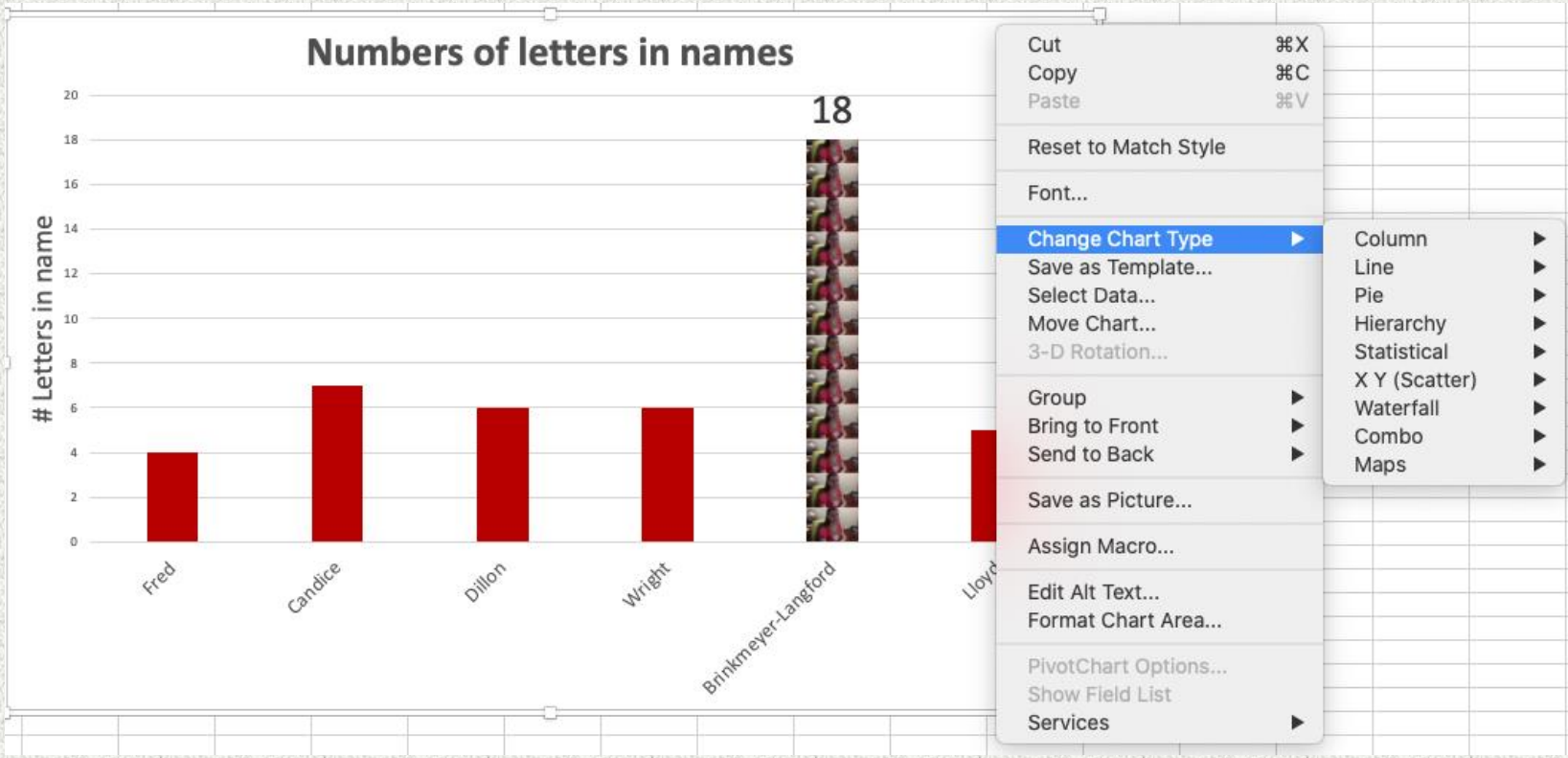
Bar charts

- Here, I have selected one data point (Brinkmeyer-Langford).
  - Format Data Point lets me change color, pattern, etc. just for that one bar.
  - Add Data Label shows the values for each selected data point (or for everything, if you choose).



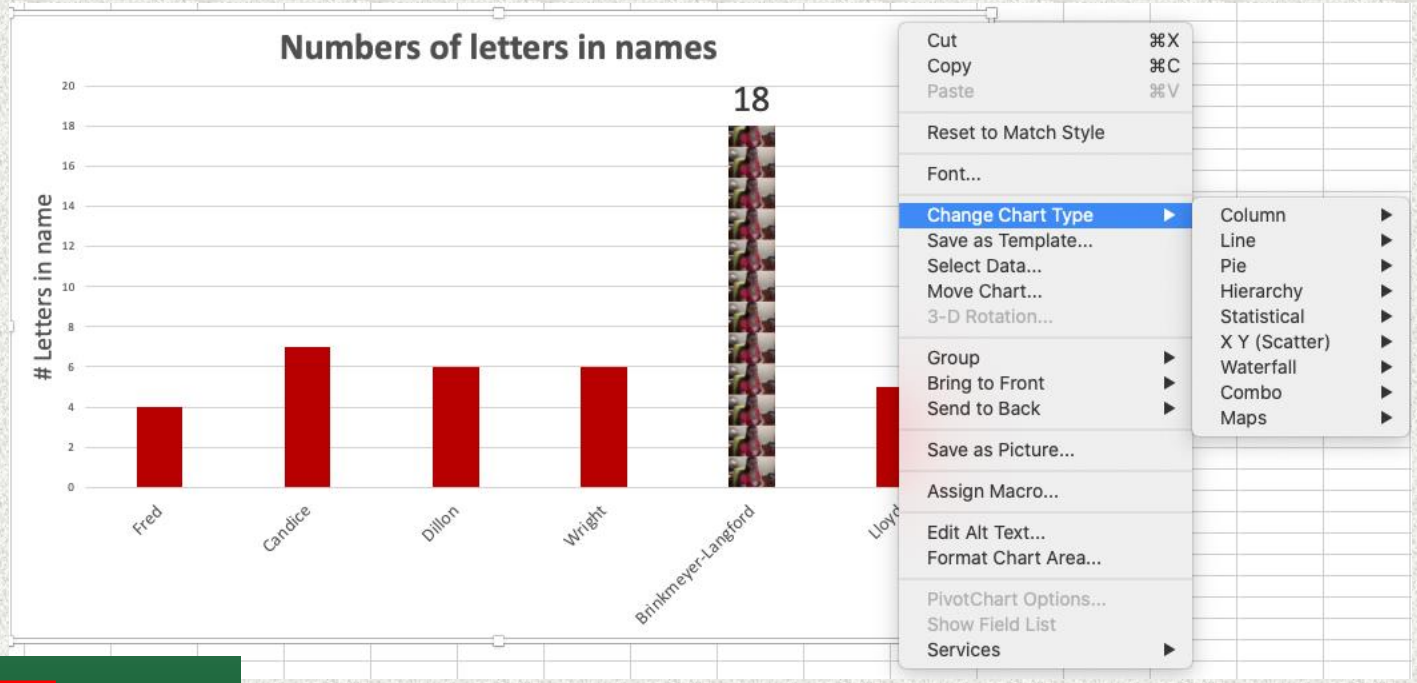
Changing your chart

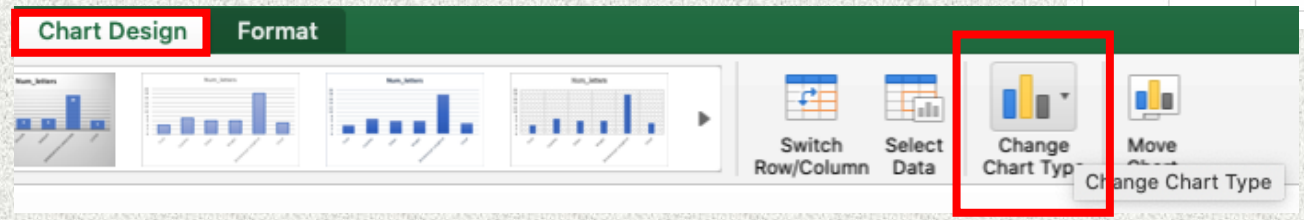
- Changing your chart type entirely:
  - Select the whole chart, right click, **Change Chart Type**



Changing your chart

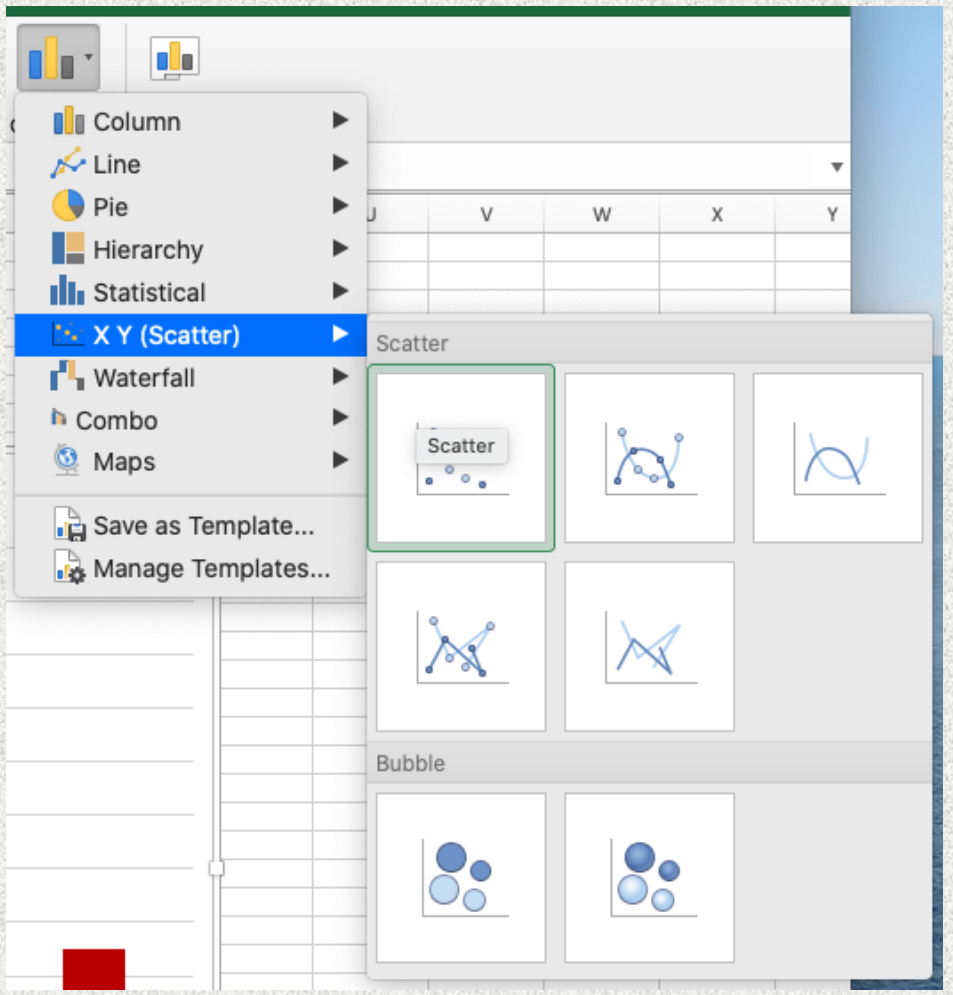
- Changing your chart type entirely:
  - Select the whole chart, right click, **Change Chart Type**
- OR, menu at top of spreadsheet:
  - **Chart Design**
  - **Change Chart Type**





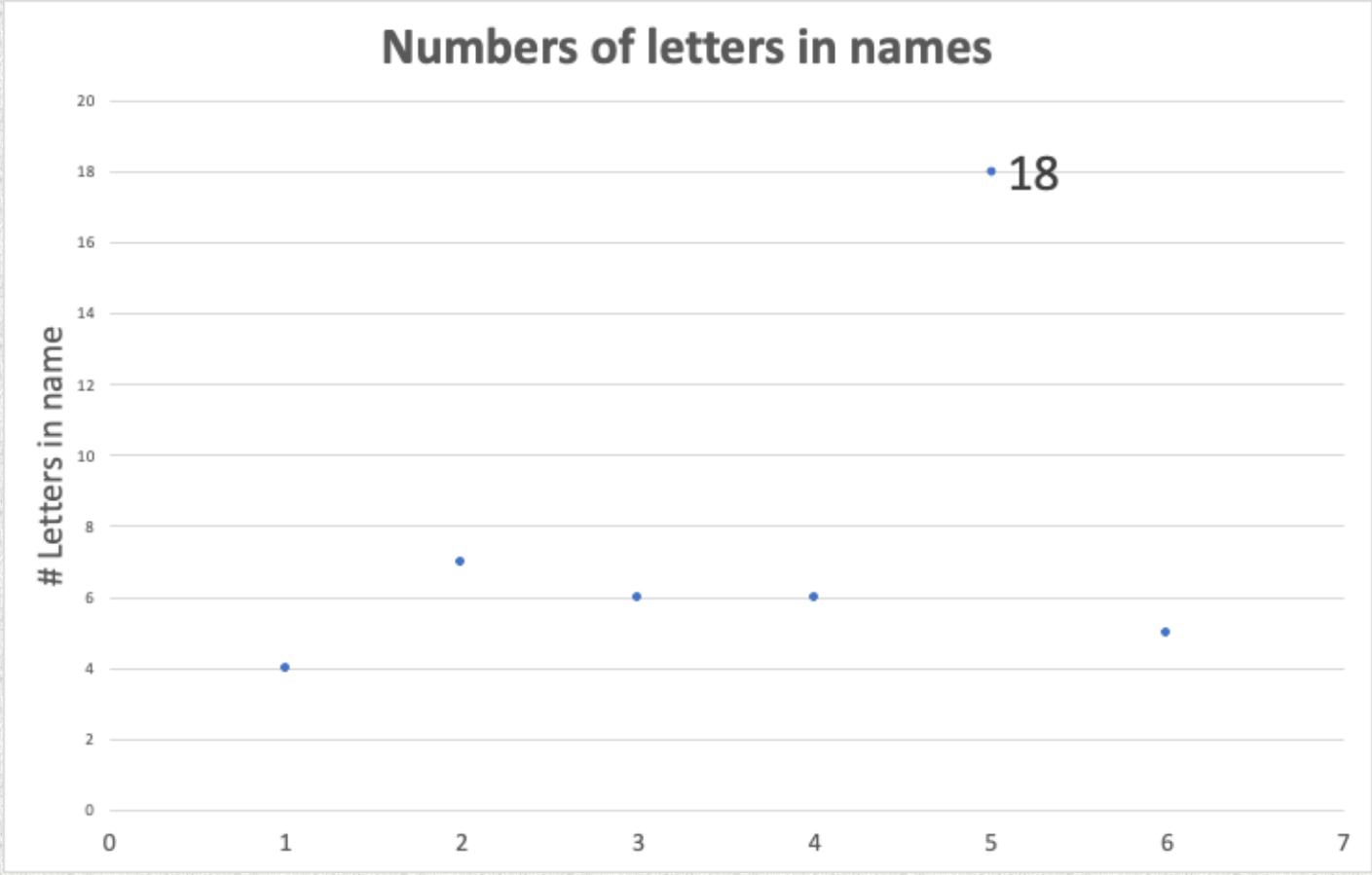
Changing your chart

- Let's try scatter plotting...



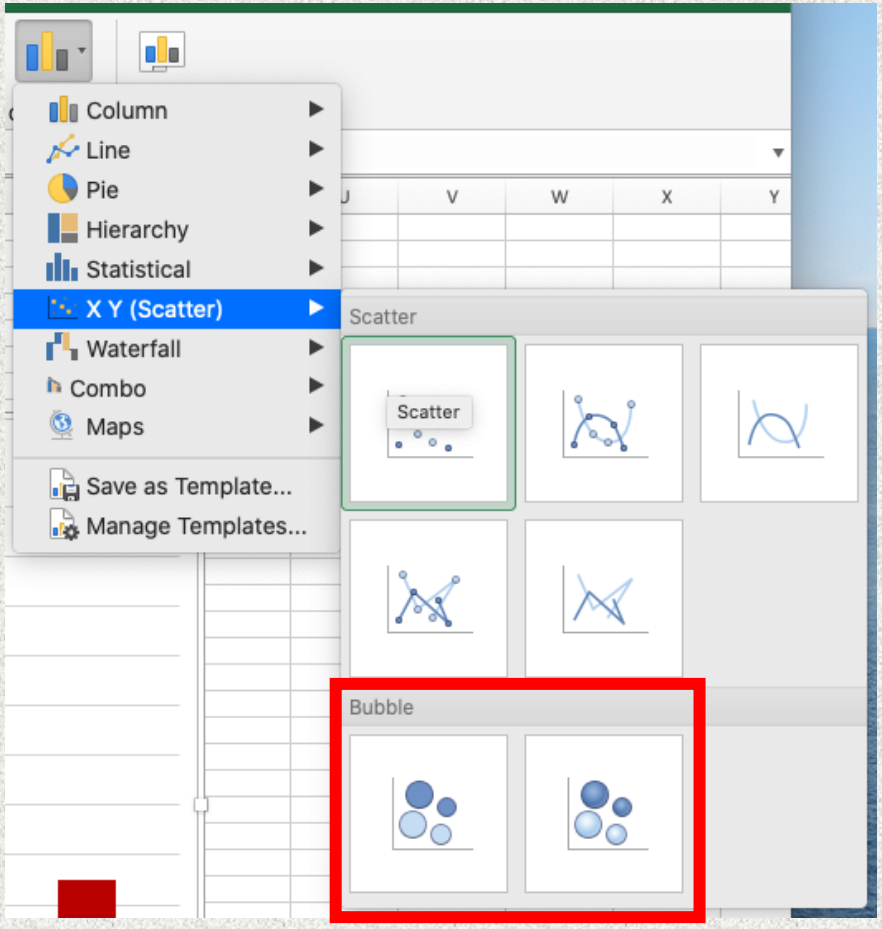
Changing your chart

- Let's try scatter plotting...
- Not ideal here.



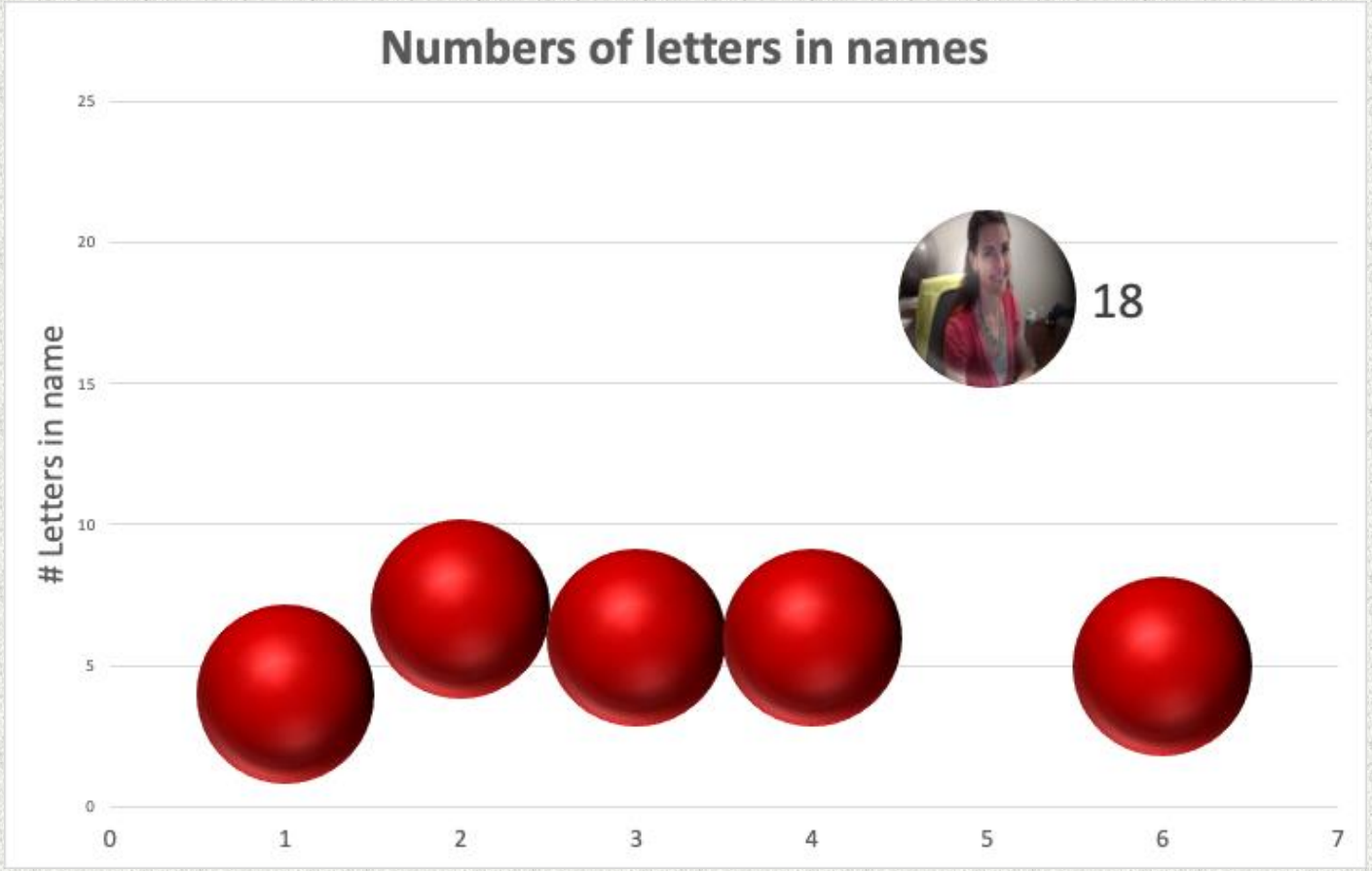
Changing your chart

- Bubble scatter plot?!



