

# Excel

(ACTION TASKS)

data-1-1-3-StarterBook.xlsx

## ZOOM

look at the bottom right of the screen, where you'll see a bar with a slider—a negative (minus) sign on the left and a positive (plus) sign on the right. To the right of the plus sign is a percentage. This percentage indicates the magnification of the text. The text can be enlarged by dragging the small white circle to the right, or by clicking the "+" sign until the text is a comfortable size (in some versions of Excel, this may appear as a gray vertical bar instead).

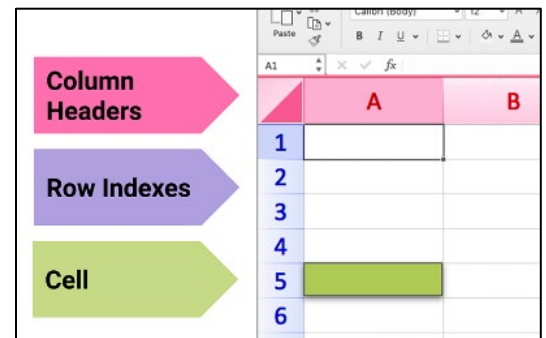
## WORKSHEETS . TABULAR DATA

When you open the data file in Excel, what you're looking at is a **WORKSHEET**, also referred to as a sheet. At a glance, we can see that the data is arranged in rows and columns. Data in this format is called **TABULAR DATA**. **Tabular data** is data that is displayed in a column and row format. This format isn't limited to Excel spreadsheets; any data displayed as a table is considered to be tabular. This includes digital tables on a website and printed tables in a textbook.

## PARTS OF A WORKSHEET

The letters along the top of the columns represent the **column headers**, and the numbers to the left of the rows are the **row indexes**. The headers and indexes help us identify where each data point is located.

In the following image, the cell is located at the intersection of column A, row 5, so we refer to it as A5.

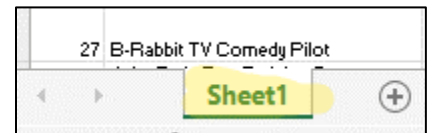


## NAME A WORKSHEET

At the bottom of the sheet is a tab labeled "Sheet1."

This is our current, or active, worksheet.

To create a new worksheet, click the plus sign (+). When multiple sheets are being used, the left and right arrows allow us to navigate between them.



Rename "Sheet1" to "Kickstarter":

Right-click the sheet name > Rename > With the current name highlighted, type the new name > [Enter]

## AUTO FIT COLUMN/ROW TO DATA SIZE

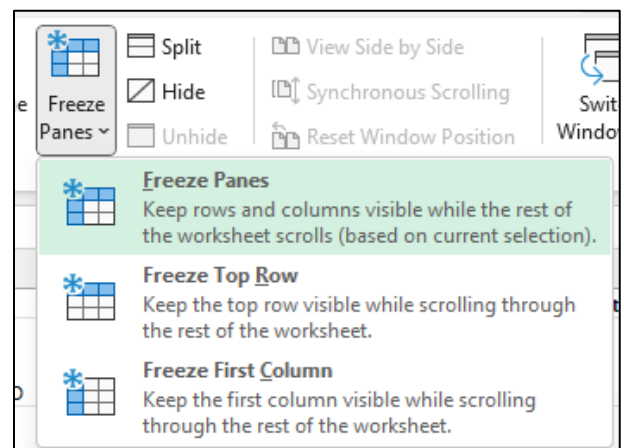
- > Place your cursor over the small line between 2 columns/rows. The cursor should change in appearance to a vertical bar with an arrow pointing either left or right from the center.
- > Double-click to expand the column to fit the value with the most physical width.

## FREEZE COLUMNS/ROWS

**Freezing** allows portions of a spreadsheet to be locked in place so that it is always displayed. When a column and row are frozen together, they become a **pane**. Like a window: the window can be opened and moved, but the pane stays in place. Same idea applies.

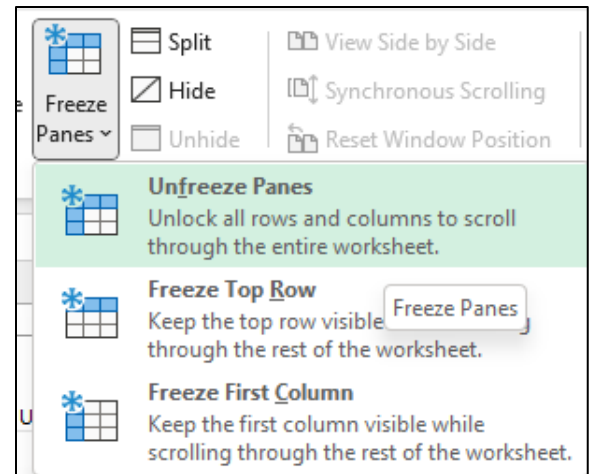
View: Window > Freeze Panes >

	A	B	C	
1	id	name	blurb	
2	122	The Time Jumper (Canceled)	My ambition for this knows no bounds. Seeing Sephoria in a live-action is a dream of mine.	1.0
3	2960	Lynnewood Hall Restoration (Canceled)	Built in the late 1800's, this 70K sq. feet estate has fallen into disrepair. Seeking to buy and convert to useful space	3.0



	A	B	J	K
1	id	name	launched_at	staff_pick
11	619	Big Data (Canceled)	1411745790	FALSE
12	163	UNDIVIDED (Working Title)	1440716654	FALSE

View > Window > Unfreeze >



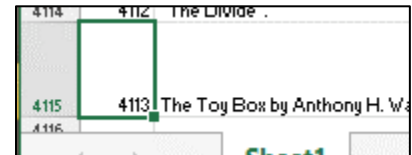
How do you approach a new experience? For instance, think about when you're eating in a restaurant for the first time. You probably peruse the menu to familiarize yourself with the food and drink options. Or consider what you do when you're in a new city. You probably take some time to survey your surroundings, taking in the sights and sounds of your new environment.

### COUNT ROW/COLUMN

A similar process needs to take place for data analysts when they look at a dataset for the first time. They need to size it up and get a feel for what they'll be working with, which is exactly what you'll do next with the Kickstarter data.

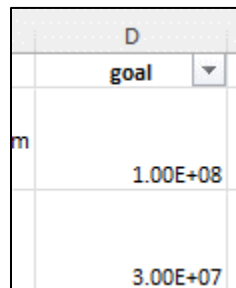
Here are a few things to keep in mind during this initial review:

- How many columns and rows are there? *CTRL and right arrow keys*

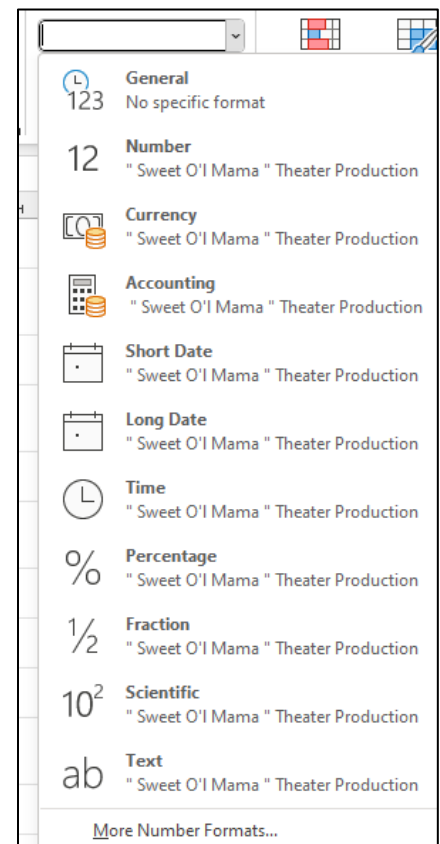


### DETERMINE DATA TYPES

- What types of data are present? *Home: Number > (click on a cell) >*
  - Change data type by selecting the cell(s) and choosing another type
  - SCIENTIFIC NOTATION – keeps numbers compact
    - $3E+07 = 3 \times 10^7 = 30,000,000$



- Is the data readable, or does it need to be converted in some way?
  - Right now, the *deadline* and *launched\_at* columns contain Unix timestamps rather than dates in a standard format.



I	J
deadline	launched_at
1437620400	1434931811
1488464683	1485872683

- How do we know the data is Unix timestamps and not random numbers?
  - data is supposed to represent a date, and
  - timestamps like these are common. But
  - to be sure that the data are timestamps, check with the [timestamp converter tool](#)

- Learn more about the use of [Unix timestamps](#)

## What is Timestamp – UNIX Timestamp: All You Need to Know

What is Timestamp?

The timestamp is a sequence of different characters or information that has been encoded to

### EpochConverter

#### Epoch & Unix Timestamp Conversion Tools

The current Unix epoch time is **1656175145**

Convert epoch to human-readable date and vice versa

1656175135 Timestamp to Human date [batch convert]

Supports Unix timestamps in seconds, milliseconds, microseconds and nanoseconds.

Mon Day Yr Hr Min Sec

to Timestamp

## INTERPRET COLUMN MEANINGS

D	E	F	G
how much money each campaign will need to succeed	how much each campaign actually made	if the campaign met its goal	where the campaign was started
goal	pledged	outcomes	country

## FILTERING

**Filters** allow us to display only the specific data that we want to focus on to filter out “noise” we aren’t interested in, or zoom in on data we want to know more about.

Let's focus first on the money raised by various campaigns. Louise estimates that her play will cost \$12,000, so we can use data from the Pledged column to research projects with a similar monetary goal.

Add a filter & sort:

Select entire column D by clicking the “D” > Data: Sort & Filter > Filter > ▼ > ‘sort Lg to sm’

D
goal

Data Review View

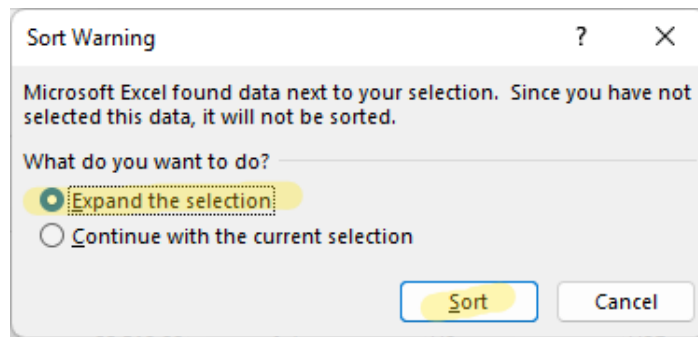
Sort

Sort

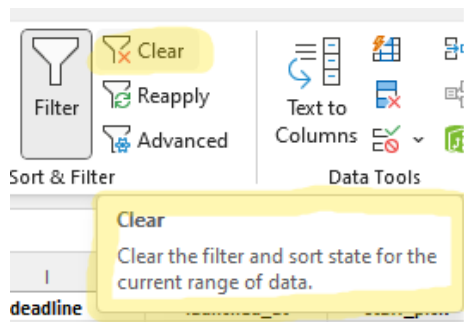
Filter

Sort & Filter

C	D
blurb	goal
<div> <div>Sort Smallest to Largest</div> <div>Sort Largest to Smallest</div> <div>Sort by Color</div> </div>	



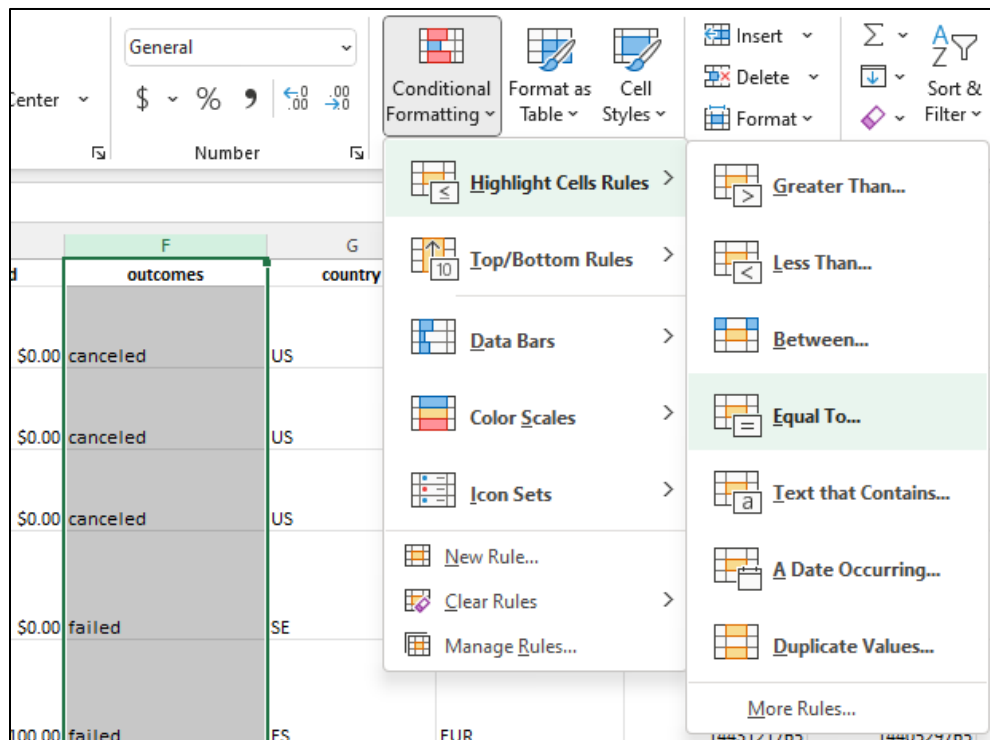
Clear the sort:

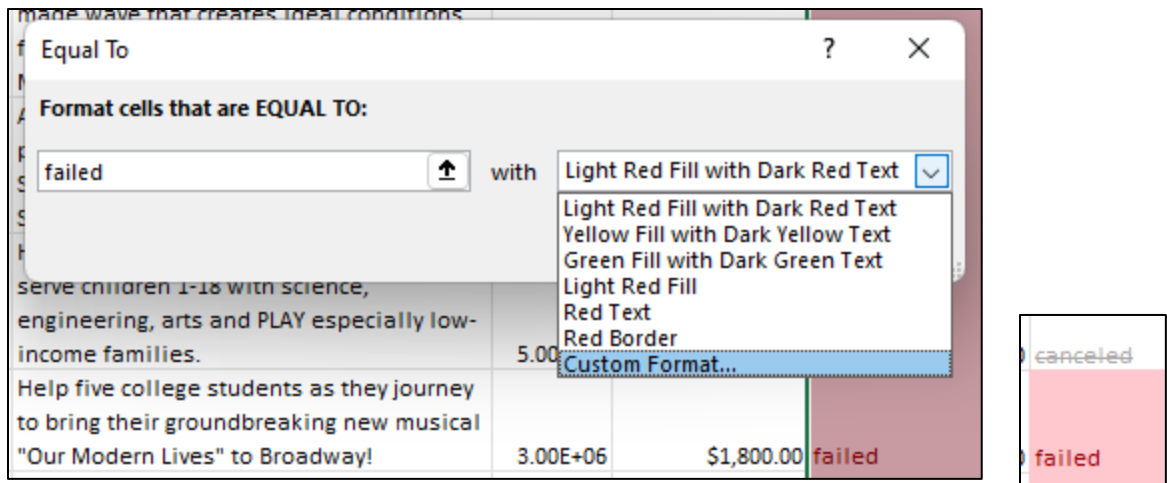


## CONDITIONAL FORMATTING

What do we mean by "visual feedback"? Imagine if traffic lights were words rather than colors. What would that experience be like? Or think of a weather map, where the strongest part of a storm is shaded an angry red. In the cases of traffic lights and weather maps, colors provide a visual link to a specific situation. We can do the same in our data worksheet with conditional formatting.

Highlight failed outcomes in red & canceled grey with strikeout





Format just like you would text in a text box

### CREATE NEW COLUMN

The ability to visually process outcomes at a glance will be very useful for Louise. Let's add a bit more customization to the sheet by creating another easy-to-interpret column. Many of the campaigns missed their goal amount by a small margin. Instead of looking at both the Goal and Pledged columns to determine the deficit, let's create a new column that contains this information: percentage funded.