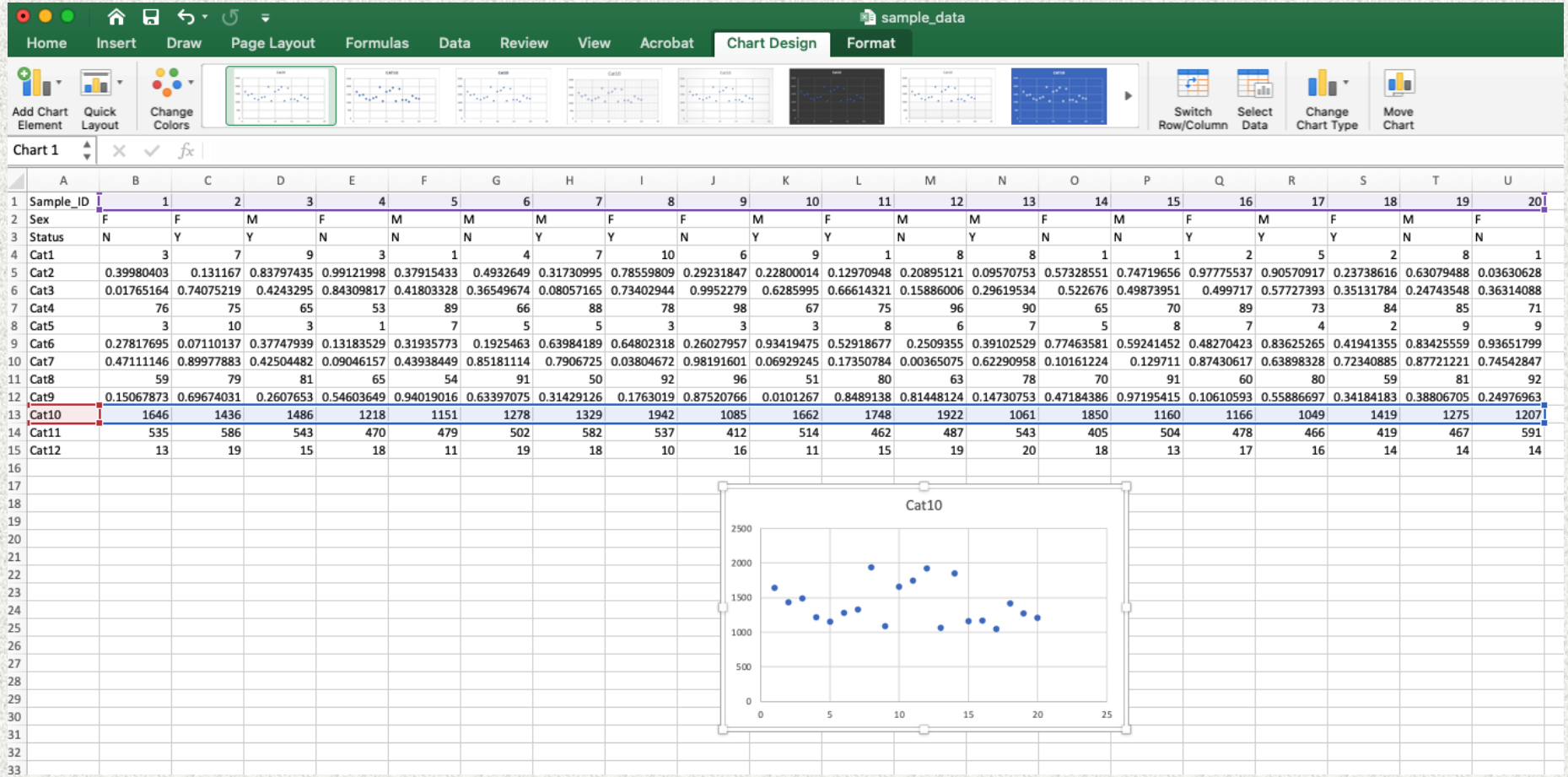
Changing your chart

- Bubble scatter plot
  - Can vary bubble sizes, colors, data labels, etc. as before



Scatterplots

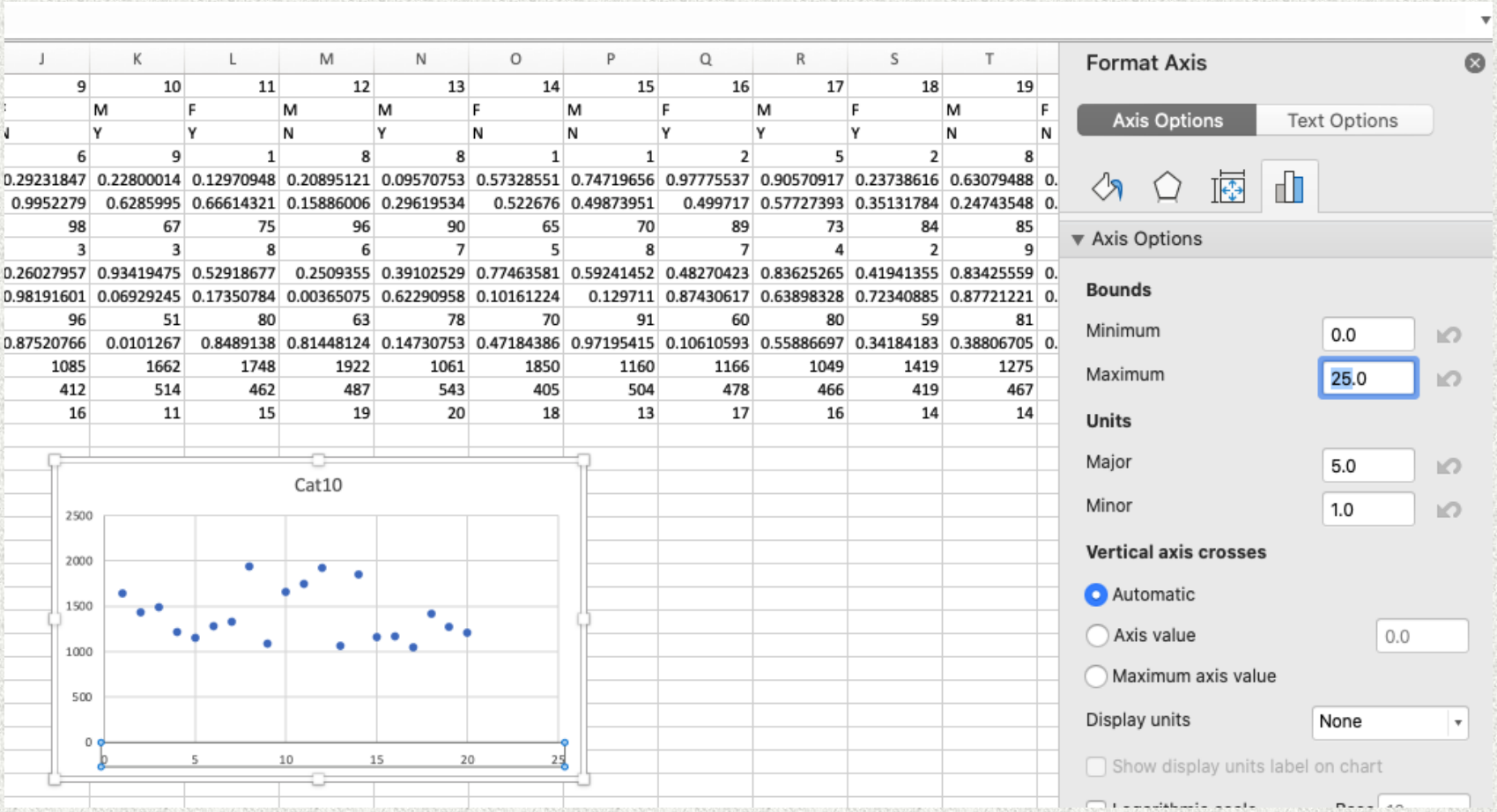
- Simple scatterplot to compare Cat10 values



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Sample_ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2	Sex	F	F	M	F	M	M	M	F	F	M	F	M	M	F	M	F	M	F	M	F
3	Status	N	Y	Y	N	N	N	Y	Y	N	Y	Y	N	Y	N	N	Y	Y	Y	N	N
4	Cat1	3	7	9	3	1	4	7	10	6	9	1	8	8	1	1	2	5	2	8	1
5	Cat2	0.39980403	0.131167	0.83797435	0.99121998	0.37915433	0.4932649	0.31730995	0.78559809	0.29231847	0.22800014	0.12970948	0.20895121	0.09570753	0.57328551	0.74719656	0.97775537	0.90570917	0.23738616	0.63079488	0.03630628
6	Cat3	0.01765164	0.74075219	0.4243295	0.84309817	0.41803328	0.36549674	0.08057165	0.73402944	0.9952279	0.6285995	0.66614321	0.15886006	0.29619534	0.522676	0.49873951	0.499717	0.57727393	0.35131784	0.24743548	0.36314088
7	Cat4	76	75	65	53	89	66	88	78	98	67	75	96	90	65	70	89	73	84	85	71
8	Cat5	3	10	3	1	7	5	5	3	3	3	8	6	7	5	8	7	4	2	9	9
9	Cat6	0.27817695	0.07110137	0.37747939	0.13183529	0.31935773	0.1925463	0.63984189	0.64802318	0.26027957	0.93419475	0.52918677	0.2509355	0.39102529	0.77463581	0.59241452	0.48270423	0.83625265	0.41941355	0.83425559	0.93651799
10	Cat7	0.47111146	0.89977883	0.42504482	0.09046157	0.43938449	0.85181114	0.7906725	0.03804672	0.98191601	0.06929245	0.17350784	0.00365075	0.62290958	0.10161224	0.129711	0.87430617	0.63898328	0.72340885	0.87721221	0.74542847
11	Cat8	59	79	81	65	54	91	50	92	96	51	80	63	78	70	91	60	80	59	81	92
12	Cat9	0.15067873	0.69674031	0.2607653	0.54603649	0.94019016	0.63397075	0.31429126	0.1763019	0.87520766	0.0101267	0.8489138	0.81448124	0.14730753	0.47184386	0.97195415	0.10610593	0.55886697	0.34184183	0.38806705	0.24976963
13	Cat10	1646	1436	1486	1218	1151	1278	1329	1942	1085	1662	1748	1922	1061	1850	1160	1166	1049	1419	1275	1207
14	Cat11	535	586	543	470	479	502	582	537	412	514	462	487	543	405	504	478	466	419	467	591
15	Cat12	13	19	15	18	11	19	18	10	16	11	15	19	20	18	13	17	16	14	14	14

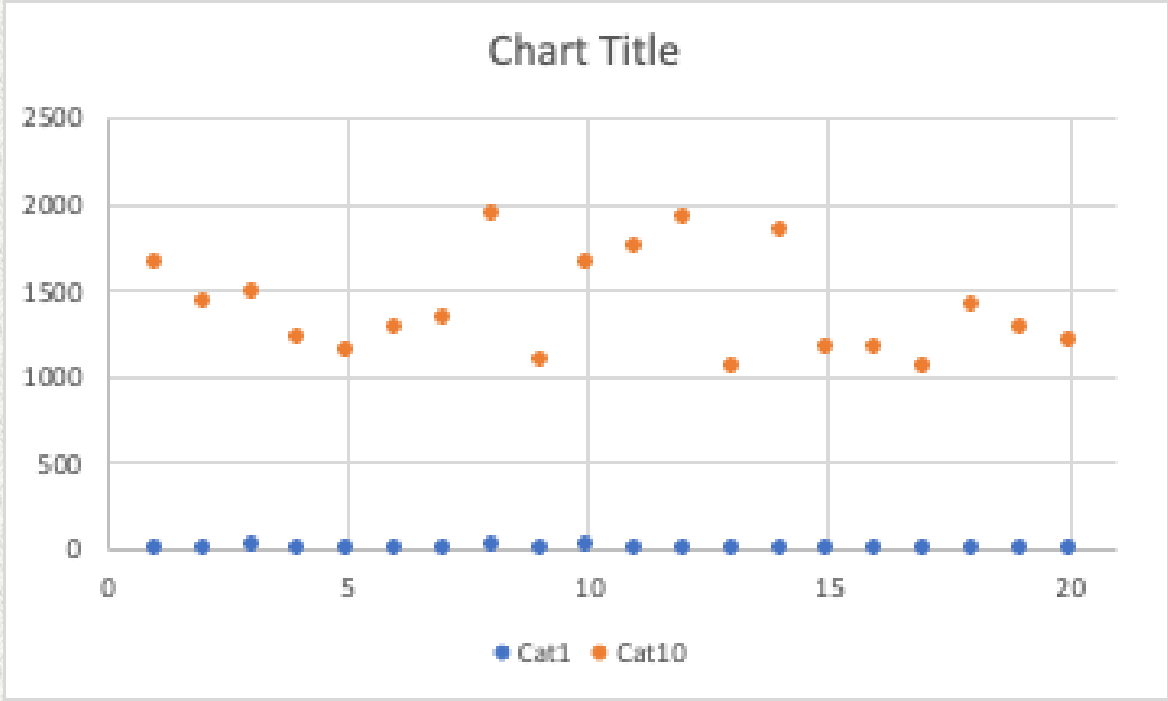
Scatterplots

- Format chart as before…
  - Here, updating x-axis bounds.



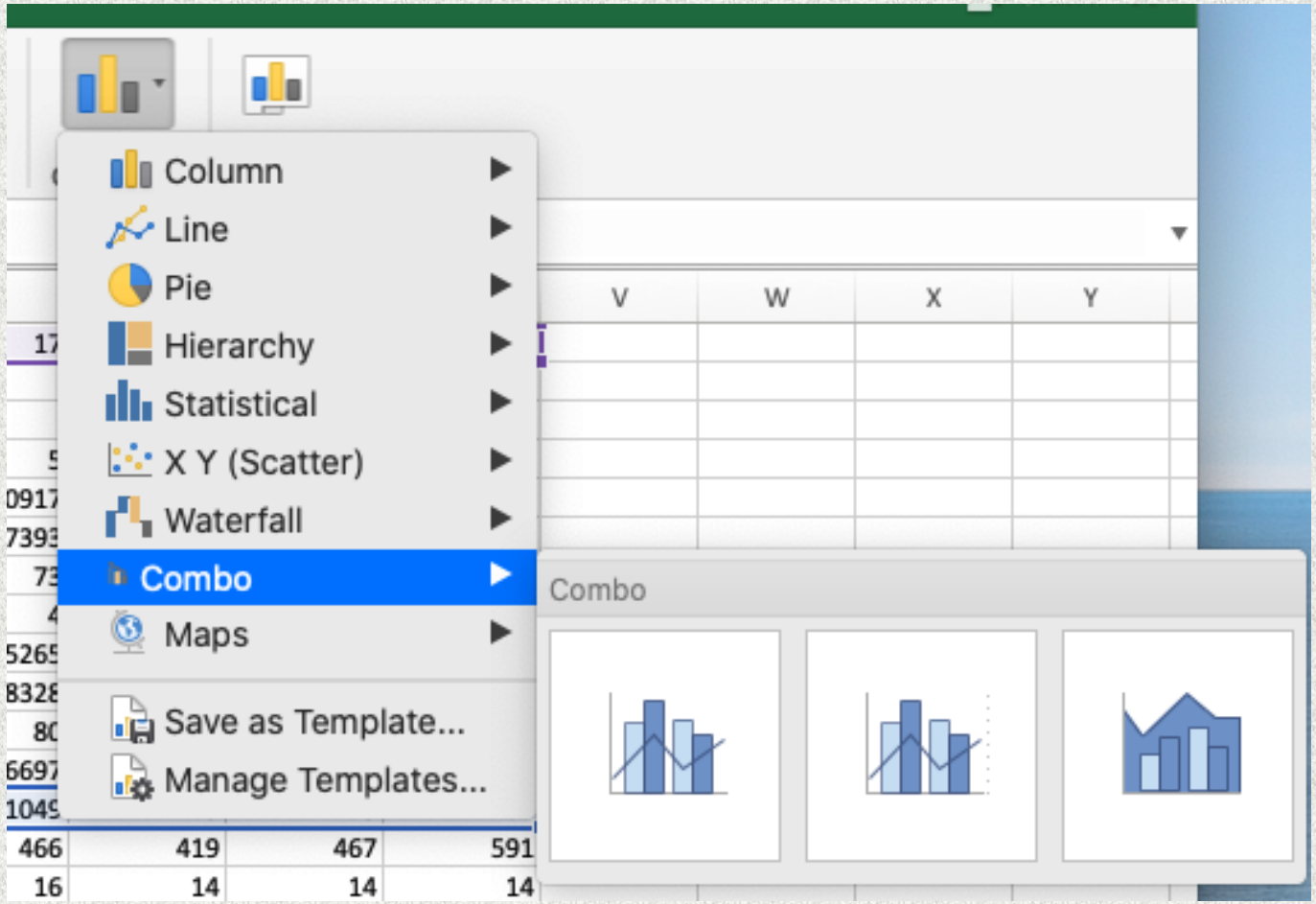
Scatterplots

- We can add >1 categories to a scatterplot.
- Excel makes them different colors by default, but this can be changed using the Format menus as before.
- But here, Cat1 and Cat10 are on very different scales…



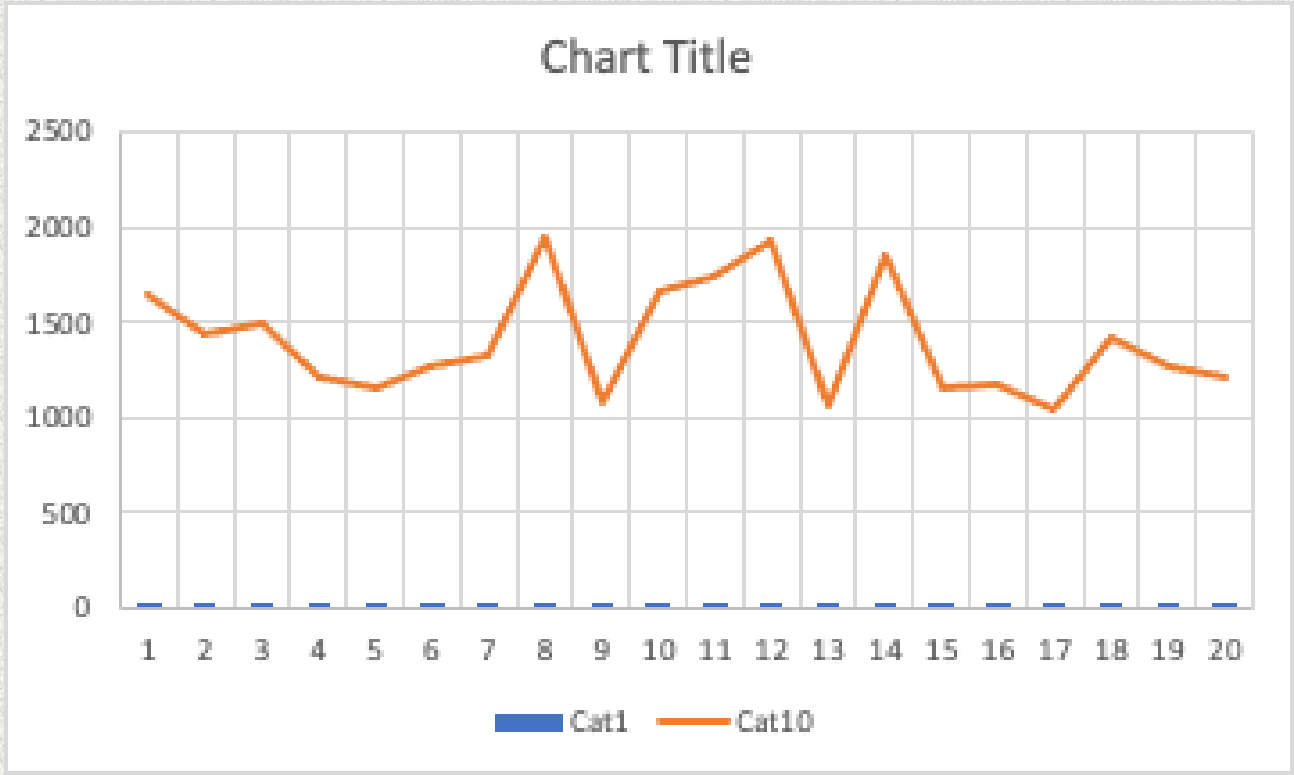
Adding a secondary axis to show data with different scales

- To represent data on separate axes (e.g. y- and z-z-axes), we start by changing the chart type to **Combo**.



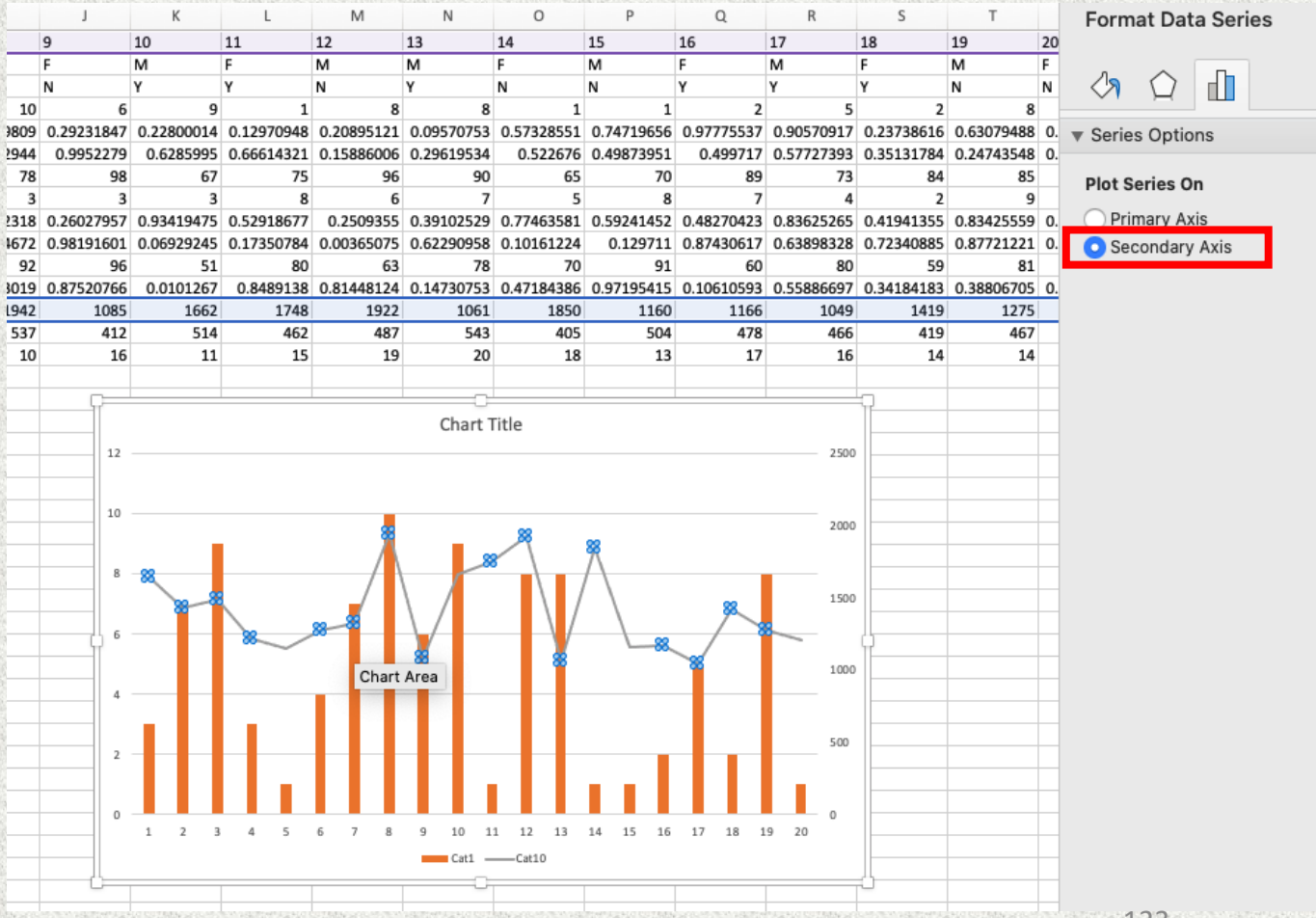
Adding a secondary axis

- Here is what it looks like at first:



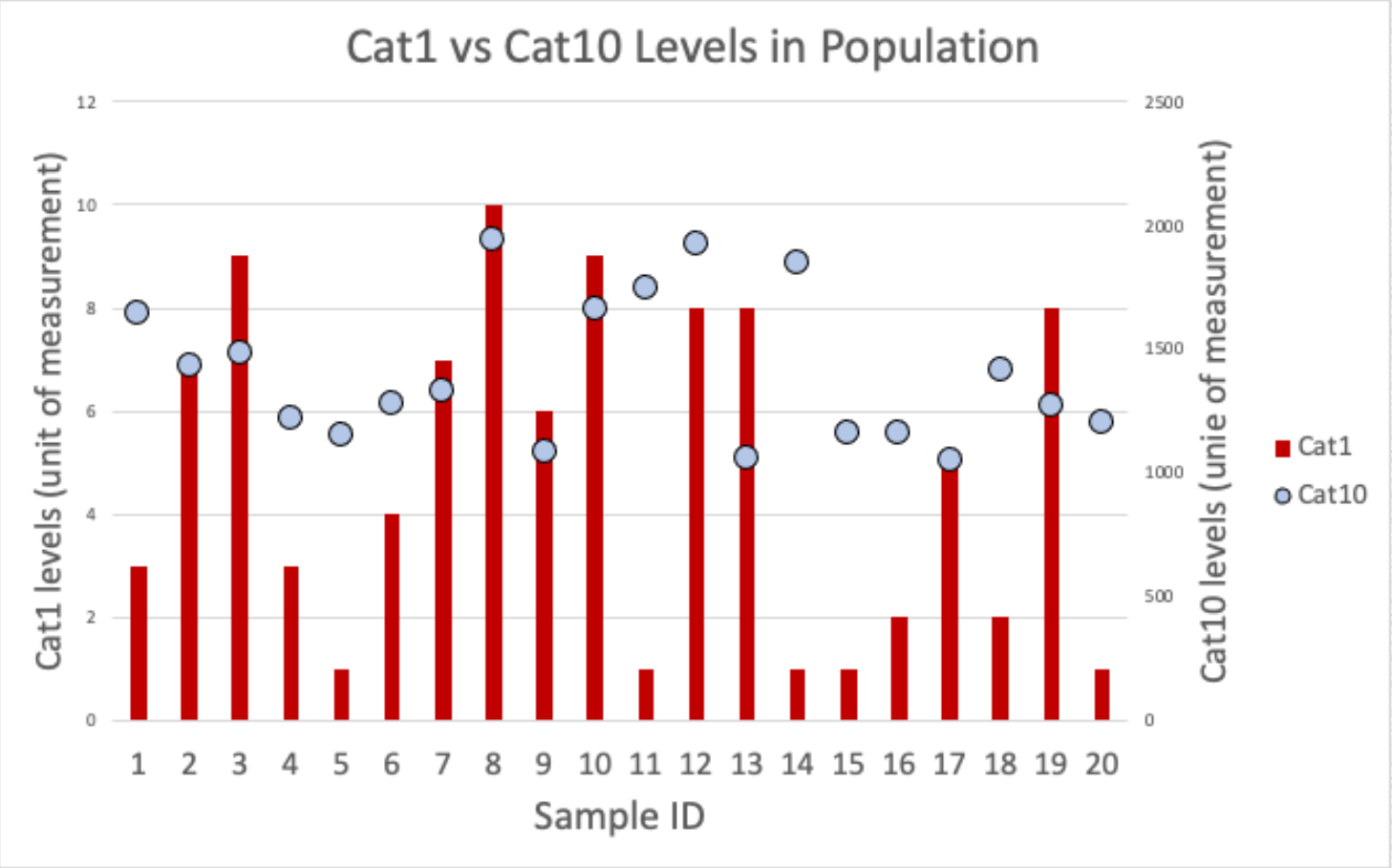
Adding a secondary axis

- Select one of the two categories.
- Right click – **Format Data Series**
- Plot series on **Secondary Axis**



Adding a secondary axis

- Format further to tell the story effectively.



Stacked bar charts

- Here, I have calculated the average values for each of the 12 categories in exposed females and unexposed females.

C16 =AVERAGEIFS(sample_values!B4:U4,sample_values!B$2:U$2,"F",sample_values!B$3:U$3,"N")

	A	B	C
1	Is #8 female?	YES	
2	Females	10	
3	Males	10	
4	Exposed	10	
5	Unexposed	10	
6	Female+Exp	5	
7	Cat1 avg F+Exp	4.4	
8	Cat1 avg F+Unexp	2.8	
9	Difference: Exp-Unexp	1.6	
10	Cat7: individual 8	0.03804672	
11	Cat7: individual 8 (v2)	0.03804672	
12			
13			
14			
15		Cat avg F+Ex	Cat avg F+Unexp
16	Cat1	4.4	2.8
17	Cat2	0.45232322	0.45858686
18	Cat3	0.59839193	0.54835892
19	Cat4	80.2	72.6
20	Cat5	6	4.2
21	Cat6	0.43008582	0.47628912
22	Cat7	0.54180968	0.47810595
23	Cat8	74	76.4
24	Cat9	0.43398075	0.45870727
25	Cat10	1542.2	1401.2
26	Cat11	496.4	482.6
27	Cat12	15	15.8

	A	B	C	D	E
1	Sample_ID	1	2	3	4
2	Sex	F	F	M	F
3	Status	N	Y	Y	N
4	Cat1	3	7	9	3
5	Cat2	0.39980403	0.131167	0.83797435	0.99121998
6	Cat3	0.01765164	0.74075219	0.4243295	0.84309817
7	Cat4	76	75	65	53
8	Cat5	3	10	3	1
9	Cat6	0.27817695	0.07110137	0.37747939	0.13183529
10	Cat7	0.47111146	0.89977883	0.42504482	0.09046157
11	Cat8	59	79	81	65
12	Cat9	0.15067873	0.69674031	0.2607653	0.54603649
13	Cat10	1646	1436	1486	1218
14	Cat11	535	586	543	470
15	Cat12	13	19	15	18
16					

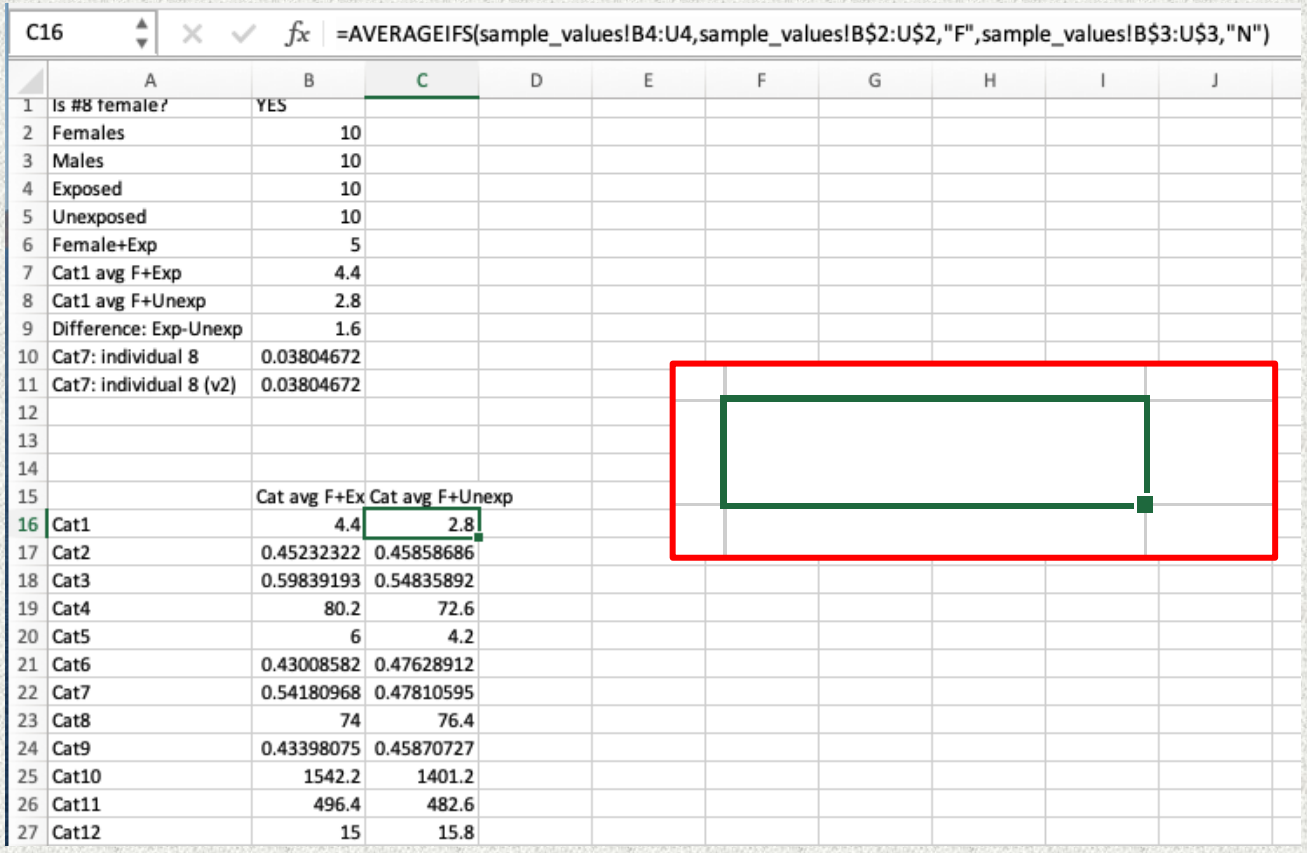
Notice where the **$** symbols are in the formula…

$ indicates what you do NOT want to change when you "drag and fill"

For example: column identity.

Stacked bar charts: Drag and fill

- ”Drag and fill”
  - Move your mouse cursor over lower right corner until it becomes “+”,
  - Hold down the right mouse button
  - drag and fill in the same formula for multiple cells located immediately below (or to the right).



	A	B	C	D	E	F	G	H	I	J
1	Is #8 female?	YES								
2	Females	10								
3	Males	10								
4	Exposed	10								
5	Unexposed	10								
6	Female+Exp	5								
7	Cat1 avg F+Exp	4.4								
8	Cat1 avg F+Unexp	2.8								
9	Difference: Exp-Unexp	1.6								
10	Cat7: individual 8	0.03804672								
11	Cat7: individual 8 (v2)	0.03804672								
12										
13										
14										
15		Cat avg F+Ex	Cat avg F+Unexp							
16	Cat1	4.4	2.8							
17	Cat2	0.45232322	0.45858686							
18	Cat3	0.59839193	0.54835892							
19	Cat4	80.2	72.6							
20	Cat5	6	4.2							
21	Cat6	0.43008582	0.47628912							
22	Cat7	0.54180968	0.47810595							
23	Cat8	74	76.4							
24	Cat9	0.43398075	0.45870727							
25	Cat10	1542.2	1401.2							
26	Cat11	496.4	482.6							
27	Cat12	15	15.8							

QUESTION 4

- I used AVERAGEIFS to calculate the Cat1 values, and then dragged the “fill handle” (lower right corner) to fill in the formula for all the other categories. Which of these formulas should I see for the selected cell?

A.=AVERAGEIFS(B4:U4,sample_values!$B2:$U$2,”F”,sample_values!$B$3:$U$3,”Y”)

B.=AVERAGEIFS(sample_values!B4:U4, sample_values!$B2:$U$2,”F”,sample_values!$B$3:$U$3,”Y”)

C.=AVERAGEIFS(sample_values!$B11:$U11,sample_values!$B$2:$U$2,"F",sample_values!$B$3:$U$3,"Y")

D.=AVERAGEIFS(sample_values!B$11:U$11,sample_values!$B$2:$U$2,"F",sample_values!$B$3:$U$3,"Y")

	A	B	C	D	E
1	Sample_ID	1	2	3	4
2	Sex	F	F	M	F
3	Status	N	Y	Y	N
4	Cat1	3	7	9	3
5	Cat2	0.39980403	0.131167	0.83797435	0.99121998
6	Cat3	0.01765164	0.74075219	0.4243295	0.84309817
7	Cat4	76	75	65	53
8	Cat5	3	10	3	1
9	Cat6	0.27817695	0.07110137	0.37747939	0.13183529
10	Cat7	0.47111146	0.89977883	0.42504482	0.09046157
11	Cat8	59	79	81	65
12	Cat9	0.15067873	0.69674031	0.2607653	0.54603649
13	Cat10	1646	1436	1486	1218
14	Cat11	535	586	543	470
15	Cat12	13	19	15	18
16					

sample_values | Sheet1

	A	B	C
15		Cat avg F+Exp	Cat avg F+Unexp
16	Cat1	4.4	2.8
17	Cat2	0.452323217	0.458586855
18	Cat3	0.598391934	0.548358917
19	Cat4	80.2	72.6
20	Cat5	6	4.2
21	Cat6	0.430085816	0.476289122
22	Cat7	0.541809681	0.478105953
23	Cat8	74	76.4
24	Cat9	0.433980755	0.458707272
25	Cat10	1542.2	1401.2
26	Cat11	496.4	482.6
27	Cat12	15	15.8

sample_values | Sheet1

QUESTION 4

- I used AVERAGEIFS to calculate the Cat1 values, and then dragged the “fill handle” (lower right corner) to fill in the formula for all the other categories. Which of these formulas should I see for the selected cell?

A.=AVERAGEIFS(B4:U4,sample_values!$B2:$U$2,”F”,sample_values!$B$3:$U$3,”Y”)

B.=AVERAGEIFS(sample_values!B4:U4, sample_values!$B2:$U$2,”F”,sample_values!$B$3:$U$3,”Y”)

C.=AVERAGEIFS(sample_values!$B11:$U11,sample_values!$B$2:$U$2,"F",sample_values!$B$3:$U$3,"Y")

D.=AVERAGEIFS(sample_values!B$11:U$11,sample_values!$B$2:$U$2,"F",sample_values!$B$3:$U$3,"Y")

	A	B	C	D	E
1	Sample_ID	1	2	3	4
2	Sex	F	F	M	F
3	Status	N	Y	Y	N
4	Cat1	3	7	9	3
5	Cat2	0.39980403	0.131167	0.83797435	0.99121998
6	Cat3	0.01765164	0.74075219	0.4243295	0.84309817
7	Cat4	76	75	65	53
8	Cat5	3	10	3	1
9	Cat6	0.27817695	0.07110137	0.37747939	0.13183529
10	Cat7	0.47111146	0.89977883	0.42504482	0.09046157
11	Cat8	59	79	81	65
12	Cat9	0.15067873	0.69674031	0.2607653	0.54603649
13	Cat10	1646	1436	1486	1218
14	Cat11	535	586	543	470
15	Cat12	13	19	15	18
16					

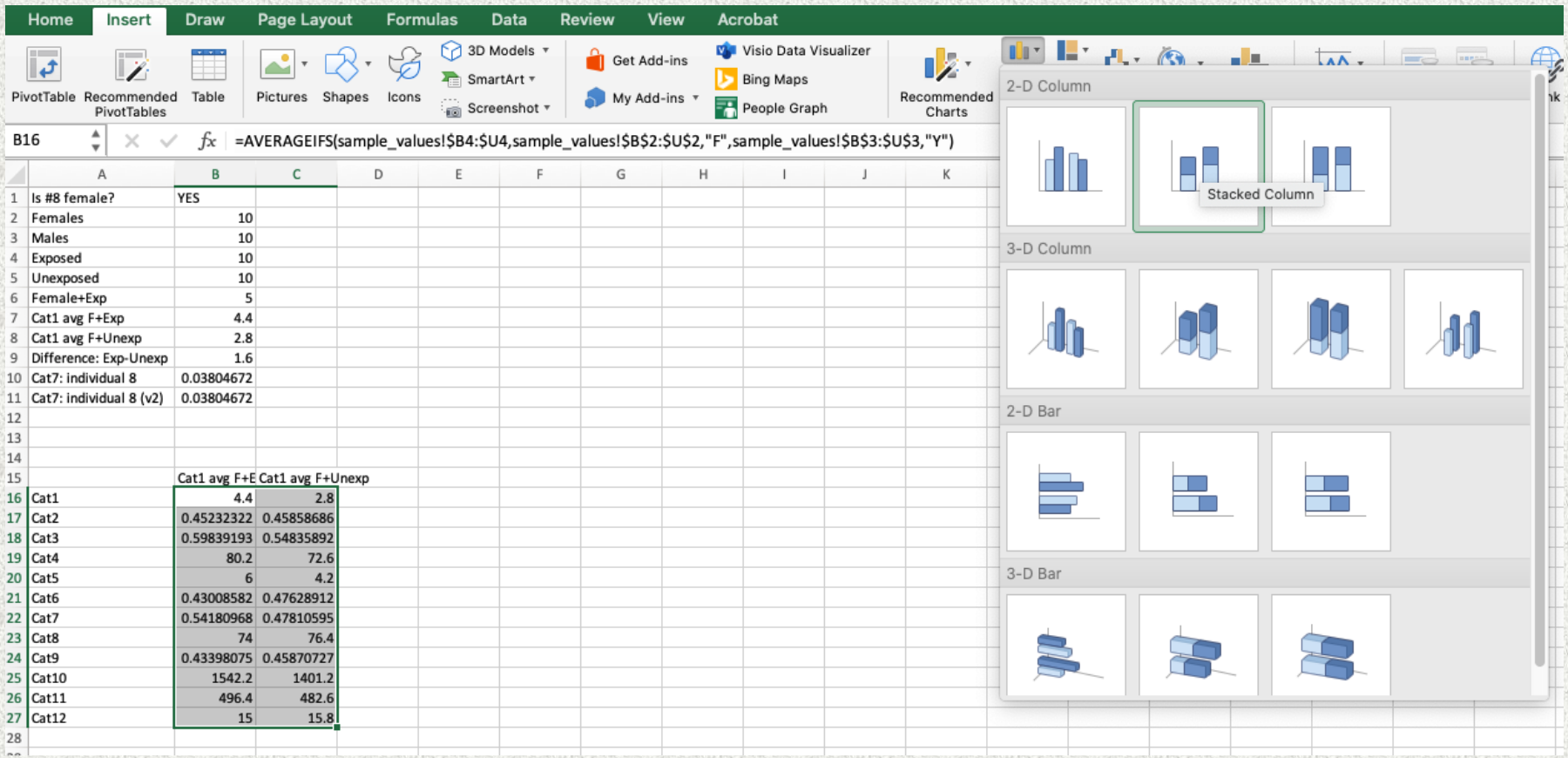
sample_values | Sheet1

	A	B	C
15		Cat avg F+Exp	Cat avg F+Unexp
16	Cat1	4.4	2.8
17	Cat2	0.452323217	0.458586855
18	Cat3	0.598391934	0.548358917
19	Cat4	80.2	72.6
20	Cat5	6	4.2
21	Cat6	0.430085816	0.476289122
22	Cat7	0.541809681	0.478105953
23	Cat8	74	76.4
24	Cat9	0.433980755	0.458707272
25	Cat10	1542.2	1401.2
26	Cat11	496.4	482.6
27	Cat12	15	15.8

sample_values | Sheet1

Stacked bar charts

- Select data → Insert → 2-D Column → Stacked Column

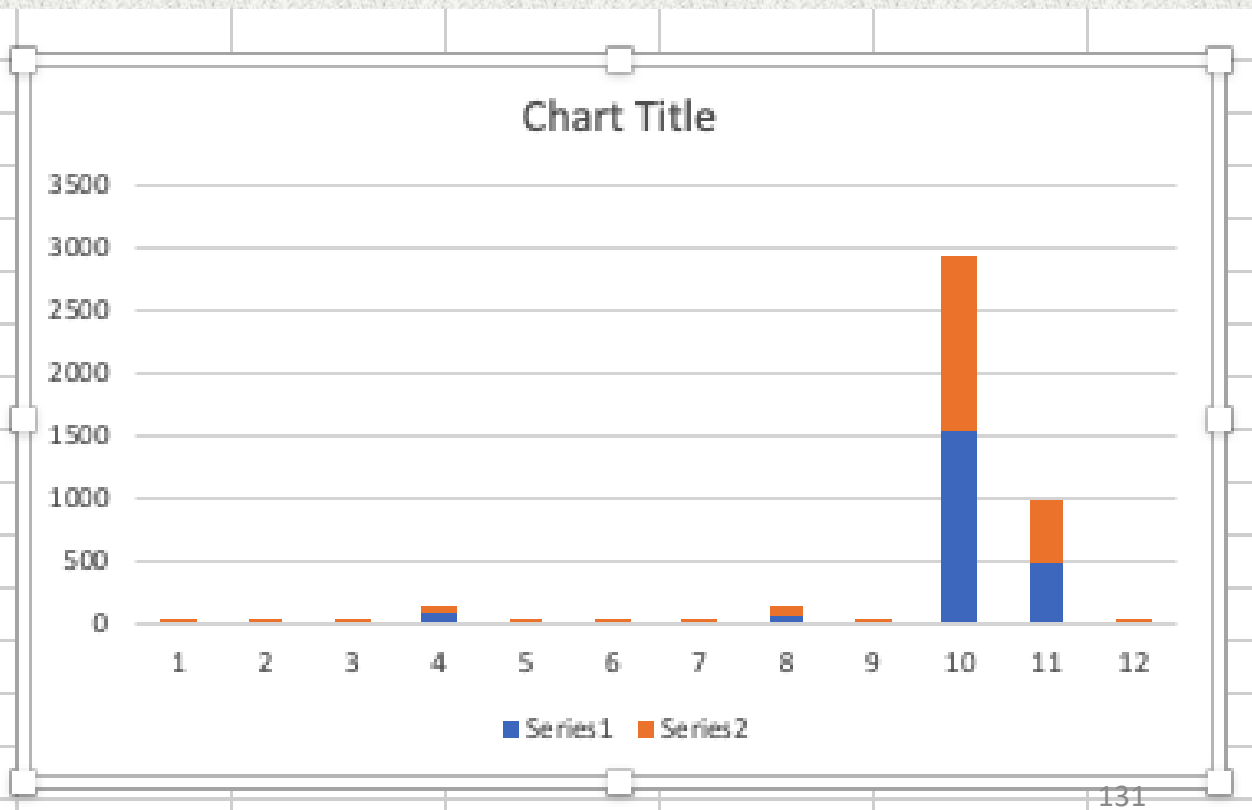


	A	B	C
1	Is #8 female?	YES	
2	Females	10	
3	Males	10	
4	Exposed	10	
5	Unexposed	10	
6	Female+Exp	5	
7	Cat1 avg F+Exp	4.4	
8	Cat1 avg F+Unexp	2.8	
9	Difference: Exp-Unexp	1.6	
10	Cat7: individual 8	0.03804672	
11	Cat7: individual 8 (v2)	0.03804672	
12			
13			
14			
15		Cat1 avg F+E	Cat1 avg F+Unexp
16	Cat1	4.4	2.8
17	Cat2	0.45232322	0.45858686
18	Cat3	0.59839193	0.54835892
19	Cat4	80.2	72.6
20	Cat5	6	4.2
21	Cat6	0.43008582	0.47628912
22	Cat7	0.54180968	0.47810595
23	Cat8	74	76.4
24	Cat9	0.43398075	0.45870727
25	Cat10	1542.2	1401.2
26	Cat11	496.4	482.6
27	Cat12	15	15.8
28			