$\text{=ROUND}(E2/D2*100,0)$

$\text{= ROUND formula ( pledged } \div \text{ goal } \times 100 \text{ , \# of decimal places to show )}$

### APPLY THE FORMULA TO THE ENTIRE COLUMN

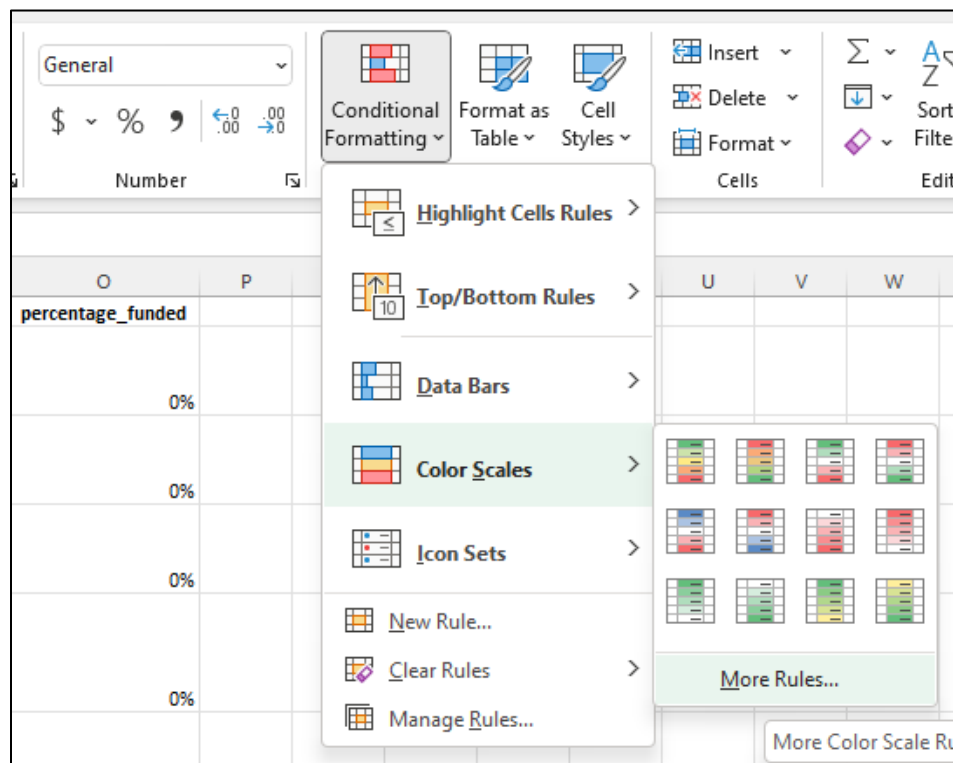
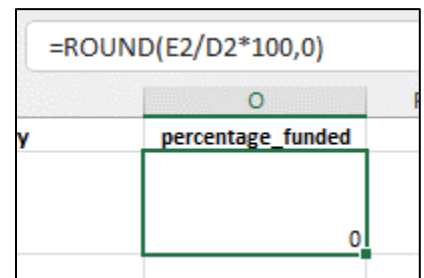
hover over lower right of cell O2, crosshairs should appear > double-click