Stacked bar charts

- The data does not fit well using the same y-axis.
- Can transform the data in some way, or add a secondary axis for 1 category, or…

	Cat1 avg F+E	Cat1 avg F+Unexp
Cat1	4.4	2.8
Cat2	0.45232322	0.45858686
Cat3	0.59839193	0.54835892
Cat4	80.2	72.6
Cat5	6	4.2
Cat6	0.43008582	0.47628912
Cat7	0.54180968	0.47810595
Cat8	74	76.4
Cat9	0.43398075	0.45870727
Cat10	1542.2	1401.2
Cat11	496.4	482.6
Cat12	15	15.8

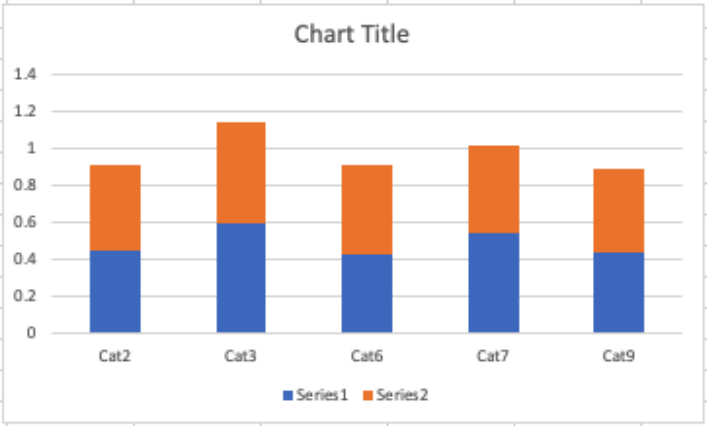


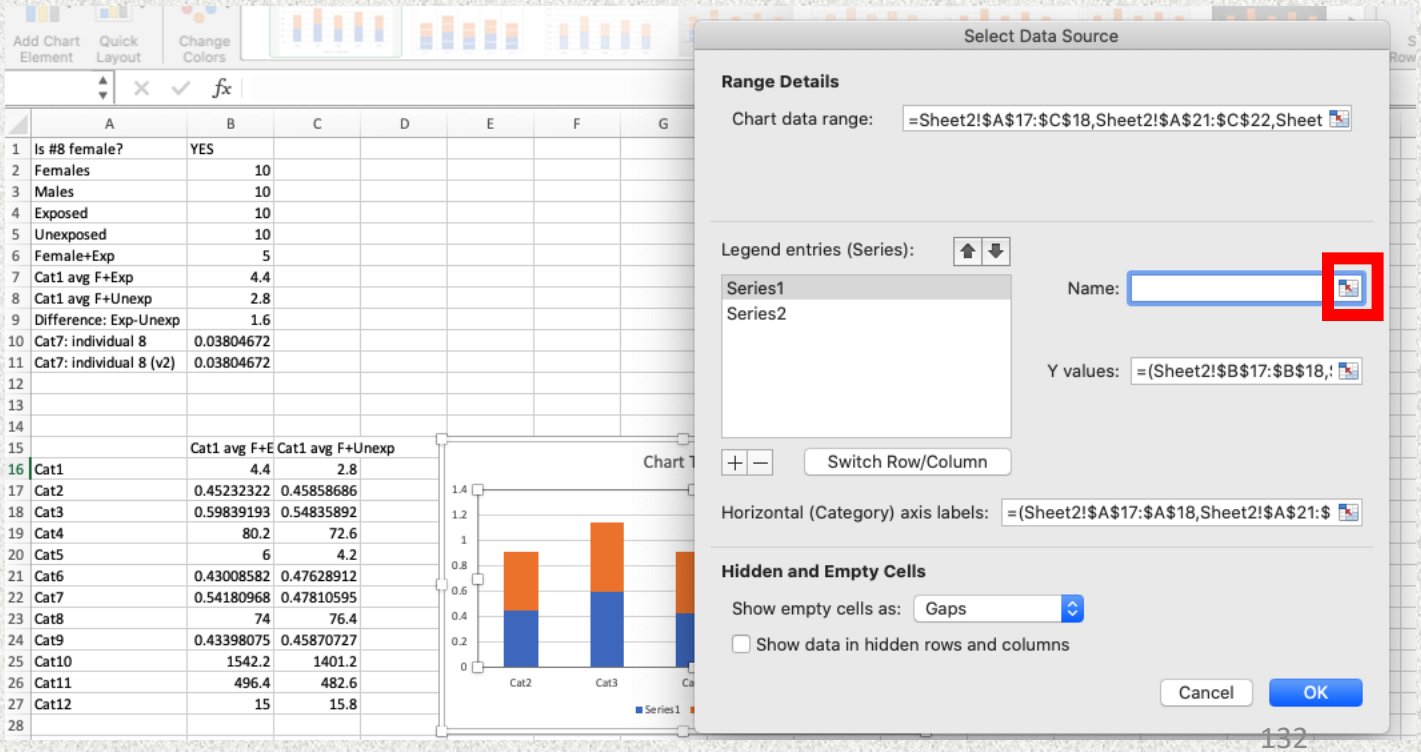
Stacked bar charts

- Much better!

- Next, update the Series labels by clicking in the **Name** area…

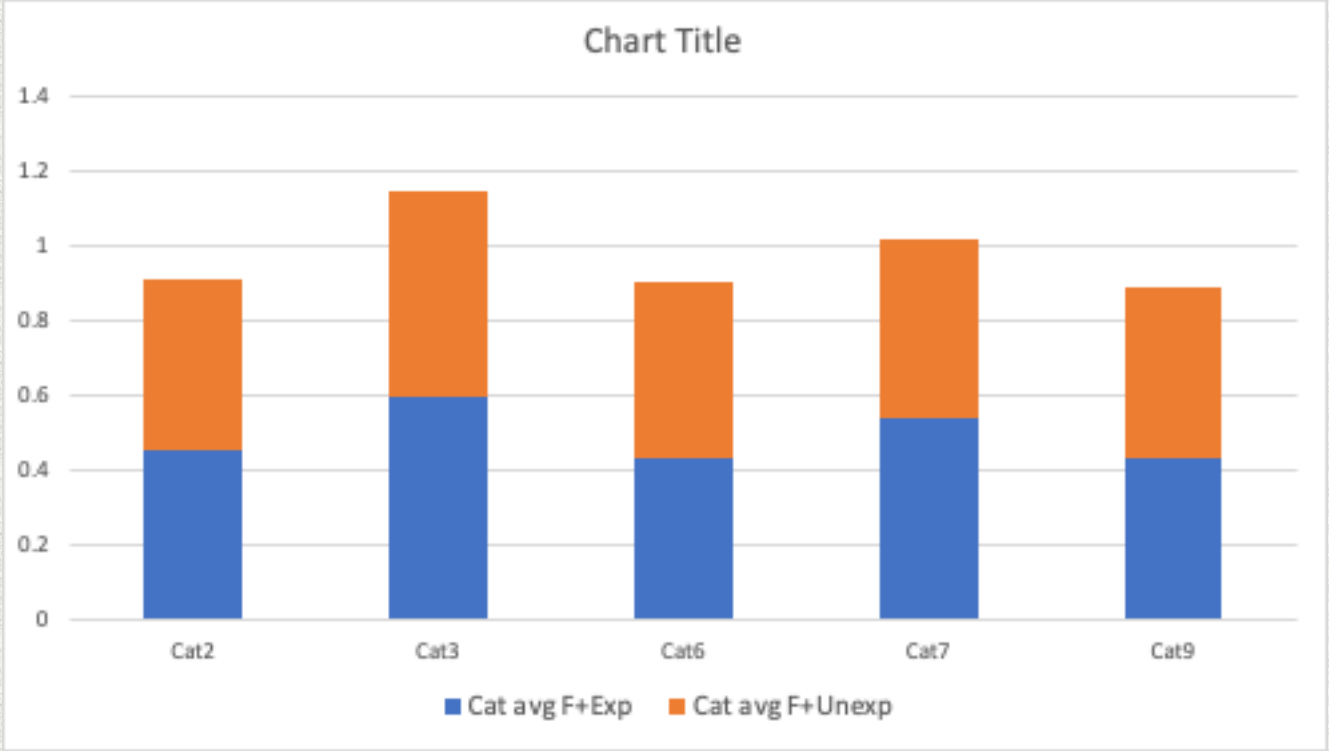
	Cat1 avg F+E	Cat1 avg F+Unexp
Cat1	4.4	2.8
Cat2	0.45232322	0.45858686
Cat3	0.59839193	0.54835892
Cat4	80.2	72.6
Cat5	6	4.2
Cat6	0.43008582	0.47628912
Cat7	0.54180968	0.47810595
Cat8	74	76.4
Cat9	0.43398075	0.45870727
Cat10	1542.2	1401.2
Cat11	496.4	482.6
Cat12	15	15.8





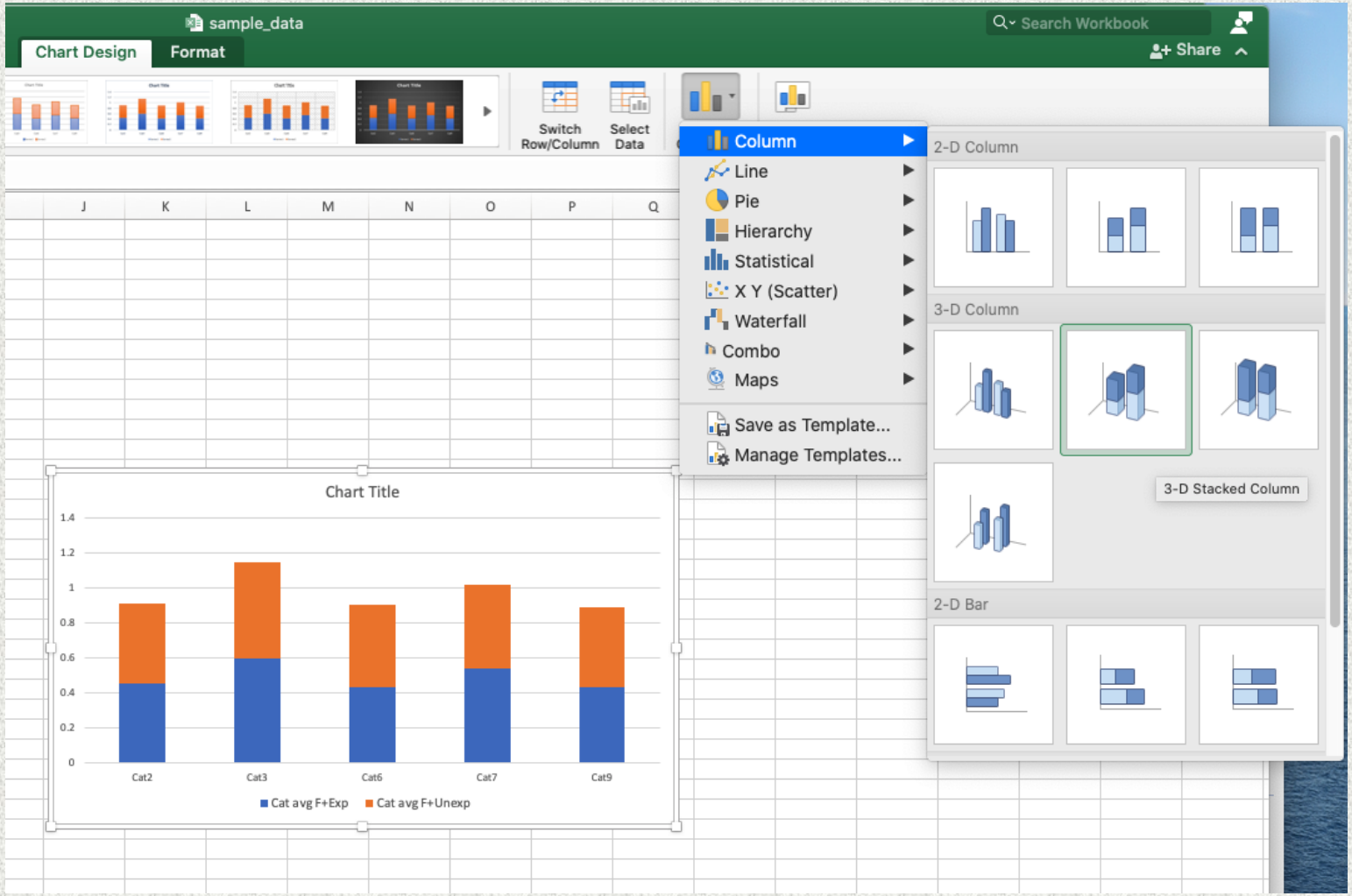
Stacked bar charts

- Now we can show how each category was affected by exposure in females.



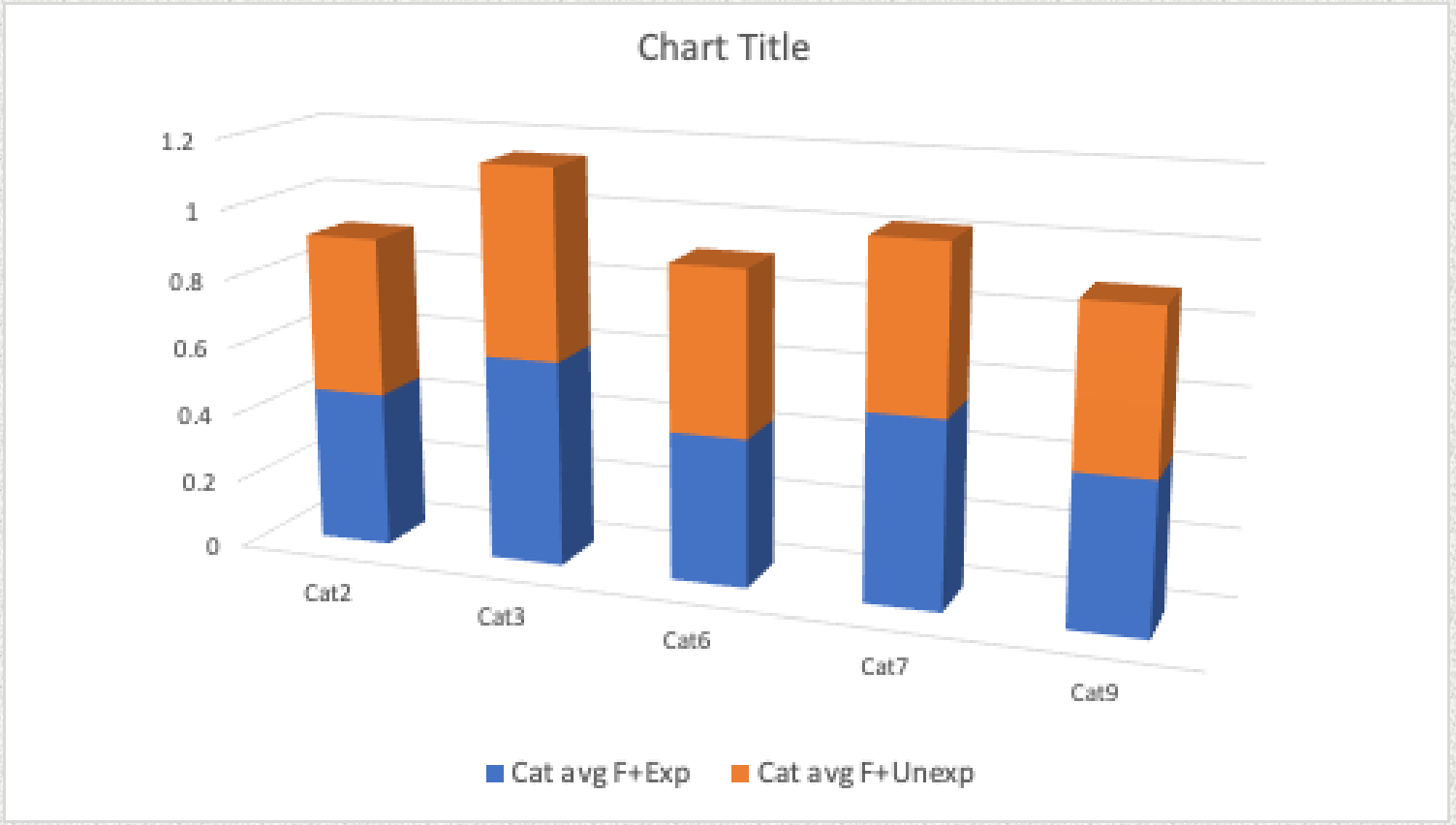
3-D graphs

- Excel offers 3-D options for different types of charts.
- You can start off with 3-D, or you can try it on an existing chart
  - Select the chart
  - **Change Chart Type**
  - Select the 3-D option you want



3-D graphs

- You can change colors, fonts, labels, etc. as before.
- **Can't add a secondary axis.**



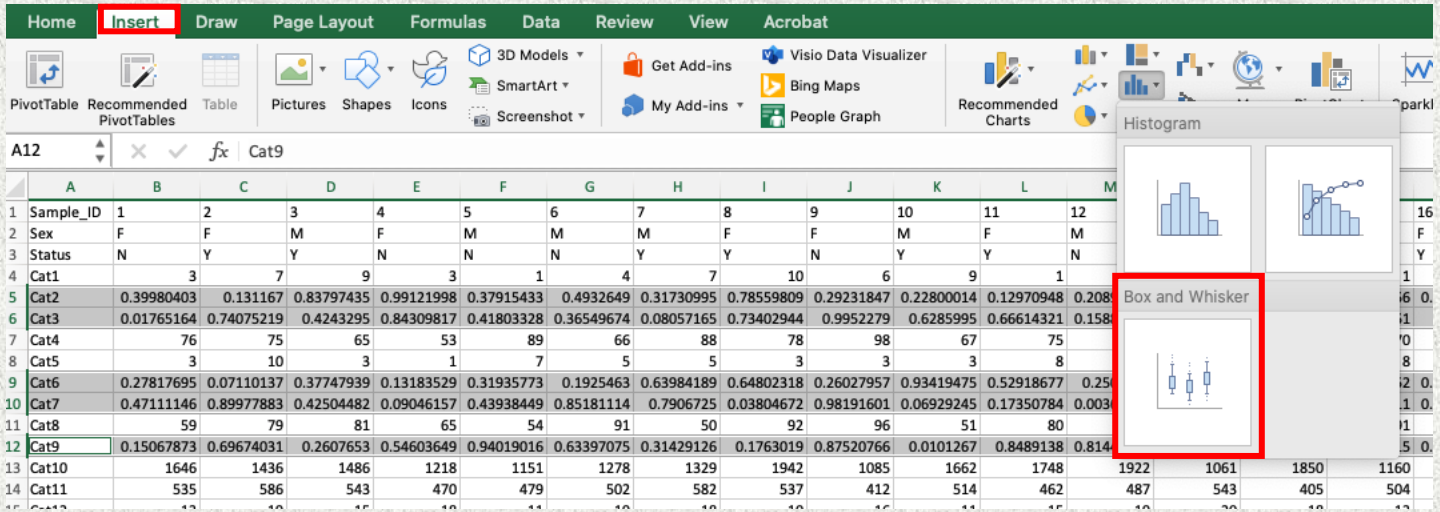
Box and whisker plots

- Select the data you want to plot.
  - Here, I've chosen data that will produce histograms in the same scale range.

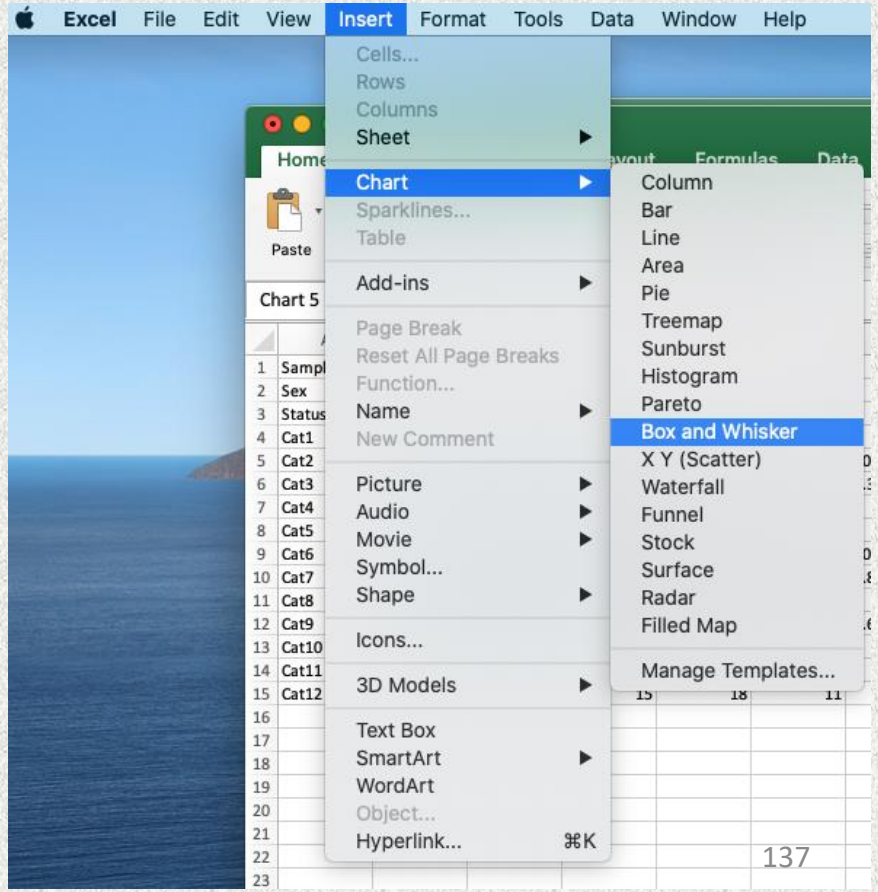
Sample_ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Sex	F	F	M	F	M	M	M	F	F	M	F	M	M	F	M	F	M	F	M	F
Status	N	Y	Y	N	N	N	Y	Y	N	Y	Y	N	Y	N	N	Y	Y	Y	N	N
Cat1	3	7	9	3	1	4	7	10	6	9	1	8	8	1	1	2	5	2	8	1
Cat2	0.39980403	0.131167	0.83797435	0.99121998	0.37915433	0.4932649	0.31730995	0.78559809	0.29231847	0.22800014	0.12970948	0.20895121	0.09570753	0.57328551	0.74719656	0.97775537	0.90570917	0.23738616	0.63079488	0.03630628
Cat3	0.01765164	0.74075219	0.4243295	0.84309817	0.41803328	0.36549674	0.08057165	0.73402944	0.9952279	0.6285995	0.66614321	0.15886006	0.29619534	0.522676	0.49873951	0.499717	0.57727393	0.35131784	0.24743548	0.36314088
Cat4	76	75	65	53	89	66	88	78	98	67	75	96	90	65	70	89	73	84	85	71
Cat5	3	10	3	1	7	5	5	3	3	3	8	6	7	5	8	7	4	2	9	9
Cat6	0.27817695	0.07110137	0.37747939	0.13183529	0.31935773	0.1925463	0.63984189	0.64802318	0.26027957	0.93419475	0.52918677	0.2509355	0.39102529	0.77463581	0.59241452	0.48270423	0.83625265	0.41941355	0.83425559	0.93651799
Cat7	0.47111146	0.89977883	0.42504482	0.09046157	0.43938449	0.85181114	0.7906725	0.03804672	0.98191601	0.06929245	0.17350784	0.00365075	0.62290958	0.10161224	0.129711	0.87430617	0.63898328	0.72340885	0.87721221	0.74542847
Cat8	59	79	81	65	54	91	50	92	96	51	80	63	78	70	91	60	80	59	81	92
Cat9	0.15067873	0.69674031	0.2607653	0.54603649	0.94019016	0.63397075	0.31429126	0.1763019	0.87520766	0.0101267	0.8489138	0.81448124	0.14730753	0.47184386	0.97195415	0.10610593	0.55886697	0.34184183	0.38806705	0.24976963
Cat10	1646	1436	1486	1218	1151	1278	1329	1942	1085	1662	1748	1922	1061	1850	1160	1166	1049	1419	1275	1207
Cat11	535	586	543	470	479	502	582	537	412	514	462	487	543	405	504	478	466	419	467	591
Cat12	13	19	15	18	11	19	18	10	16	11	15	19	20	18	13	17	16	14	14	14

Histograms: Box and Whisker plot

- As before: **Insert** → your chart of choice (Box and Whisker plot here).

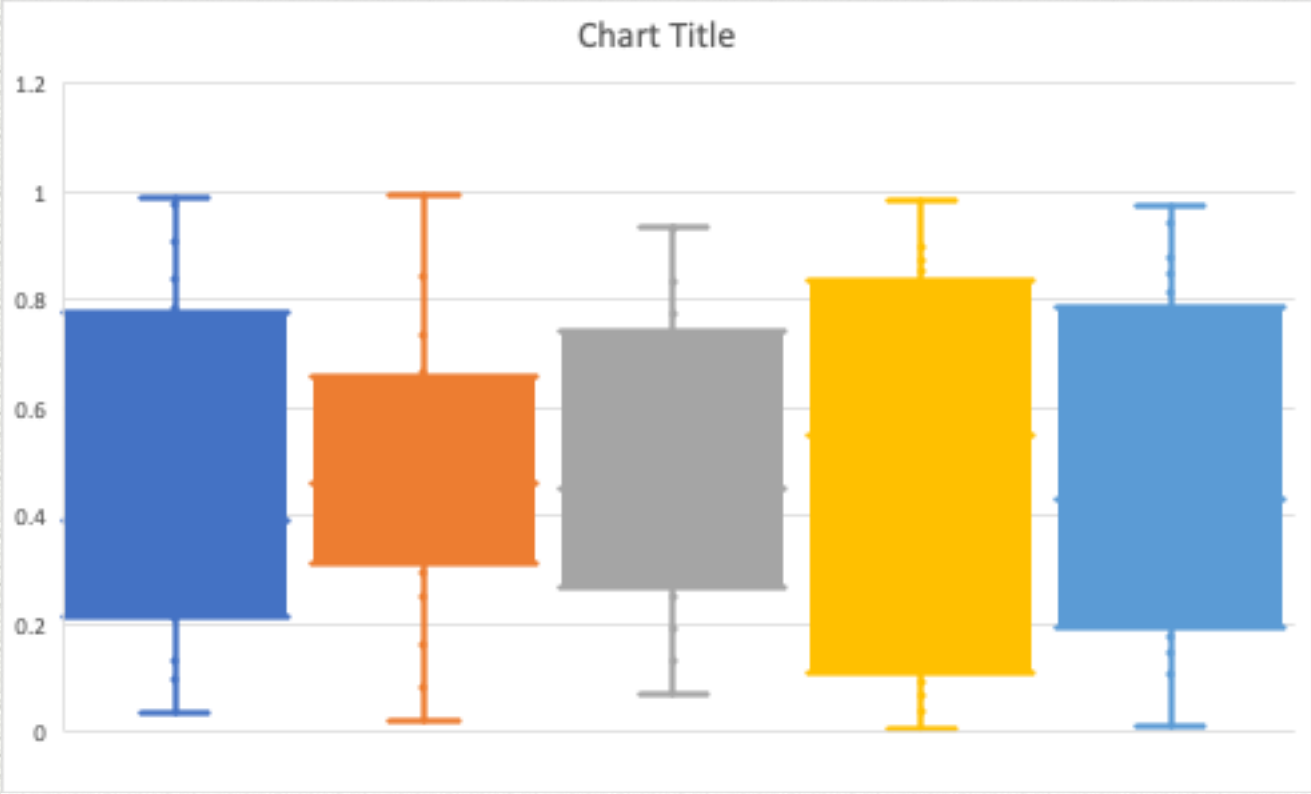


	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Sample_ID	1	2	3	4	5	6	7	8	9	10	11	12
2	Sex	F	F	M	F	M	M	M	F	F	M	F	M
3	Status	N	Y	Y	N	N	N	Y	Y	N	Y	Y	N
4	Cat1	3	7	9	3	1	4	7	10	6	9	1	
5	Cat2	0.39980403	0.131167	0.83797435	0.99121998	0.37915433	0.4932649	0.31730995	0.78559809	0.29231847	0.22800014	0.12970948	0.208
6	Cat3	0.01765164	0.74075219	0.4243295	0.84309817	0.41803328	0.36549674	0.08057165	0.73402944	0.9952279	0.6285995	0.66614321	0.158
7	Cat4	76	75	65	53	89	66	88	78	98	67	75	
8	Cat5	3	10	3	1	7	5	5	3	3	3	8	
9	Cat6	0.27817695	0.07110137	0.37747939	0.13183529	0.31935773	0.1925463	0.63984189	0.64802318	0.26027957	0.93419475	0.52918677	0.25
10	Cat7	0.47111146	0.89977883	0.42504482	0.09046157	0.43938449	0.85181114	0.7906725	0.03804672	0.98191601	0.06929245	0.17350784	0.003
11	Cat8	59	79	81	65	54	91	50	92	96	51	80	
12	Cat9	0.15067873	0.69674031	0.2607653	0.54603649	0.94019016	0.63397075	0.31429126	0.1763019	0.87520766	0.0101267	0.8489138	0.814
13	Cat10	1646	1436	1486	1218	1151	1278	1329	1942	1085	1662	1748	1922
14	Cat11	535	586	543	470	479	502	582	537	412	514	462	487



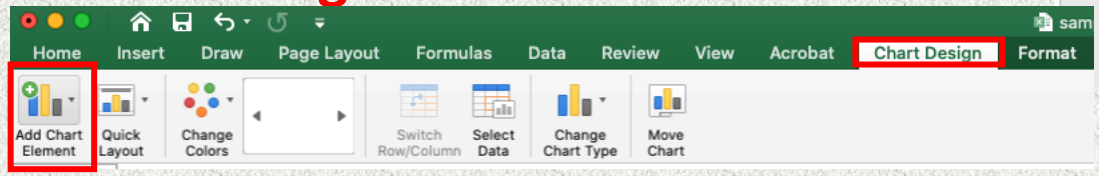
Histograms: Box and Whisker plot

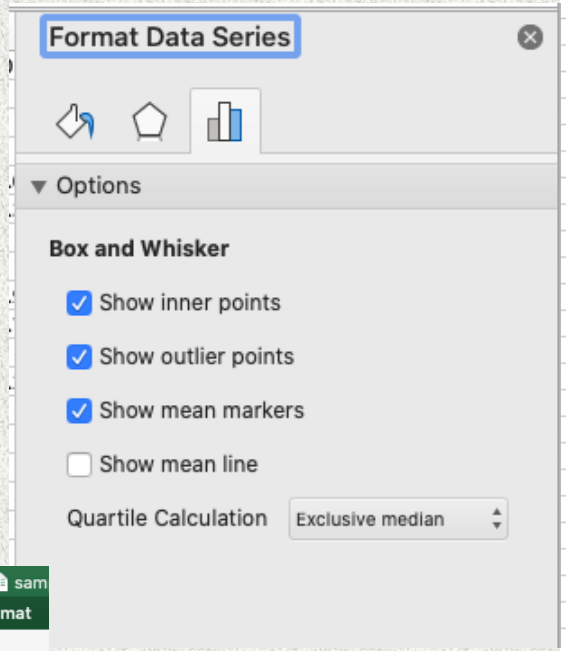
- Caution: Nothing is labeled by default.

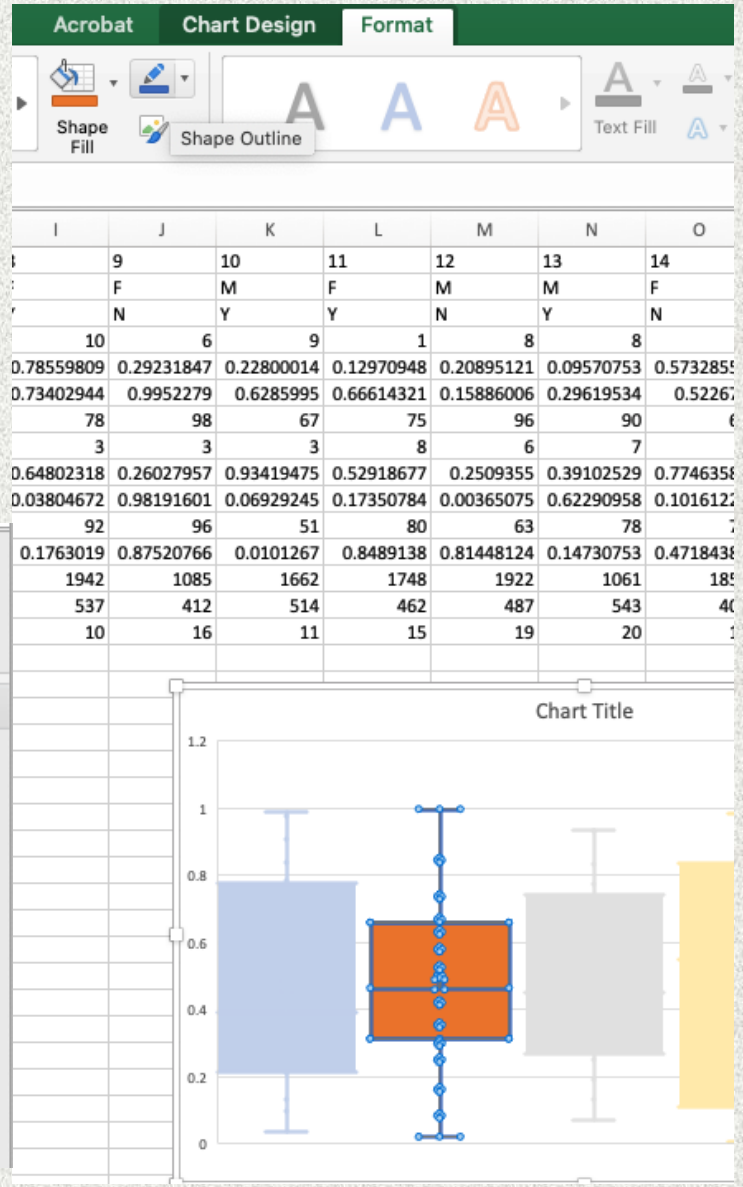


Histograms: Box and Whisker plot

- **Format → Shape Outline** to show individual data points and median line.
- We can change what points are shown by right clicking a plot and choosing **Format Data Series**.
  - We can only edit one plot at a time.
- Add Legends, Labels, Titles, etc. the same way as other charts:
  - **Chart Design** --> **Add Chart Element**







Miscellaneous tips and tricks

If I copy these two cells...

Name Box	A	B
1	Is #8 female?	YES
2	Females	10
3	Males	10
4	Exposed	10
5	Unexposed	10
6	Female+Exp	5
7	Cat1 avg F+Exp	4.4
8	Cat1 avg F+Unexp	2.8
9	Difference: Exp-Unexp	1.6
10	Cat7: individual 8	0.03804672
11	Cat7: individual 8 (v2)	0.03804672

And try to paste them directly below:
ERROR!

	A	B
1	Is #8 female?	YES
2	Females	10
3	Males	10
4	Exposed	10
5	Unexposed	10
6	Female+Exp	5
7	Cat1 avg F+Exp	4.4
8	Cat1 avg F+Unexp	2.8
9	Difference: Exp-Unexp	1.6
10	Cat7: individual 8	0.03804672
11	Cat7: individual 8 (v2)	0.03804672
12		
13		
14		
15		
16	Cat1 avg F+Exp	#DIV/0!
17	Cat1 avg F+Unexp	#DIV/0!
18		

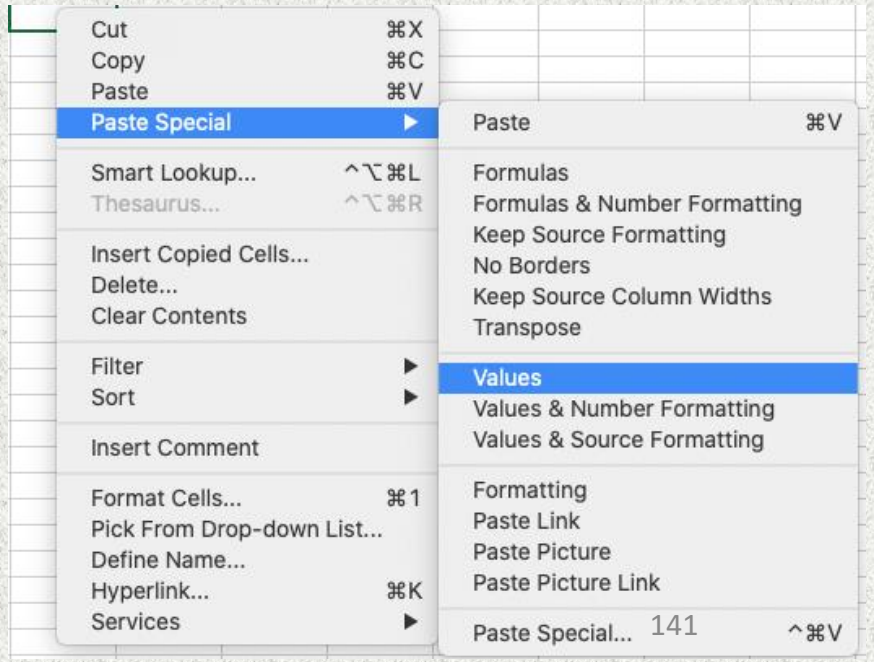
The “addresses” in the formulas of the cells update automatically.

Excel will either give an error message (if it can’t calculate the formula because some value is missing), or you’ll get a bogus result.

Miscellaneous tips and tricks

Solution: add $ before any part you don't want to change.

Another solution: if you don't need Excel to recalculate anything (for example, you're just copying and pasting some data in another place):



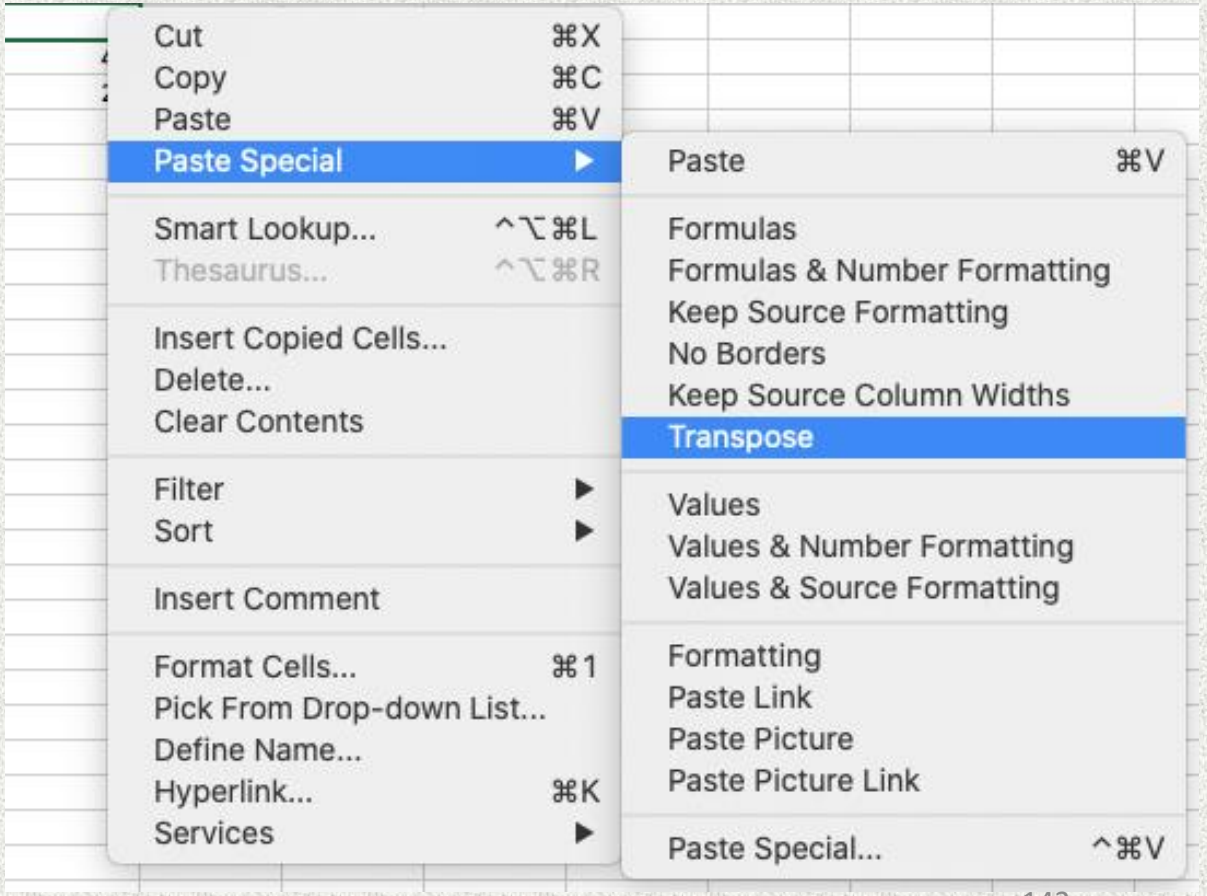
Copy the data, then right click where you want to paste it, and choose **Paste Special → Values**