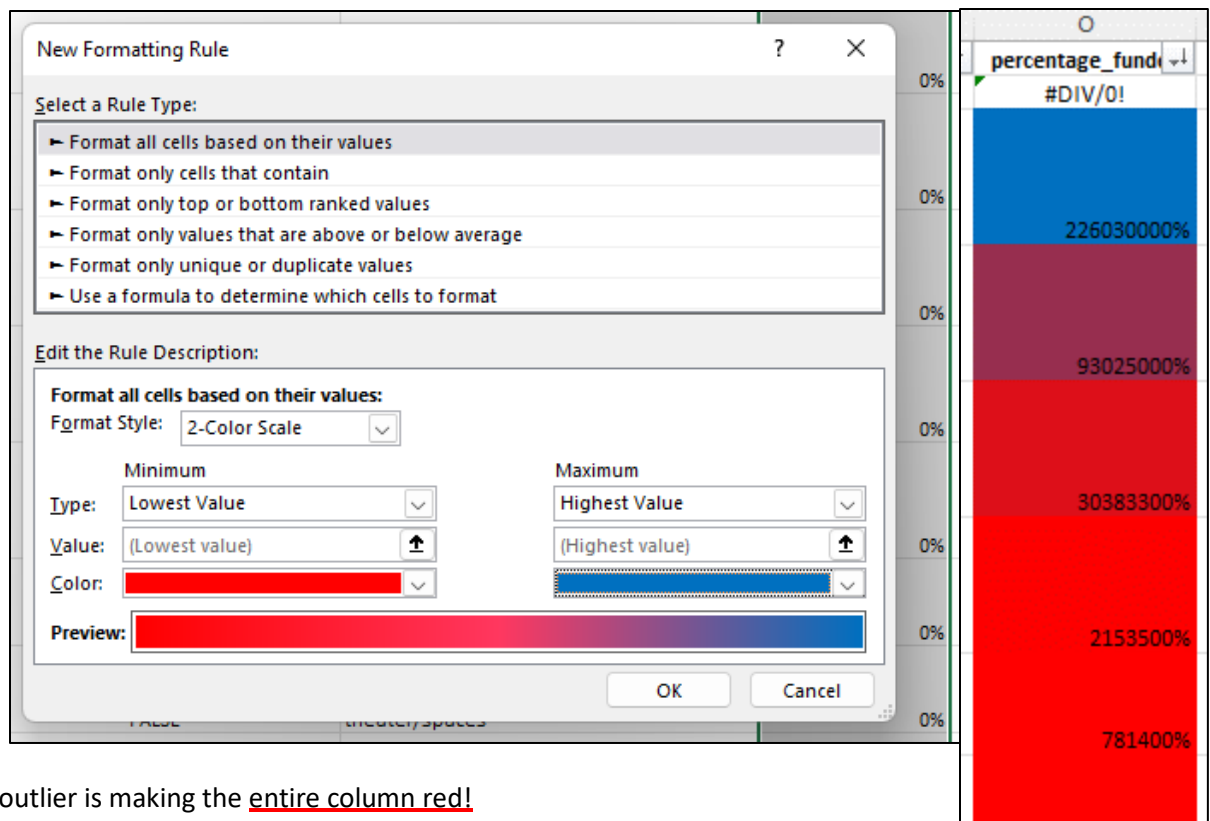
### CHANGE TO % DATA TYPE

select column O > Home: Number > %

### ADD COLOR SCALE

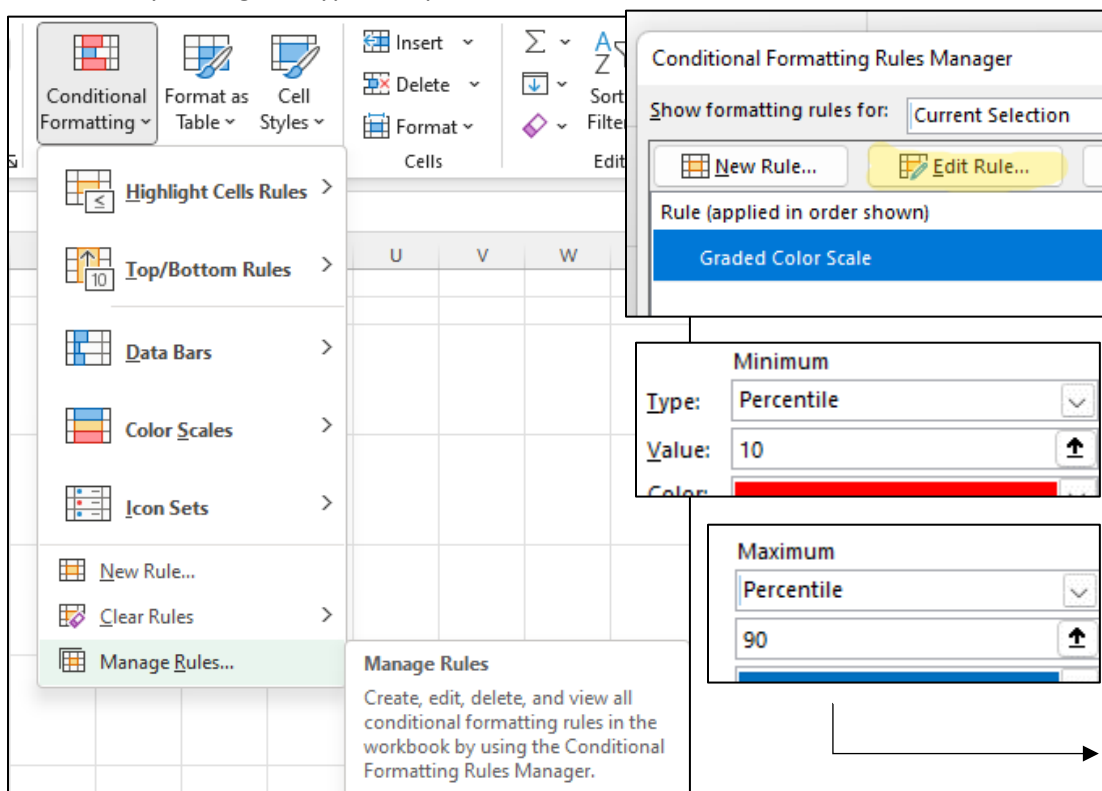
select column O > Home: Number > Conditional Formatting > Color Scales > More Rules





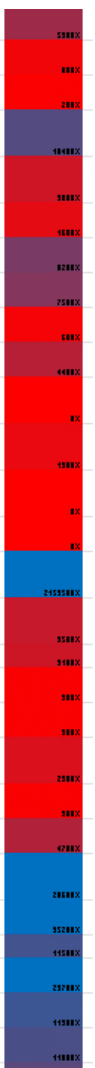
Wow! This outlier is making the entire column red!

DEAL WITH OUTLIERS by editing the Types to “percentiles”



Set incentive levels by determining how much money people have pledged to campaigns historically.

- Create a column that will provide a quick look at the average donation.
- Enter the formula & apply to entire column  
=ROUND(E2/L2,2)



=ROUND formula ( pledged ÷ backers\_count , # of decimal places to show )

## ERROR DETECTED !

### #DIV/0

Kickstarter requires every campaign to have a fundraising goal. However, not every campaign has backers, which means, in some cases, there is no number to divide by in the formula. Our formula, =ROUND(E2/L2,2) uses data from the Pledged and Backers columns. Let's look at row 124 and plug in the numbers ourselves. Now, our formula becomes =ROUND(0/0,2). If we were to take out a calculator and try to divide 0 by 0, we'd get an error there, too. No wonder it's not working correctly. The #DIV/0! error occurs because numbers are not divisible by zero.

While this error doesn't hinder our research, we can and should clean it up...

### =IFERROR(value,value\_if\_error)

Catches errors and replaces them with a user-defined input.

## NESTING FORMULAS

Add a bit of a twist by nesting this formula and the ROUND formula.

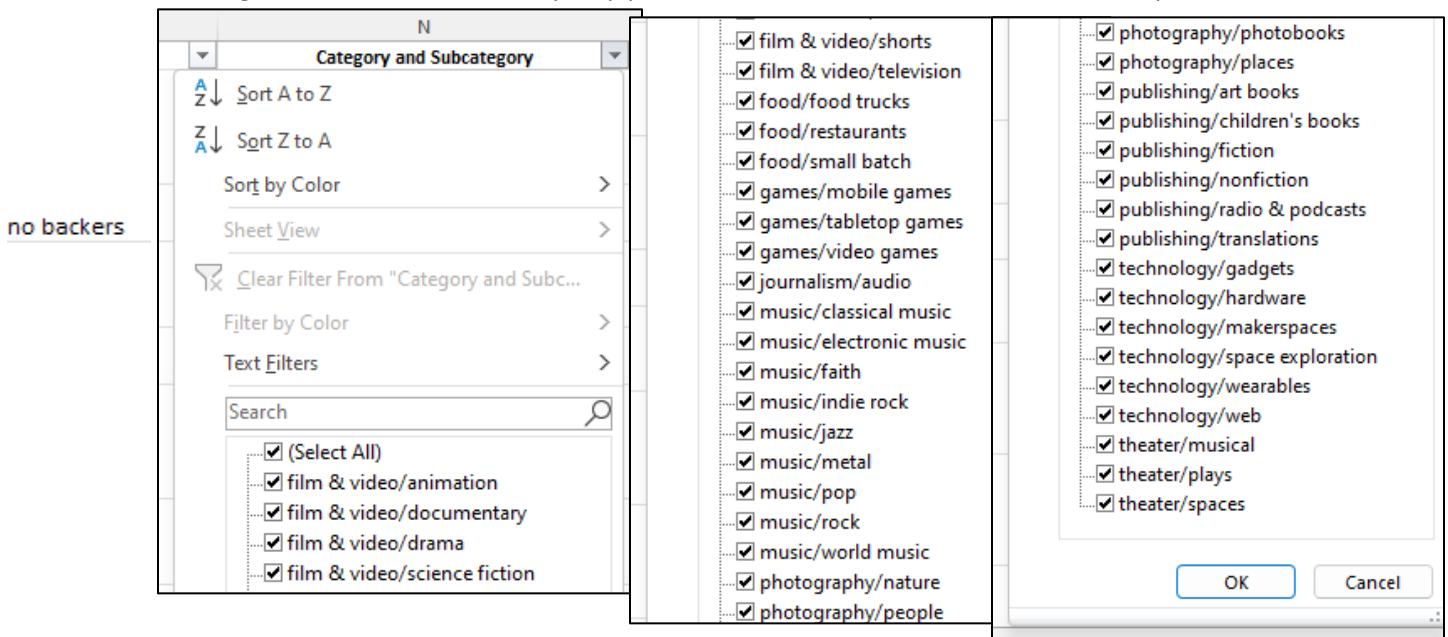
New formula:

=IFERROR(ROUND(E2/L2,2),0)

= IFERROR ( ROUND formula ( pledged ÷ backers\_count , # of decimal places to show ) , change to # )

## VIEW ALL UNIQUE VALUES

Notice any differences? Similarities? This will allow a more focused view of our category of choice by trimming down the data and eliminating what we don't need. Easy way to do this is click on the filter to see all the options.



## REPLACE ERROR VALUES WITH TEXT

Scroll to find the first cell of the Average Donation column with a 0 in it and replace the 0 with text, such as "no backers." Then, apply the updated formula to the entire column and view the result.

=IFERROR(ROUND(E2844/L2844,2),"no backers")

There are many different ways to address errors with the IFERROR formula. Because the column we're working with is a numerical data type, change the formula back to 0.

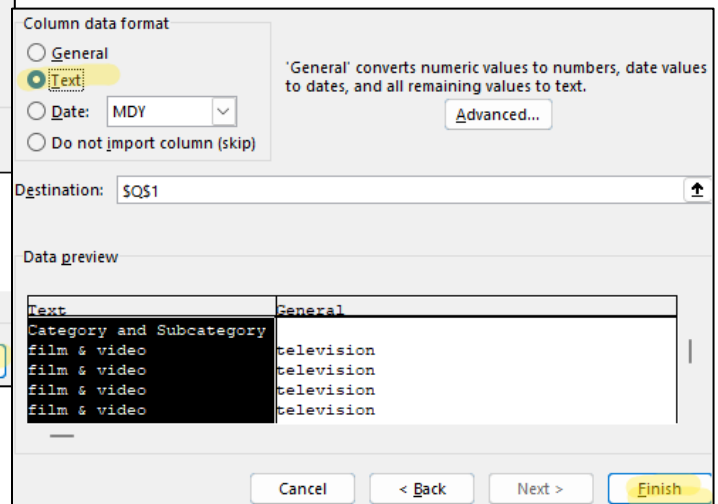
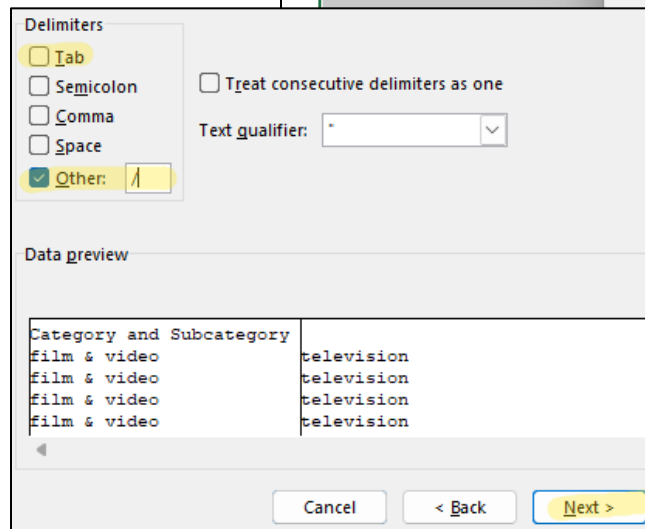
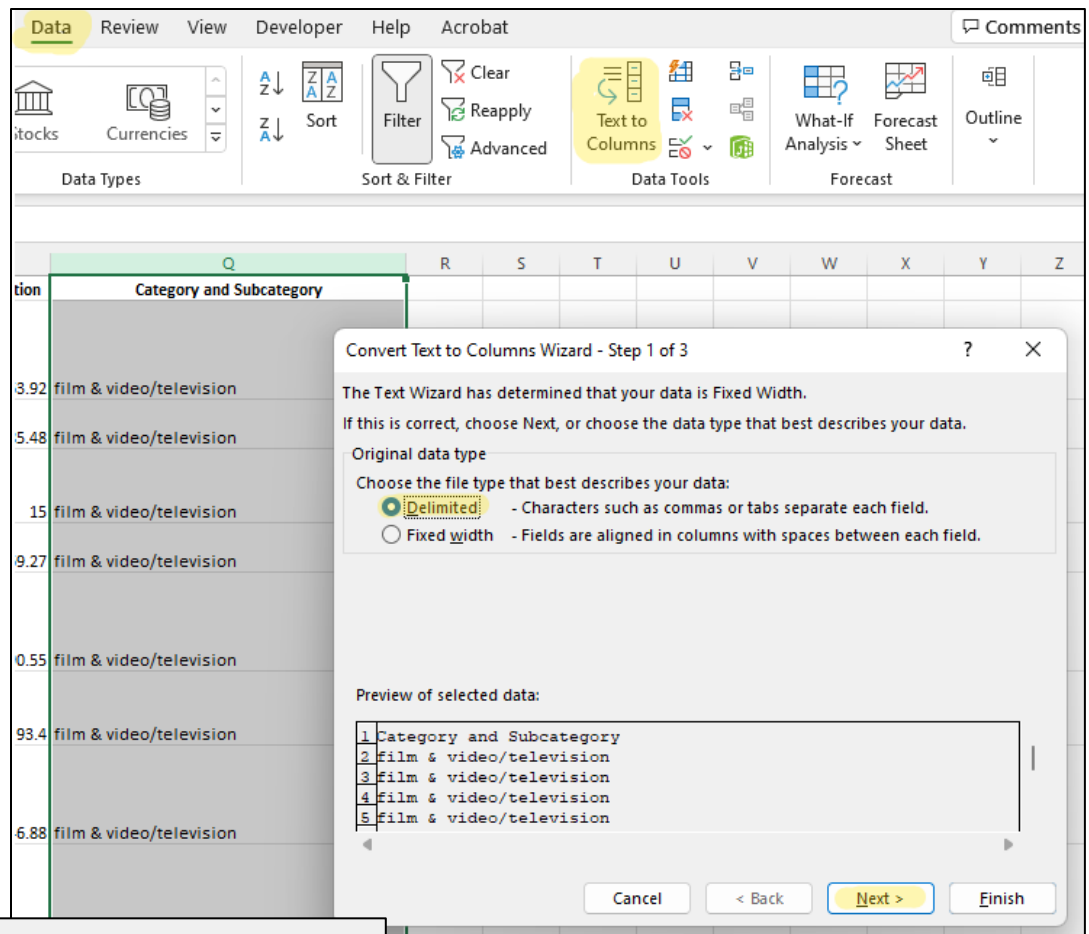
60.67
no backers
no backers
23.33



## SPLIT COLUMNS

make our data more detailed by splitting the Category and Subcategory column into two distinct columns: "Parent category" and "Subcategory." This gives us additional data to use in our analysis.

Clear all active filters >  
Copy the "Category and Subcategory" column & paste it into the next empty column (Q) >  
Highlight column Q ----->



Q	R
Category and Subcategory	
film & video	television

Q	R
Parent Category	Subcategory
film & video	television

Rename columns >



## PIVOT TABLES

condenses data into a summary that delivers information based on our questions.

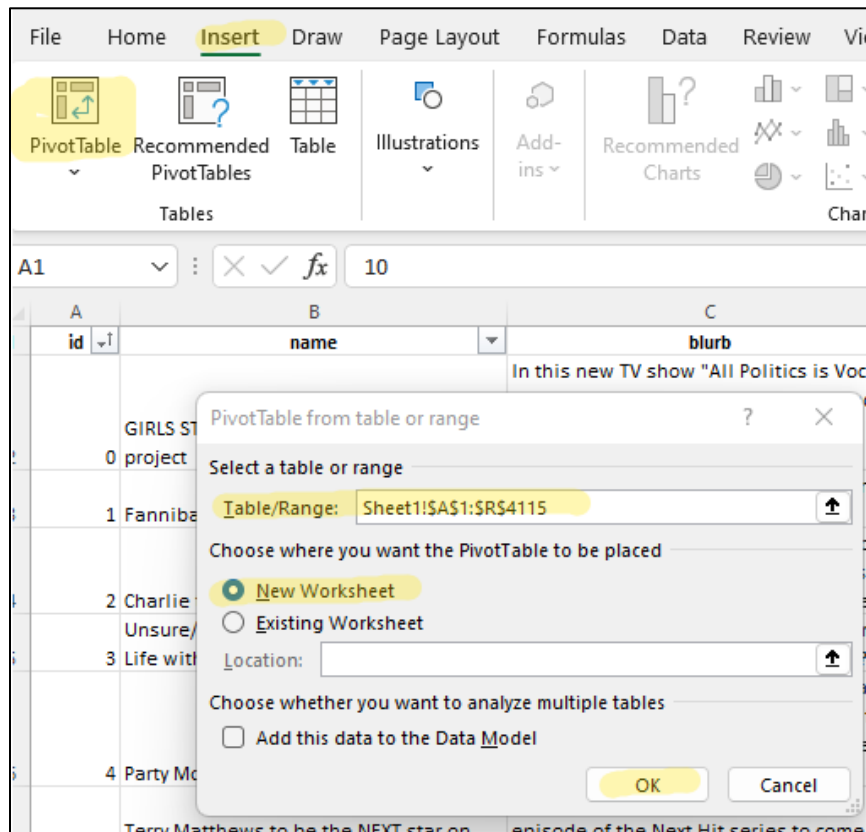
allow us to pick and choose the data we want to analyze and then tweak it with visual customizations. Pivot tables also let us continue to tweak the view by filtering our chosen data after it's been set to a graph.