Miscellaneous tips and tricks

To change the direction of some list: **copy** the list (NOT cut)…

Then **Paste Special** --> **Transpose**

	A	
1	Sample_ID	1
2	Sex	F
3	Status	N
4	Cat1	
5	Cat2	0.39
6	Cat3	0.01
7	Cat4	
8	Cat5	
9	Cat6	0.27
10	Cat7	0.47
11	Cat8	
12	Cat9	0.15
13	Cat10	
14	Cat11	
15	Cat12	
16		

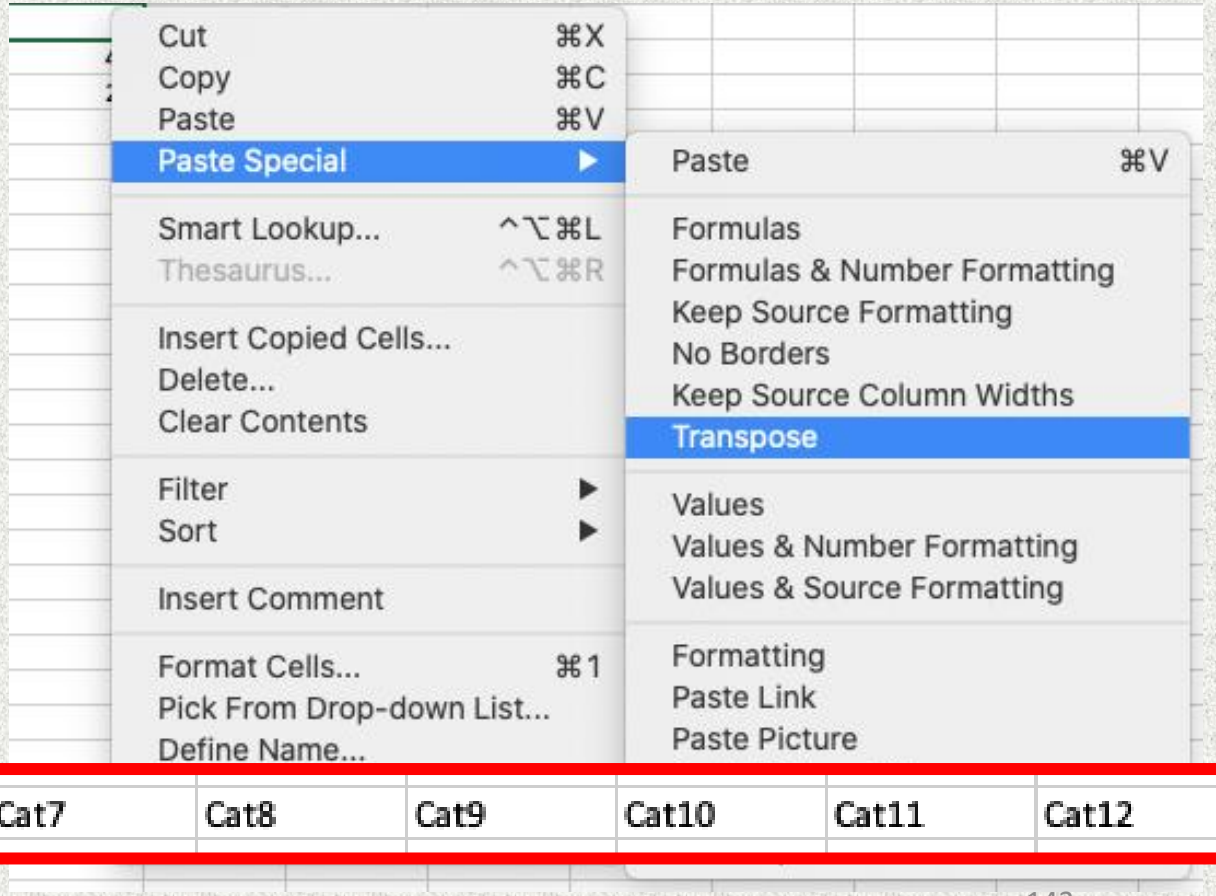


Miscellaneous tips and tricks

To change the direction of some list: **copy** the list (NOT cut)…

Then **Paste Special** --> **Transpose**

	A	
1	Sample_ID	1
2	Sex	F
3	Status	N
4	Cat1	
5	Cat2	0.39
6	Cat3	0.01
7	Cat4	
8	Cat5	
9	Cat6	0.27
10	Cat7	0.47
11	Cat8	
12	Cat9	0.15
13	Cat10	
14	Cat11	
15	Cat12	
16		

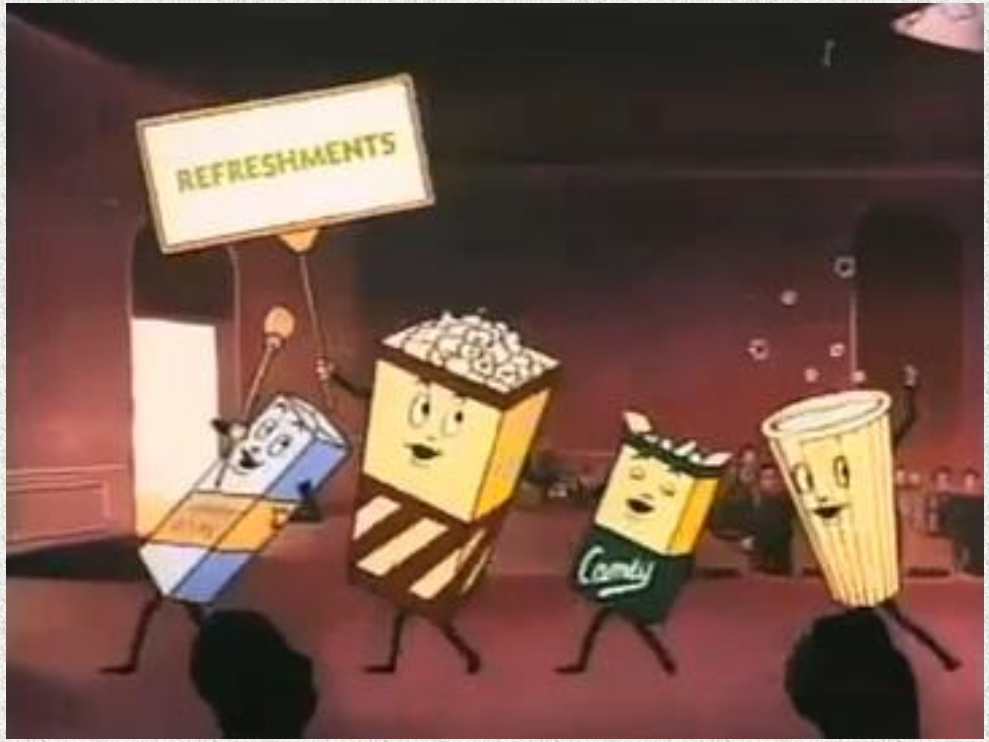


Cat1	Cat2	Cat3	Cat4	Cat5	Cat6	Cat7	Cat8	Cat9	Cat10	Cat11	Cat12

Miscellaneous tips and tricks…

- Excel can also calculate the minimum, maximum, and average values, along with standard deviation, for a data range: =MIN(data_range); =MAX(data_range); =AVERAGE(data_range); =STDEV(data_range)

Next up: Analyzing Big(ish) Data in Excel



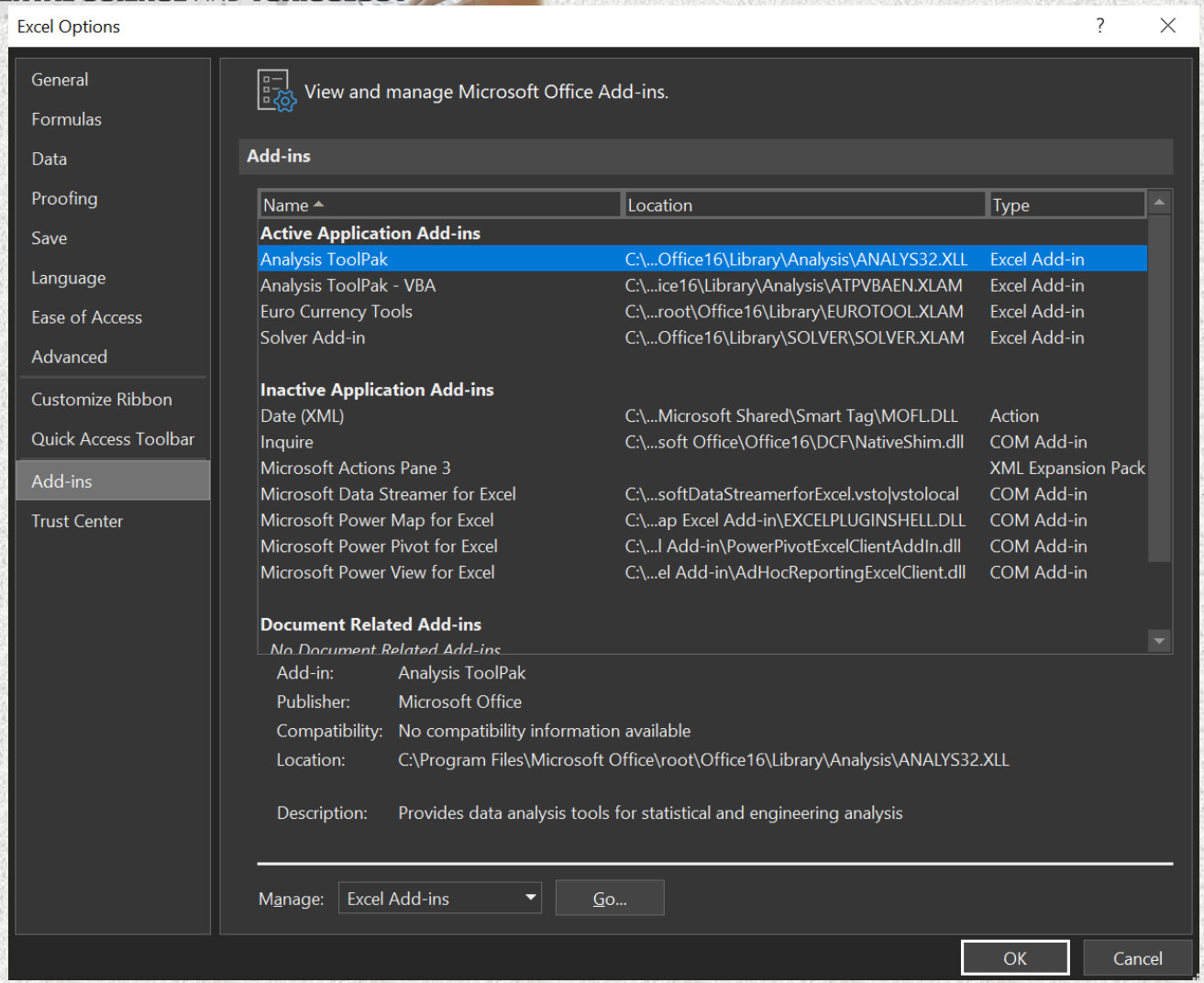
Analyzing Big(ish) Data in Excel

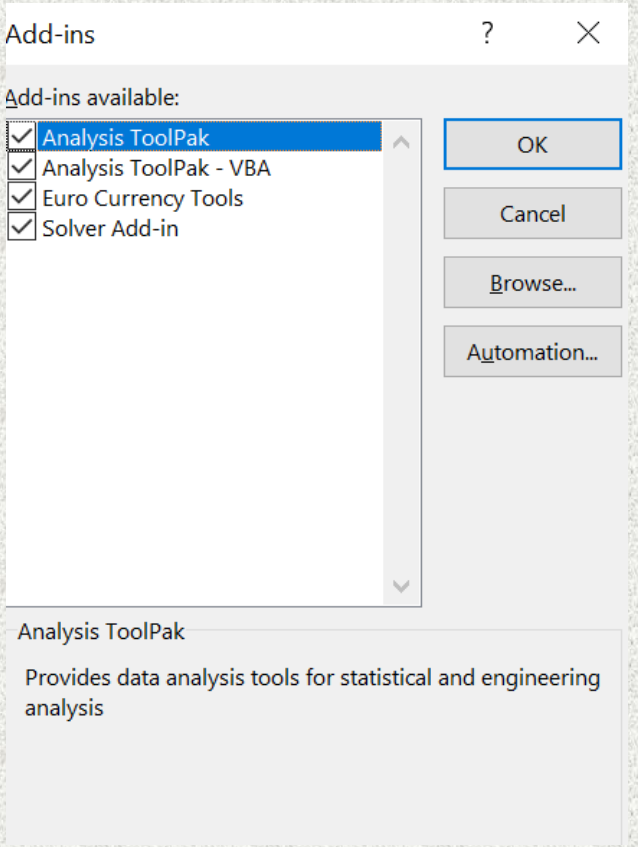
Dillon Lloyd

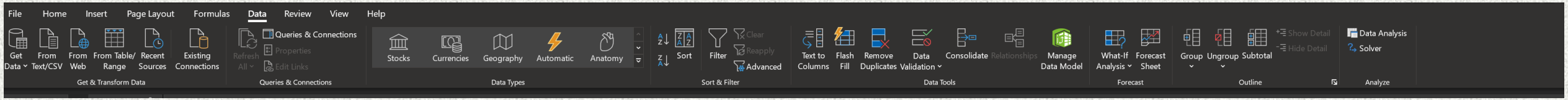
North Carolina State University

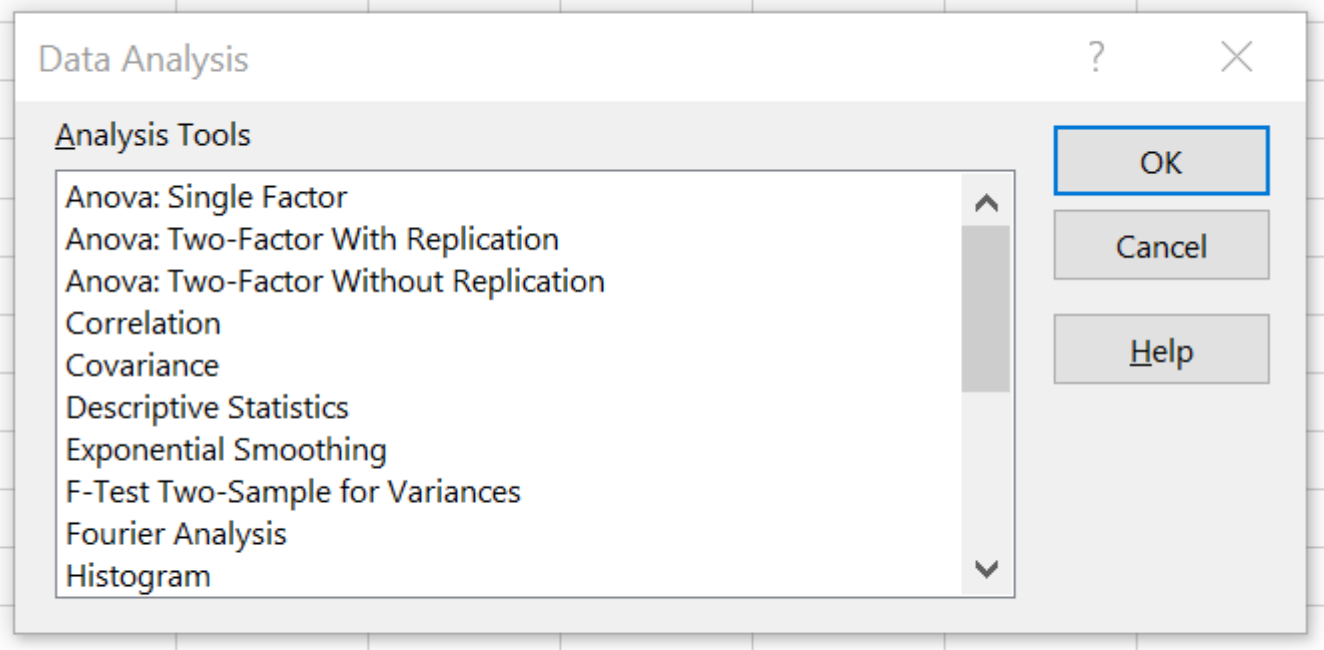
Analysis ToolPak Installation

- The Analysis ToolPak is an extension of Excel that can be used to perform common statistical tests quickly and easily
- Instructions on installation for both Windows and Mac can be found at the link below
  - https://support.microsoft.com/en-us/office/load-the-analysis-toolpak-in-excel-6a63e598-cd6d-42e3-9317-6b40ba1a66b4


Excel Options
General
Formulas
Data
Proofing
Save
Language
Ease of Access
Advanced
Customize Ribbon
Quick Access Toolbar
Add-ins
Trust Center
View and manage Microsoft Office Add-ins.
Add-ins
Name
Location
Type
Active Application Add-ins
Analysis ToolPak C:\...Office16\Library\Analysis\ANALYS32.XLL Excel Add-in
Analysis ToolPak - VBA C:\...ice16\Library\Analysis\ATPVBAEN.XLAM Excel Add-in
Euro Currency Tools C:\...root\Office16\Library\EUROTOOL.XLAM Excel Add-in
Solver Add-in C:\...Office16\Library\SOLVER\SOLVER.XLAM Excel Add-in
Inactive Application Add-ins
Date (XML) C:\...Microsoft Shared\Smart Tag\MOFL.DLL Action
Inquire C:\...soft Office\Office16\DCF\NativeShim.dll COM Add-in
Microsoft Actions Pane 3 XML Expansion Pack
Microsoft Data Streamer for Excel C:\...softDataStreamerforExcel.vsto|vstolocal COM Add-in
Microsoft Power Map for Excel C:\...ap Excel Add-in\EXCELPLUGINSHELL.DLL COM Add-in
Microsoft Power Pivot for Excel C:\...l Add-in\PowerPivotExcelClientAddIn.dll COM Add-in
Microsoft Power View for Excel C:\...el Add-in\AdHocReportingExcelClient.dll COM Add-in
Document Related Add-ins
No Document Related Add-ins
Add-in: Analysis ToolPak
Publisher: Microsoft Office
Compatibility: No compatibility information available
Location: C:\Program Files\Microsoft Office\root\Office16\Library\Analysis\ANALYS32.XLL
Description: Provides data analysis tools for statistical and engineering analysis
Manage: Excel Add-ins
Go...
OK
Cancel


Add-ins
Add-ins available:
Analysis ToolPak
Analysis ToolPak - VBA
Euro Currency Tools
Solver Add-in
OK
Cancel
Browse...
Automation...
Analysis ToolPak
Provides data analysis tools for statistical and engineering analysis

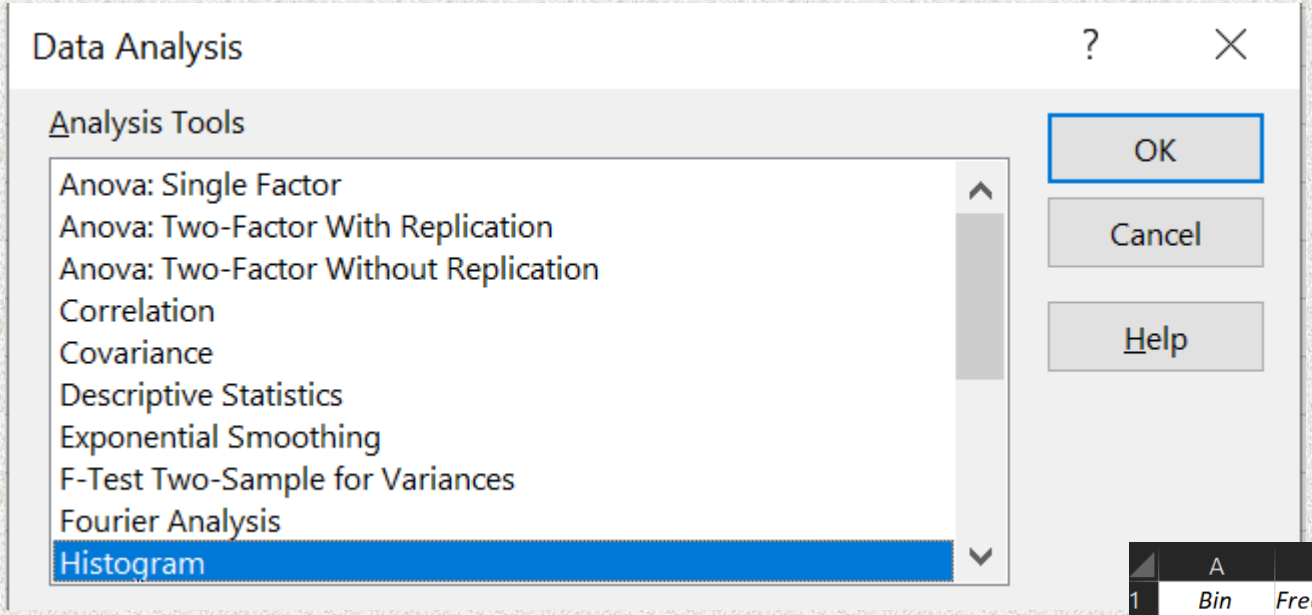

File
Home
Insert
Page Layout
Formulas
Data
Review
View
Help
Get Data
From Text/CSV
From Web
From Table/ Range
Recent Sources
Existing Connections
Get & Transform Data
Refresh All
Queries & Connections
Properties
Edit Links
Queries & Connections
Stocks
Currencies
Geography
Automatic
Anatomy
Data Types
Sort
Filter
Clear
Reapply
Advanced
Sort & Filter
Text to Columns
Flash Fill
Remove Duplicates
Data Validation
Consolidate
Relationships
Manage Data Model
Data Tools
What-If Analysis
Forecast Sheet
Forecast
Group
Ungroup
Subtotal
Show Detail
Hide Detail
Outline
Data Analysis
Solver
Analyze