Ask yourself the following:

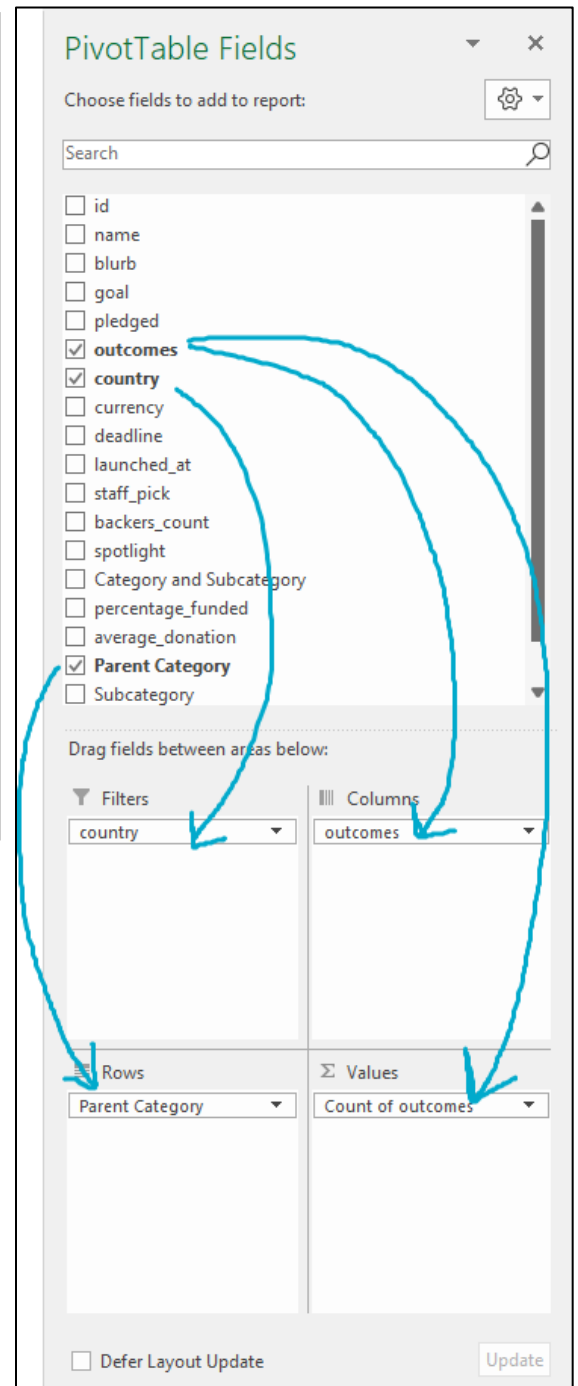
- Which data do we want to see summarized?
- How do we want the data to be presented?

## CREATE A PIVOT TABLE

1. Start a new



2. Drag & Drop fields to report



Creates:

	A	B	C	D	E	F
1	country	(All)				
2						
3	Count of outcomes	Column Labels				
4	Row Labels	canceled	failed	live	successful	Grand Total
5	film & video	40	180		300	520
6	food	20	140	6	34	200
7	games		140		80	220
8	journalism	24				24
9	music	20	120	20	540	700
10	photography		117		103	220
11	publishing	30	127		80	237
12	technology	178	213		209	600
13	theater	37	493	24	839	1393
14	Grand Total	349	1530	50	2185	4114
15						

Play around with the options:

	A	B	C	D	E	F
1						
2	country	(All)				
3						
4	Count of average_donation	Column Labels				
5	Row Labels	canceled	failed	live	successful	Grand Total
6	film & video	40	180		300	520
7	animation		100			100
8	documentary				180	180
9	drama		80			80
10	science fiction	40				40
11	shorts				60	60
12	television				60	60
13	food	20	140	6	34	200
14	food trucks	20	120			140
15	restaurants		20			20
16	small batch			6	34	40
17	games		140		80	220
18	mobile games		40			40
19	tabletop games				80	80
20	video games		100			100
21	journalism	24				24
22	audio	24				24
23	music	20	120	20	540	700
24	classical music				40	40
25	electronic music				40	40
26	faith		40	20		60
27	indie rock		20		140	160
28	jazz		60			60
29	metal				20	20
30	pop				40	40
31	rock				260	260
32	world music	20				20
33	photography		117		103	220
34	nature		20			20
35	people		20			20
36	photobooks		57		103	160
37	places		20			20
38	publishing	30	127		80	237
39	art books	20				20
40	children's books		40			40
41	fiction		40			40
42	nonfiction				60	60
43	radio & podcasts				20	20

### PivotTable Fields

Choose fields to add to report:

Search

- ☐ blurb
- ☐ goal
- ☐ pledged
- ☒ outcomes
- ☒ country
- ☐ currency
- ☐ deadline
- ☐ launched\_at
- ☐ staff\_pick
- ☐ backers\_count
- ☐ spotlight
- ☐ Category and Subcategory
- ☐ percentage\_funded
- ☒ average\_donation
- ☒ Parent Category
- ☒ Subcategory

More Tables...

Drag fields between areas below:

**Filters**

country

**Columns**

outcomes

**Rows**

Parent Category

Subcategory

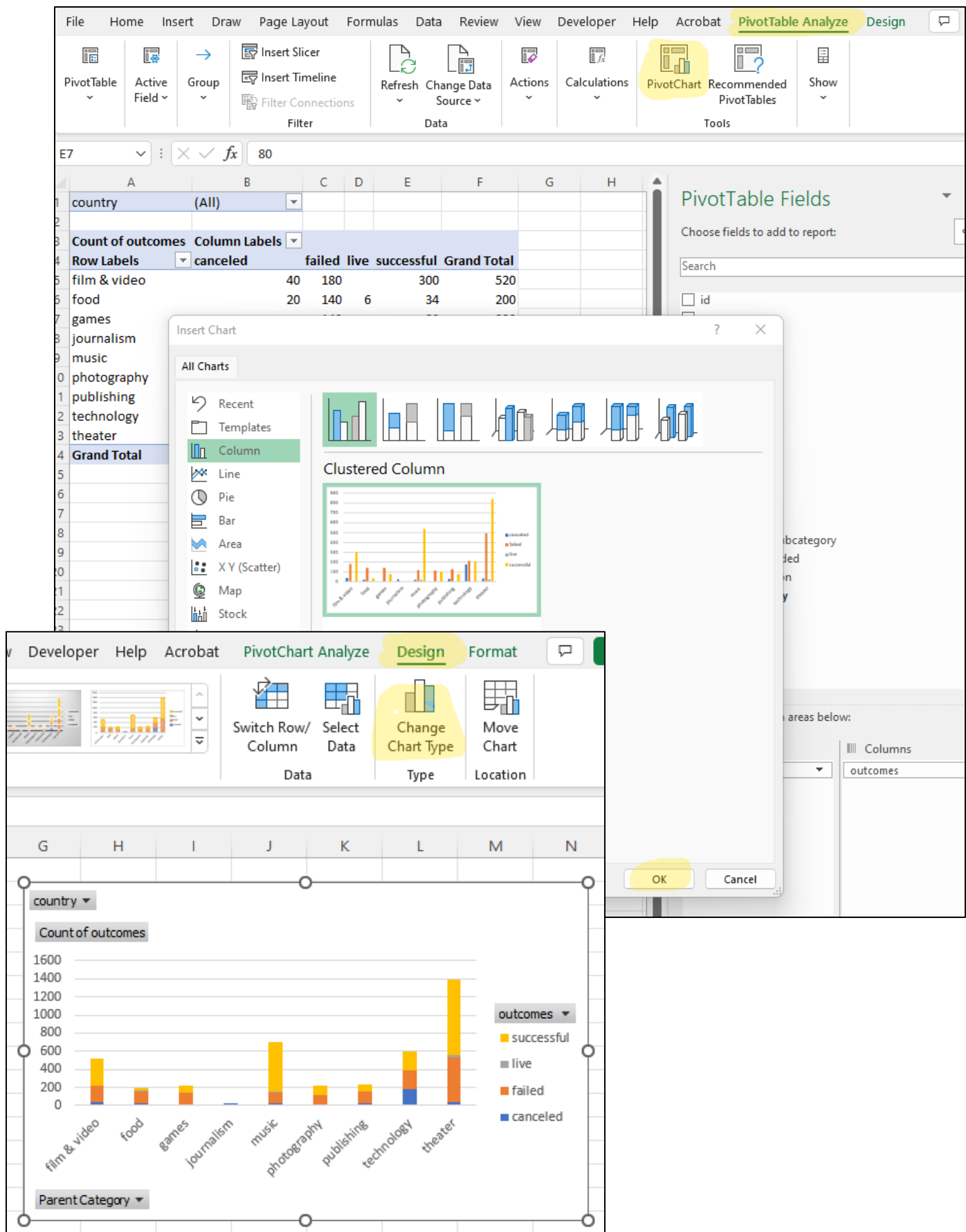
**Values**

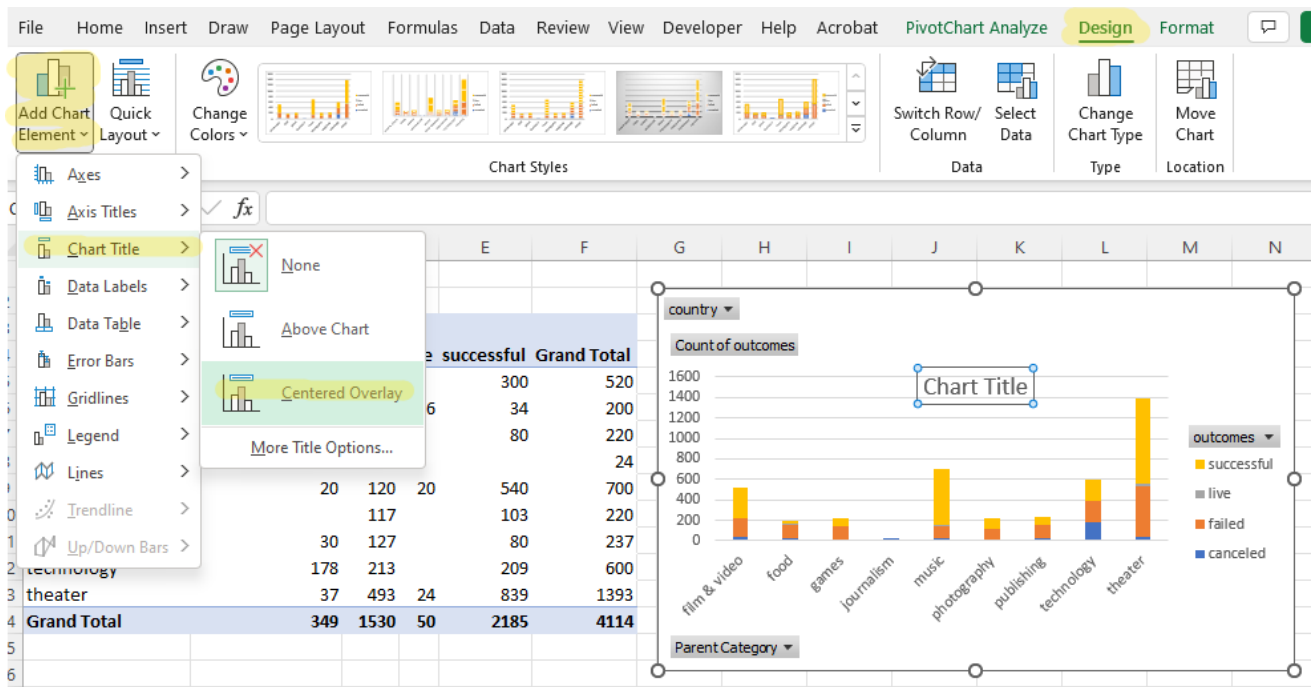
Count of average\_dona...

☐ Defer Layout Update Update

## CREAE PIVOT CHART – BAR GRAPH

When this button is clicked, Excel will automatically choose a chart based on the table data, which in our case is a clustered column.

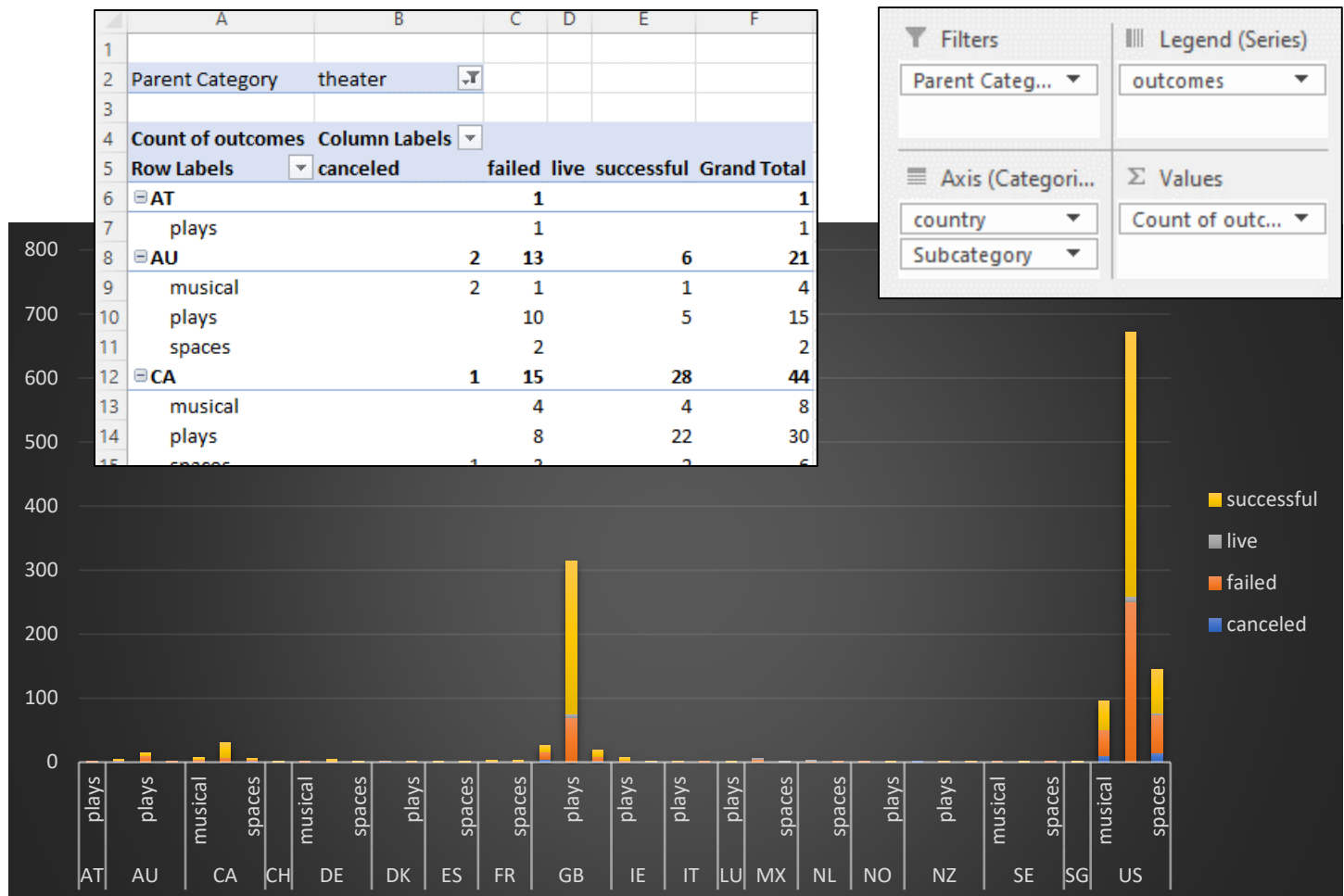




## SAVE THE CHART

to use in our report later:

- Save as-is: right click > save picture as >
- Save but edit first (to remove white space, etc.): right click > copy > paste into paint > edit > save as >

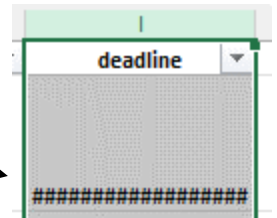


## CONVERT UNIX TIMESTAMPS TO READABLE FORMAT

Simply changing the timestamp to Date does not work.

Copy the columns > paste to empty columns > rename columns

**Formula for Unix Timestamp conversion:**  $=(((J2/60)/60)/24)+DATE(1970,1,1)$



Date that the Unix timestamps began counting from, also known as the **epoch**. Essentially, we're using the formula to figure out how many days, minutes, and seconds the timestamp translates to, and then adding it to the 1/1/1970 date.