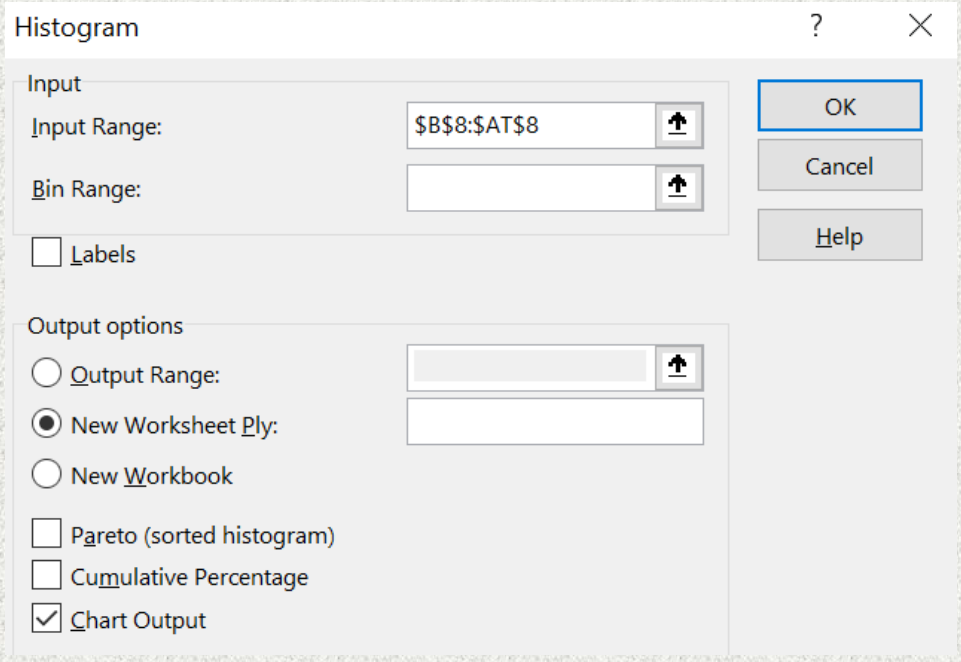

Data Analysis
Analysis Tools
Anova: Single Factor
Anova: Two-Factor With Replication
Anova: Two-Factor Without Replication
Correlation
Covariance
Descriptive Statistics
Exponential Smoothing
F-Test Two-Sample for Variances
Fourier Analysis
Histogram
OK
Cancel
Help

Histogram

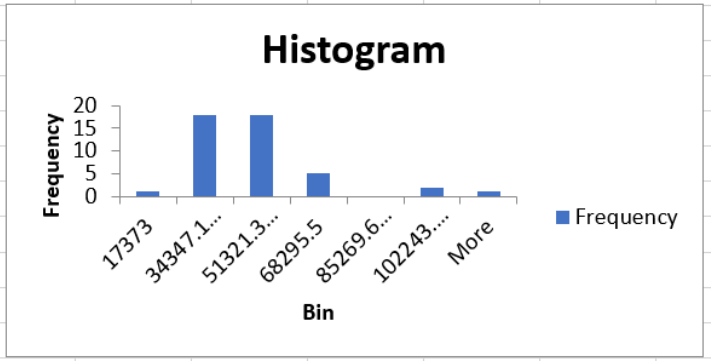
- In the Data Ribbon click data analysis
- Scroll down and click on “Histogram”
- Select the data you want to see the distribution of
- Check “Chart Output” to get a plot in the output
- Click Ok and the histogram will show up in a new sheet

Histogram





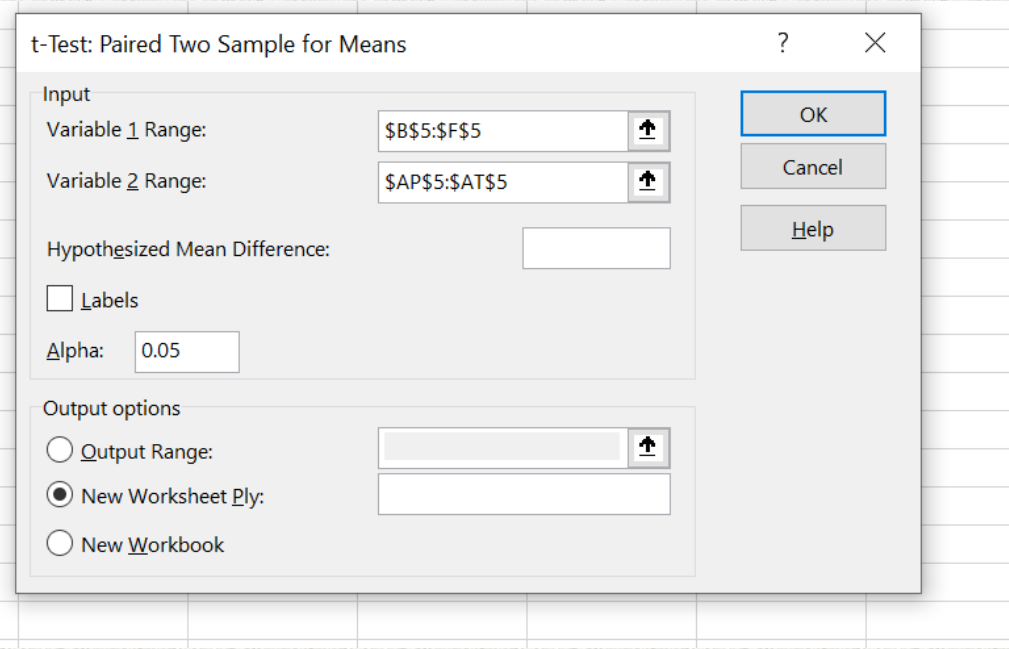
	A	B	C	D	E	F	G	H	I	J
1	*Bin*	*Frequency*								
2	17373	1								
3	34347.17	18								
4	51321.33	18								
5	68295.5	5								
6	85269.67	0								
7	102243.8	2								
8	More	1								
9										
10										
11										
12										
13										



T-Test

- Filter Data to the dose row and the row of your gene of interest
- Click Data Analysis on the data tab
- Click T-Test: Paired Two Sample for Means
- Under Variable 1 Range, select the columns in the row of your gene that correspond to the control dose
- Under Variable 2 Range, selects the columns in the row of your gene that corresponds to the max dose
- Click ok to view the results

T-Test



Control vs. Max Dose T-Test Results

	A	B	C
1	t-Test: Paired Two Sample for Means		
2			
3		*Variable 1*	*Variable 2*
4	Mean	18	30
5	Variance	32.5	45.5
6	Observations	5	5
7	Pearson Correlation	0.461584247	
8	Hypothesized Mean Difference	0	
9	df	4	
10	t Stat	-4.115966043	
11	P(T<=t) one-tail	0.007330294	
12	t Critical one-tail	2.131846786	
13	P(T<=t) two-tail	0.014660589	
14	t Critical two-tail	2.776445105	
15			

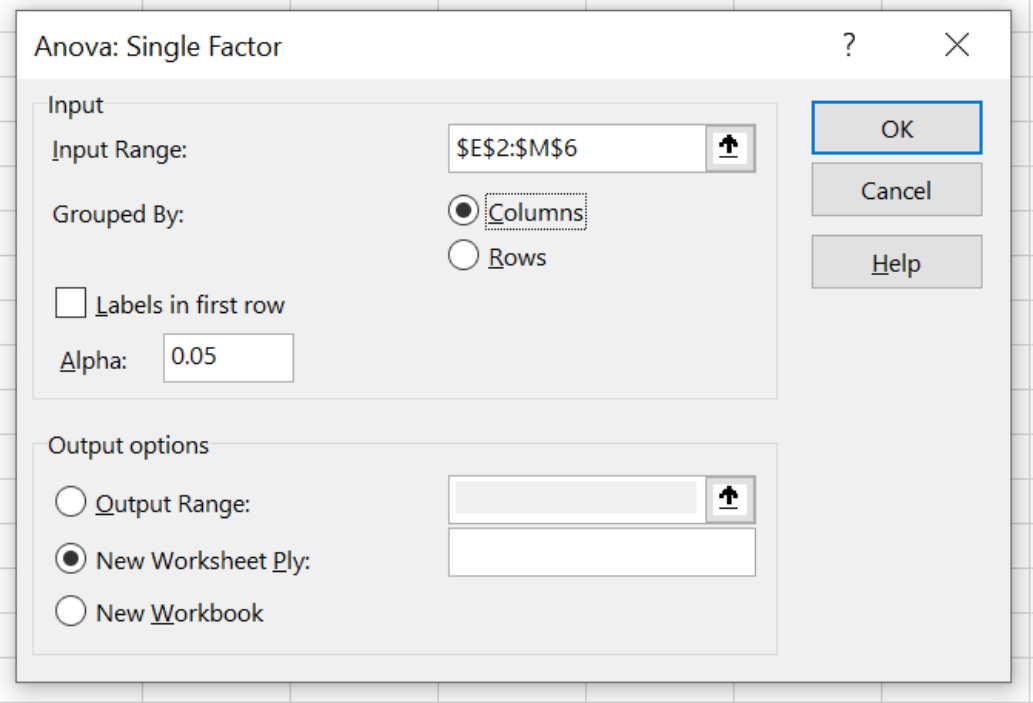
ANOVA

- Filter Data to the dose row and the row of your gene of interest
- Copy and Transpose Data and create columns for each different dose
  - ANOVA requires a bit more data manipulation due to the number of groups
  - Transpose data by highlighting the Dose and Gene of Interest Row, pressing ctrl + shift + right arrow key, and copying that highlighted data
  - Then go to Home → Paste → Paste Transpose in a new worksheet
- Click Data Analysis on the data tab
- Click on ANOVA
- Under Input Range, select the dose columns created in the previous step
- Group by columns and click ok to get the ANOVA output

ANOVA

CLASS:DOSE	ENSMUSG00000000028	Dose0	Dose0.01	Dose0.03	Dose0.1	Dose0.3	Dose1	Dose3	Dose10	Dose30
0	21	21	39	10	13	17	13	17	41	28
0	10	10	10	38	11	17	38	66	24	26
0	22	22	21	10	9	15	13	22	13	42
0	23	23	13	25	23	10	15	21	34	27
0	14	14	115	15	68	24	18	44	49	27
0.01	39									
0.01	10									
0.01	21									
0.01	13									
0.01	115									
0.03	10									
0.03	38									
0.03	10									
0.03	25									
0.03	15									
0.1	13									
0.1	11									
0.1	9									
0.1	23									
0.1	68									
0.3	17									
0.3	17									
0.3	15									
0.3	10									
0.3	24									
1	13									
1	38									
1	13									
1	15									
1	18									
3	17									
3	66									
3	22									
3	21									
3	44									
10	41									
10	24									
10	13									
10	34									
10	49									
30	28									
30	26									
30	42									
30	27									
30	27									

ANOVA

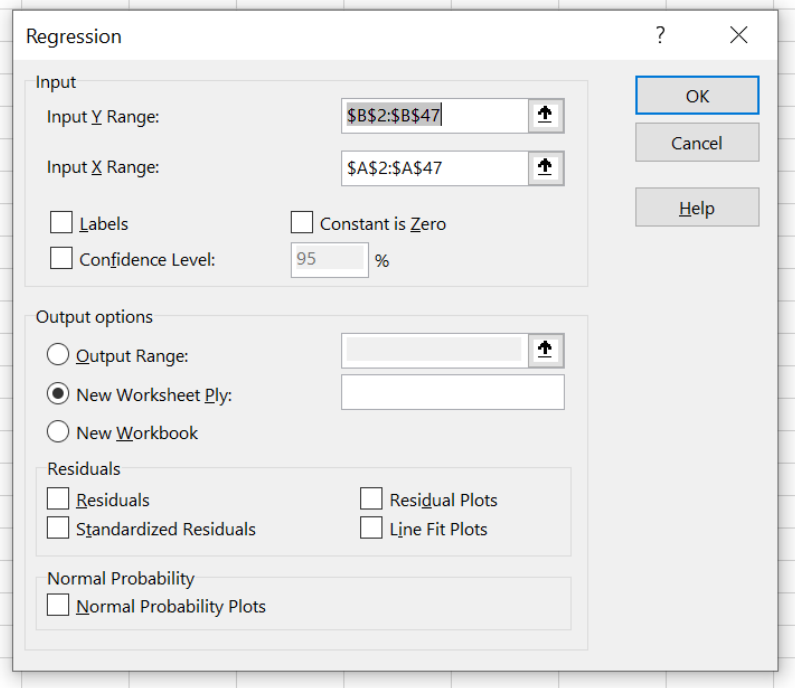


	A	B	C	D	E	F	G	H
1	Anova: Single Factor							
2								
3	SUMMARY							
4	*Groups*	*Count*	*Sum*	*Average*	*Variance*			
5	Column 1	5	90	18	32.5			
6	Column 2	5	198	39.6	1903.8			
7	Column 3	5	98	19.6	143.3			
8	Column 4	5	124	24.8	612.2			
9	Column 5	5	83	16.6	25.3			
10	Column 6	5	97	19.4	112.3			
11	Column 7	5	170	34	431.5			
12	Column 8	5	161	32.2	199.7			
13	Column 9	5	150	30	45.5			
14								
15								
16	ANOVA							
17	*Source of Variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*	
18	Between Groups	2708.578	8	338.5722	0.8691	0.550983	2.208518	
19	Within Groups	14024.4	36	389.5667				
20								
21	Total	16732.98	44					
22								
23								
24								

Regression

- Filter the data to dose and gene of interest
- Copy the dose and gene row by highlighting the first two cells and clicking ctrl+shift+right arrow
- Create a new worksheet and transpose the data In the data analysis tab select regression
- Choose the Y variable as the gene expression value, and the X variable as the dose
- Do this by highlighting the numeric value under the header row and clicking ctrl+shift+down arrow to select just the numeric cells
- Click ok to review results

Regression

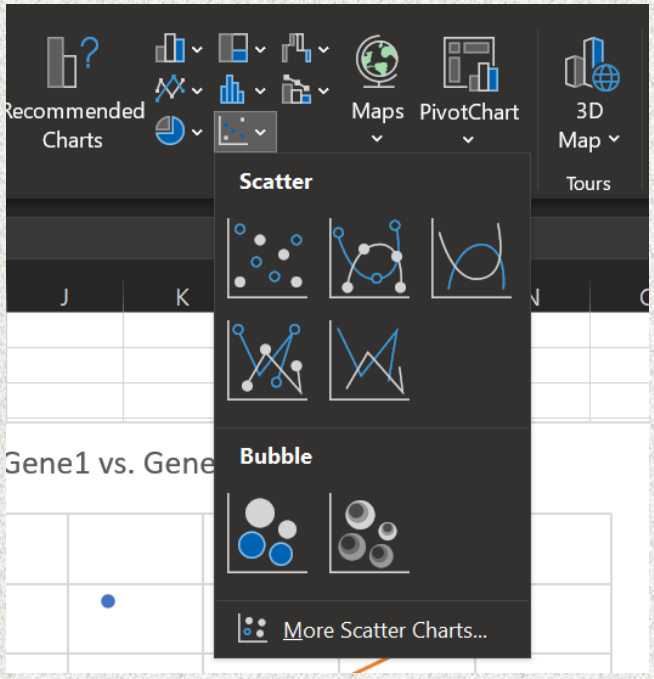


	A	B	C	D	E	F	G	H	I	J
1	SUMMARY OUTPUT									
2										
3	*Regression Statistics*									
4	Multiple R	0.111153								
5	R Square	0.012355								
6	Adjusted F	-0.01009								
7	Standard I	19.48447								
8	Observati	46								
9										
10	ANOVA									
11		*df*	*SS*	*MS*	*F*	*ignificance F*				
12	Regressior	1	208.9637	208.9637	0.550419	0.462089				
13	Residual	44	16704.36	379.6446						
14	Total	45	16913.33							
15										
16		*Coefficients*	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*	
17	Intercept	25.21011	3.237387	7.787178	8.17E-10	18.68558	31.73463	18.68558	31.73463	
18	X Variable	0.22922	0.308963	0.741902	0.462089	-0.39345	0.851894	-0.39345	0.851894	
19										
20										
21										

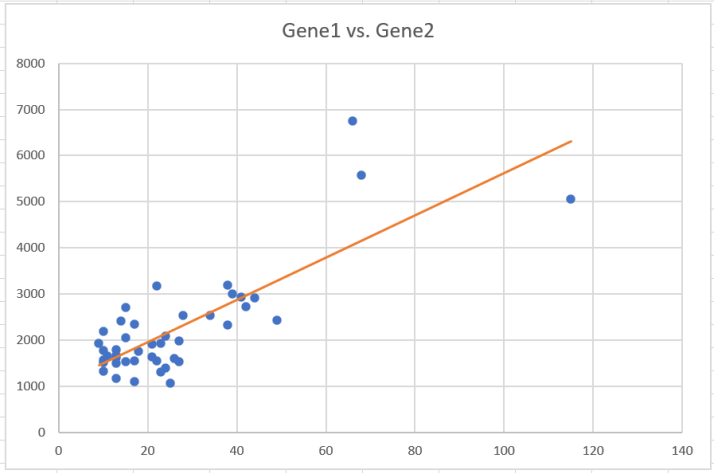
Scatter Plot

- Select Two Genes of interest and the Dose column through the filtering step and transpose them into a new worksheet
- Select the two gene columns
- Click Insert Scatter
- The Scatter Plot will then appear
- The add a trend line, right click on a point and click add trend line
- The trend line can then be edited to make it more clear
- These steps can be repeated to make a scatter plot of Dose vs. Gene Expression to visualize the regression analysis

Scatter Plot



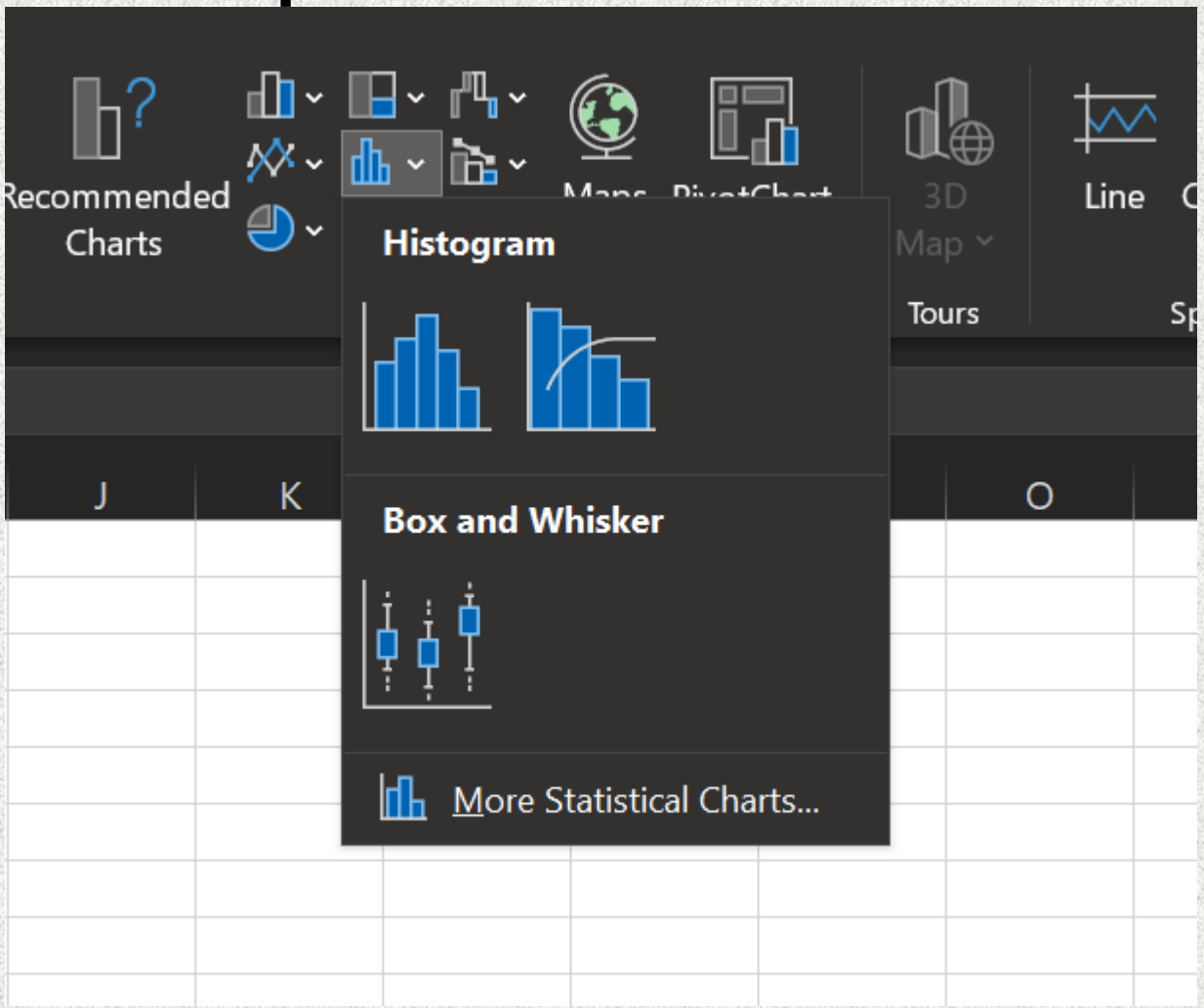
	A	B	C
1	CLASS:DOSE	ENSMUSG00000000028	ENSMUSG00000000088
2	0	21	1911
3	0	10	1770
4	0	22	1547
5	0	23	1931
6	0	14	2413
7	0.01	39	2998
8	0.01	10	2190
9	0.01	21	1912
10	0.01	13	1614
11	0.01	115	5048
12	0.03	10	1512
13	0.03	38	2321
14	0.03	10	1326
15	0.03	25	1063
16	0.03	15	1532
17	0.1	13	1164
18	0.1	11	1651
19	0.1	9	1933
20	0.1	23	1307
21	0.1	68	5580
22	0.3	17	1549
23	0.3	17	2341
24	0.3	15	2714
25	0.3	10	1569
26	0.3	24	2079
27	1	13	1503
28	1	38	3192
29	1	13	1784
30	1	15	2048
31	1	18	1752
32	3	17	1106
33	3	66	6746
34	3	22	3177
35	3	21	1629
36	3	44	2907
37	10	41	2924
38	10	24	1385
39	10	13	1669
40	10	34	2530
41	10	49	2434
42	30	28	2536
43	30	26	1592
44	30	42	2717
45	30	27	1986
46	30	27	1539
47			
48			



Boxplot

- Filter to Dose column or row and gene of interest
- Highlight gene column
- Click Insert → Statistics Chart → Box and Whisker Plot
- Right click and select Select Data
- Under series, click edit and select the data at Dose 1, then do the same for Dose 2, Dose 3 etc.
- Under Chart Design click Add Chart Element and add a legend to the plot to distinguish between doses
- Can then right click on the plot to edit other aesthetic properties

Boxplot


Recommended Charts
Maps
PivotChart
3D Map
Tours
Line
Histogram
Box and Whisker
More Statistical Charts...
J
K
O

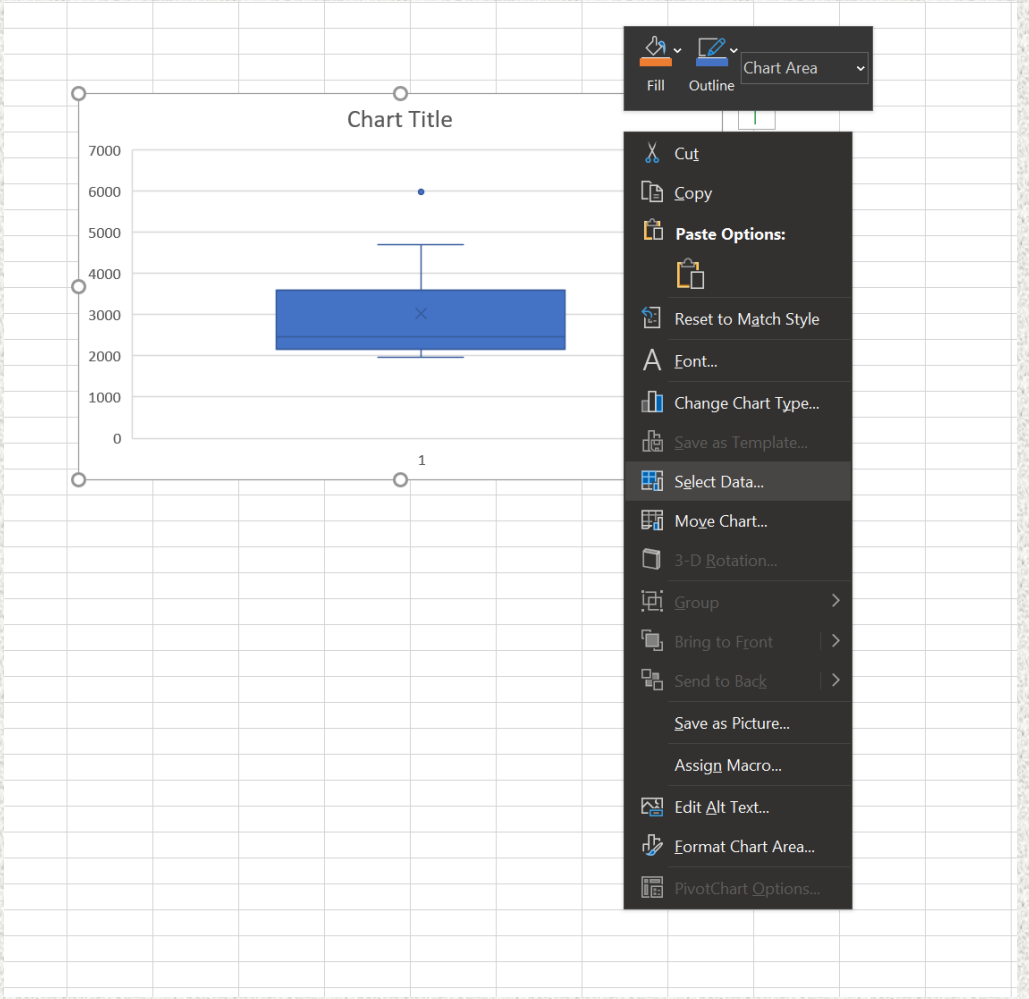
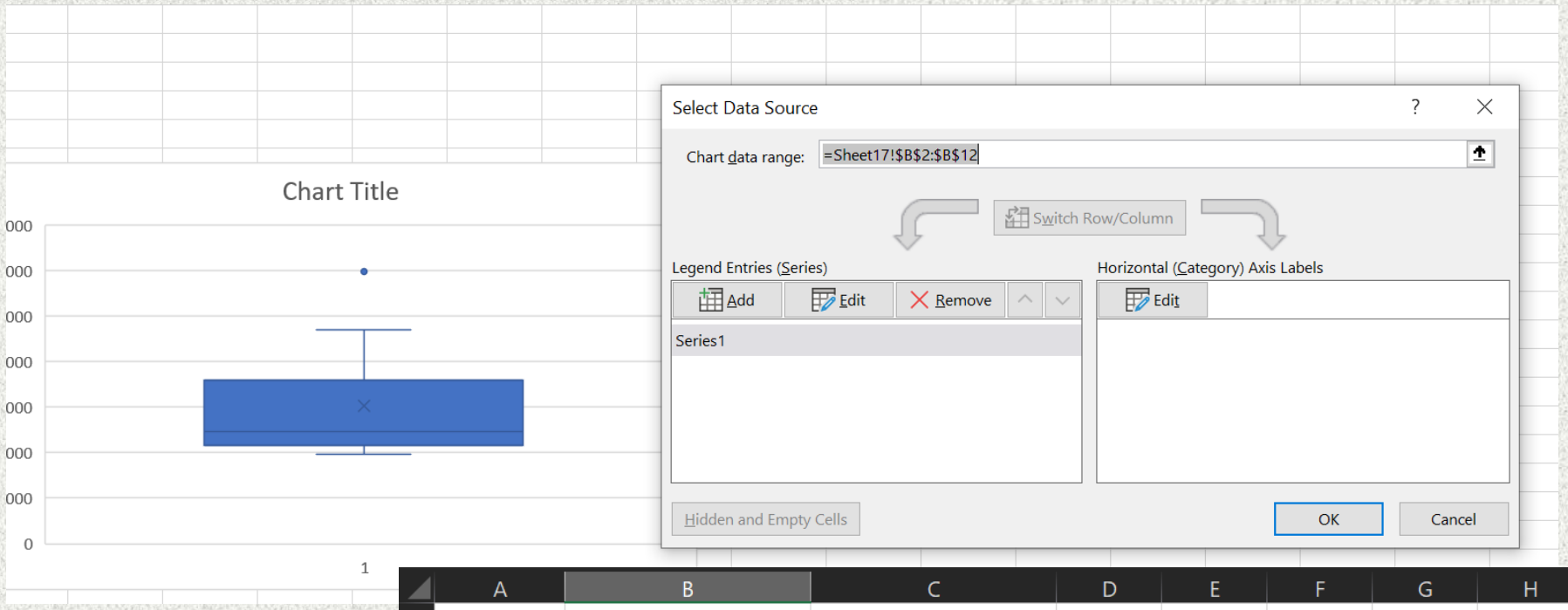


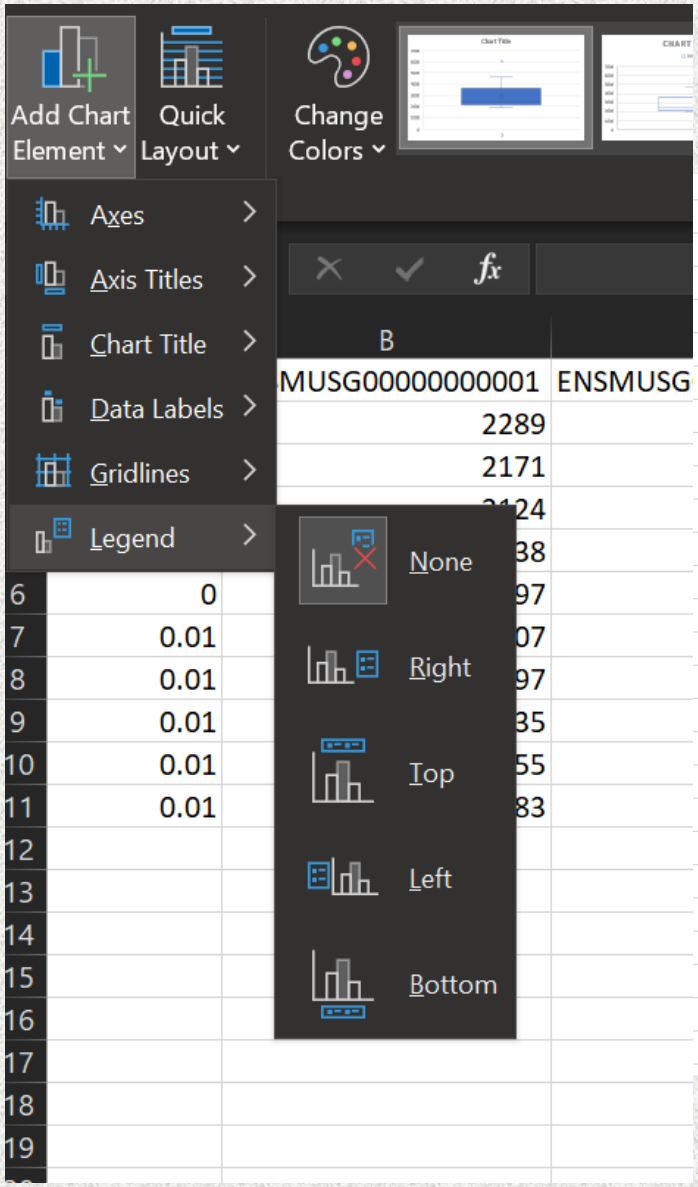
Chart Title
7000
6000
5000
4000
3000
2000
1000
0
1
Fill
Outline
Chart Area
Cut
Copy
Paste Options:
Reset to Match Style
Font...
Change Chart Type...
Save as Template...
Select Data...
Move Chart...
3-D Rotation...
Group
Bring to Front
Send to Back
Save as Picture...
Assign Macro...
Edit Alt Text...
Format Chart Area...
PivotChart Options...

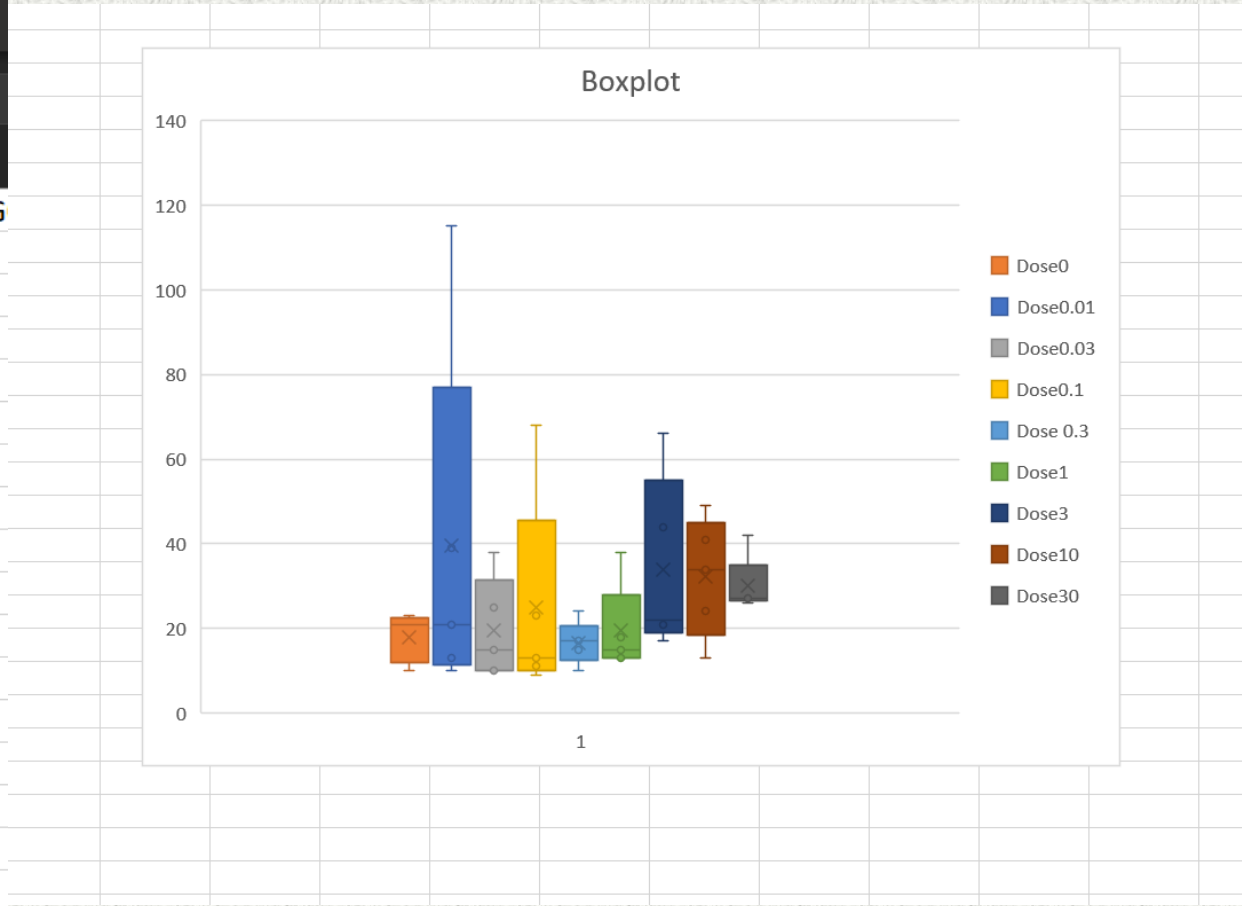
Boxplot



	A	B	C
1	CLASS:DOSE	ENSMUSG00000000001	ENSMUSG00000000049
2	0	2289	30854
3	0	2171	
4	0	2124	
5	0	2538	
6	0	2397	
7	0.01	4707	
8	0.01	2797	
9	0.01	3235	
10	0.01	1955	
11	0.01	5983	92976
12			
13			
14			
15			

Boxplot


Add Chart Element
Quick Layout
Change Colors
Axes
Axis Titles
Chart Title
Data Labels
Gridlines
Legend
None
Right
Top
Left
Bottom
B
MUSG00000000001
ENSMUSG
2289
2171
0
0.01
0.01
0.01
0.01
0.01


Boxplot
140
120
100
80
60
40
20
0
1
Dose0
Dose0.01
Dose0.03
Dose0.1
Dose 0.3
Dose1
Dose3
Dose10
Dose30

R Introduction

- R is a statistical computing language that is very flexible and can do everything we have talked about up to this point and more!
- Can create publication ready plots fairly easily
- RStudio is the primer Interactive Development Environment and has a host of helpful features that makes coding in R as simple as possible
- Can download R and RStudio from the links below
  - R: https://cran.r-project.org/
  - RStudio: https://www.rstudio.com/products/rstudio/

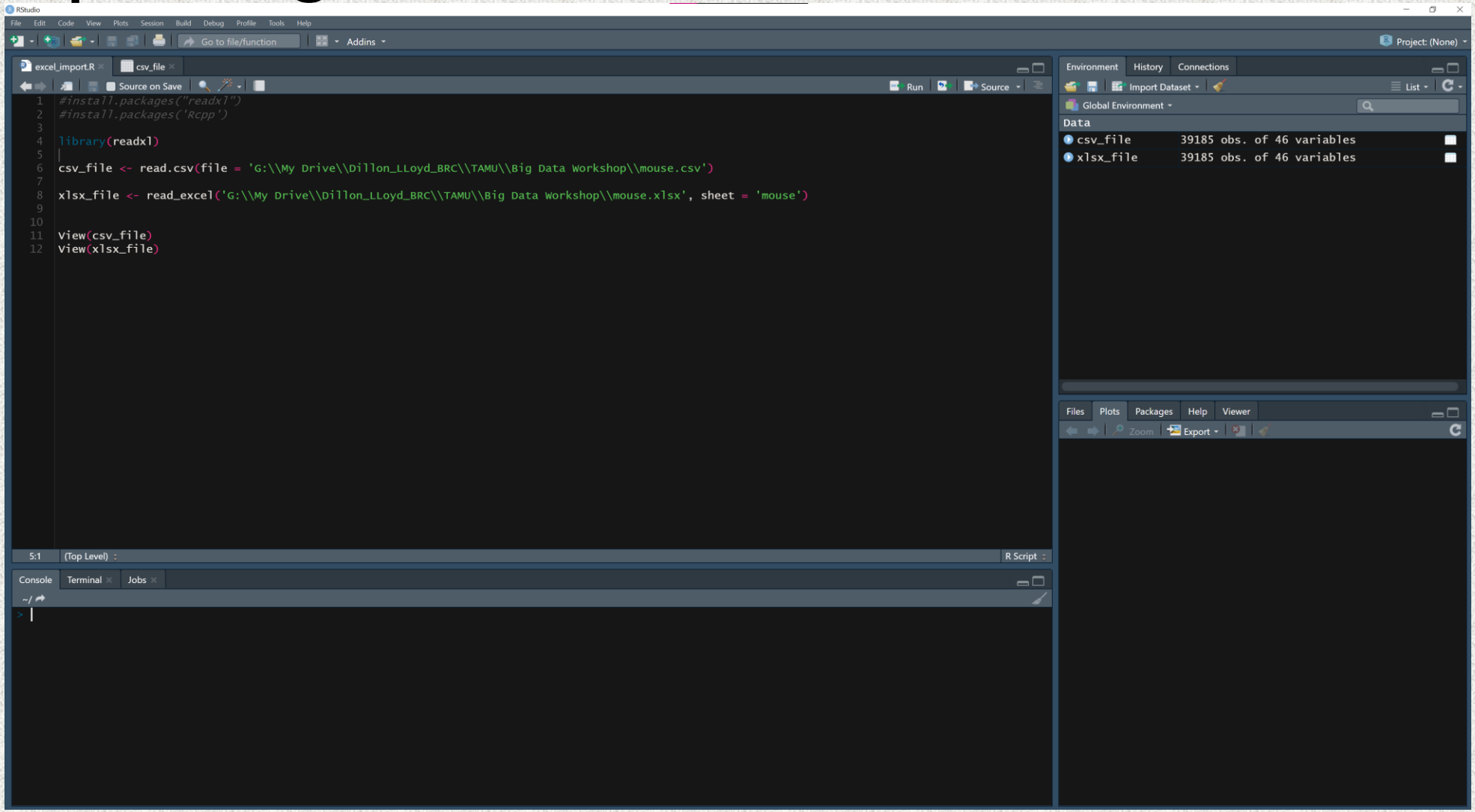
Importing Excel Data into R

- Make sure column names are consistent across files
  - Column names should not have special characters or spaces
  - Capitalization needs to be consistent, case matters
- Know what kind of file you are reading in
  - Are you reading in a csv file? Or a .xlsx file that contains multiple pages?
- Know where your data is located
  - What folder? Is it on a google drive or some other network drive?
- Have a plan for your data
  - Statistical Tests? Plots? Have a plan before you start

Importing .csv and .xlsx files into R

- First, locate the file that you want to import on your computer
- Next, open RStudio and click file → New File → R Script
- Then, type the code in the next slide
- Make sure to replace the file path in the green quotations with the location of the example file on your computer
- Highlight the code, and press run to have R read in the files
- You can view the files in the viewing pane or by printing the dataset
- Now, they will be saved into your R workspace and can be used for analysis

Importing .csv and .xlsx files into R



Importing .csv and .xlsx files into R

excel_import.R | xlsx_file | csv_file

Filter

	NAME	GSM1535917	GSM1535918	GSM1535919	GSM1535920	GSM1535921	GSM1535922	GSM1535923	GSM1535924	GSM1535925	GSM1535926	GSM1535927
1	CLASS:DOSE	0	0	0	0	0	0.01	0.01	0.01	0.01	0.01	0.03
2	ENSMUSG00000000001	2289	2171	2124	2538	2397	4707.00	2797.00	3235.00	1955.00	5983.00	1918.00
3	ENSMUSG00000000003	0	0	0	0	0	0.00	0.00	0.00	0.00	0.00	0.00
4	ENSMUSG00000000028	21	10	22	23	14	39.00	10.00	21.00	13.00	115.00	10.00
5	ENSMUSG00000000031	7	8	5	8	7	18.00	3.00	102.00	8.00	51.00	8.00
6	ENSMUSG00000000037	2	5	4	0	1	7.00	0.00	7.00	3.00	4.00	0.00
7	ENSMUSG00000000049	30854	34567	38467	34859	46227	59527.00	37223.00	42136.00	31393.00	92976.00	31710.00
8	ENSMUSG00000000056	1005	1183	1043	1417	1004	2712.00	1110.00	1318.00	1204.00	2113.00	1036.00
9	ENSMUSG00000000058	88	77	87	75	91	188.00	95.00	66.00	47.00	264.00	83.00
10	ENSMUSG00000000078	244	218	171	171	232	480.00	162.00	603.00	210.00	803.00	210.00
11	ENSMUSG00000000085	92	113	140	126	149	251.00	142.00	154.00	113.00	386.00	108.00
12	ENSMUSG00000000088	1911	1770	1547	1931	2413	2998.00	2190.00	1912.00	1614.00	5048.00	1512.00
13	ENSMUSG00000000093	35	26	44	38	55	68.00	37.00	32.00	36.00	145.00	33.00
14	ENSMUSG00000000094	0	0	0	0	0	0.00	0.00	0.00	0.00	0.00	0.00
15	ENSMUSG00000000103	0	0	0	0	0	0.00	0.00	0.00	0.00	0.00	0.00
16	ENSMUSG00000000120	22	17	14	24	38	63.00	54.00	54.00	13.00	54.00	15.00
17	ENSMUSG00000000125	0	0	0	0	0	0.00	0.00	0.00	0.00	0.00	0.00
18	ENSMUSG00000000126	13	11	6	3	9	27.00	9.00	7.00	7.00	31.00	8.00
19	ENSMUSG00000000127	94	118	143	157	107	308.00	153.00	158.00	111.00	398.00	95.00

Questions??

TEXAS A&M UNIVERSITY
Superfund Research Center

Thank you for joining us !

The next session is on October 6, 2021
2:00 – 4:00 pm Eastern US Time

"Manipulating and Displaying Big(ish) Data in R"

Fred Wright – North Carolina State University
Burcu Beykal – University of Connecticut
Allison Dickey – North Carolina Sate University