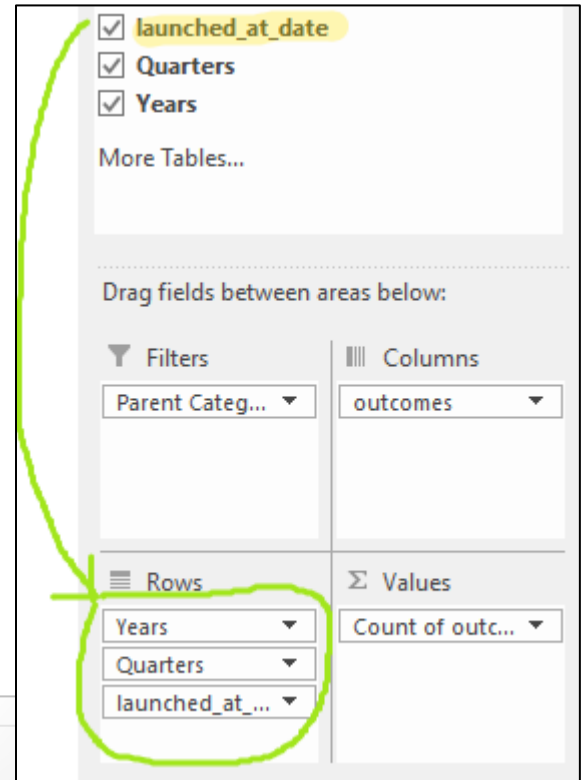
I	J
deadline	launched_at
1437620400	1434931811
1488464683	1485872683

S	T
deadline_date	launched_at_date
7/23/2015	6/22/2015
3/2/2017	1/31/2017

## CREATE PIVOT CHART - LINE GRAPH

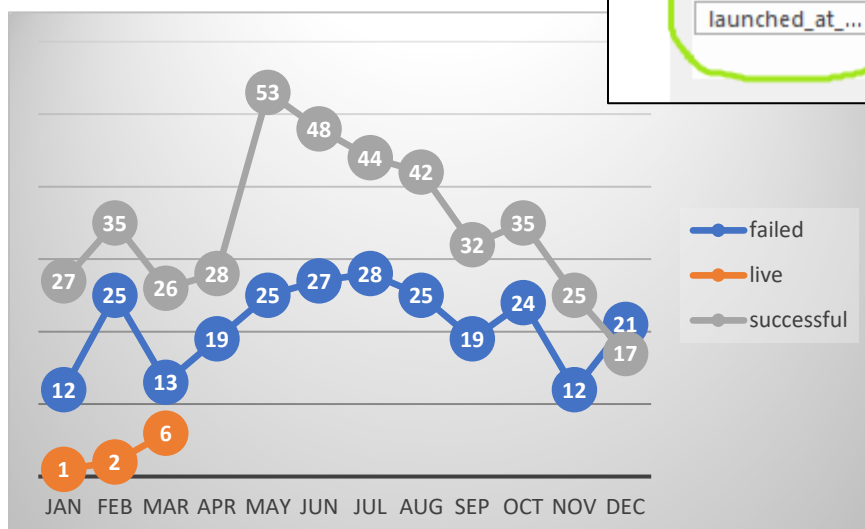
Note: when a date data type field is moved into the report, it automatically splits into sub-categories: years & quarters

Parent Category	(All)				
Count of outcomes	Column Labels				
Row Labels	canceled	failed	live	successful	Grand Total
2009	1	4		9	14
2010	1	15		49	65
2011	7	28		136	171
2012	6	60		216	282
2013	7	67		200	274
2014	80	422		474	976
2015	131	527		567	1225
2016	99	376		475	950
2017	17	31	50	59	157
Grand Total	349	1530	50	2185	4114



Can change a copied & pasted chart from Excel to Word, then still change the formatting!!!!

Click on the chart to try →



## PLAYING WITH FILTERS & DATE DATA

A		B	
1			
2	Parent Category	(All)	▼
3	Years	(All)	▼
4			
5	Count of outcomes	Column Labels	▼
6	Row Labels	▼ canceled	failed live successful Grand Total
7	Jan	34	149 2 182 367
8	Feb	27	106 18 202 353
9	Mar	28	108 30 180 346
10	Apr	27	102 192 321
11	May	26	126 234 386
12	Jun	27	147 211 385
13	Jul	43	150 194 387
14	Aug	33	134 166 333
15	Sep	24	127 147 298
16	Oct	20	149 183 352
17	Nov	37	114 183 334
18	Dec	23	118 111 252
19	Grand Total	349	1530 50 2185 4114

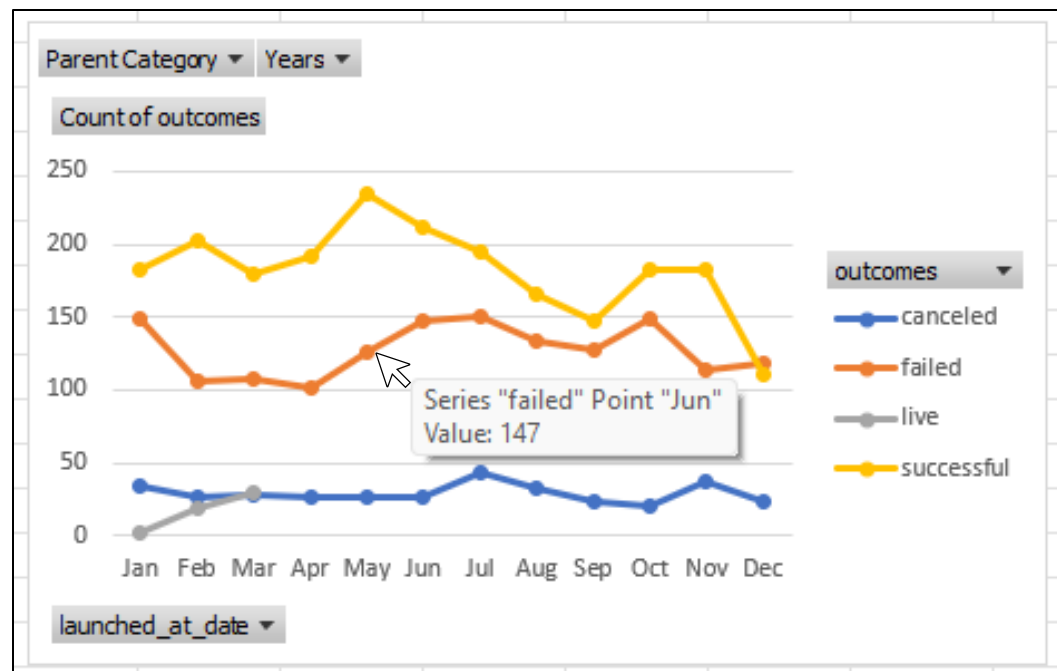
Filters

Parent Categ... ▼  
Years ▼

Legend (Series)

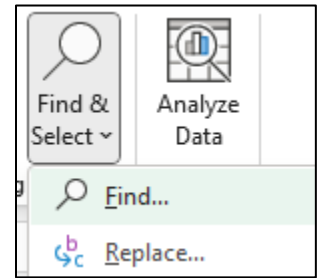
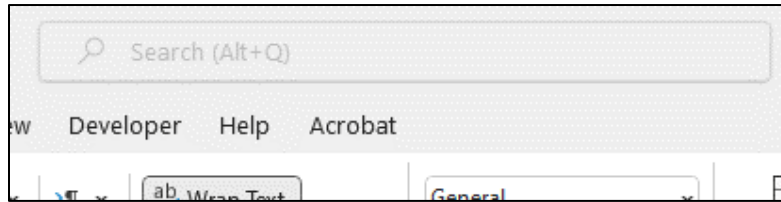
Axis (Categori...  
launched\_at\_... ▼

Σ Values  
Count of outc... ▼



## SEARCH & FIND

Single find: 3 ways: [ctrl]+F

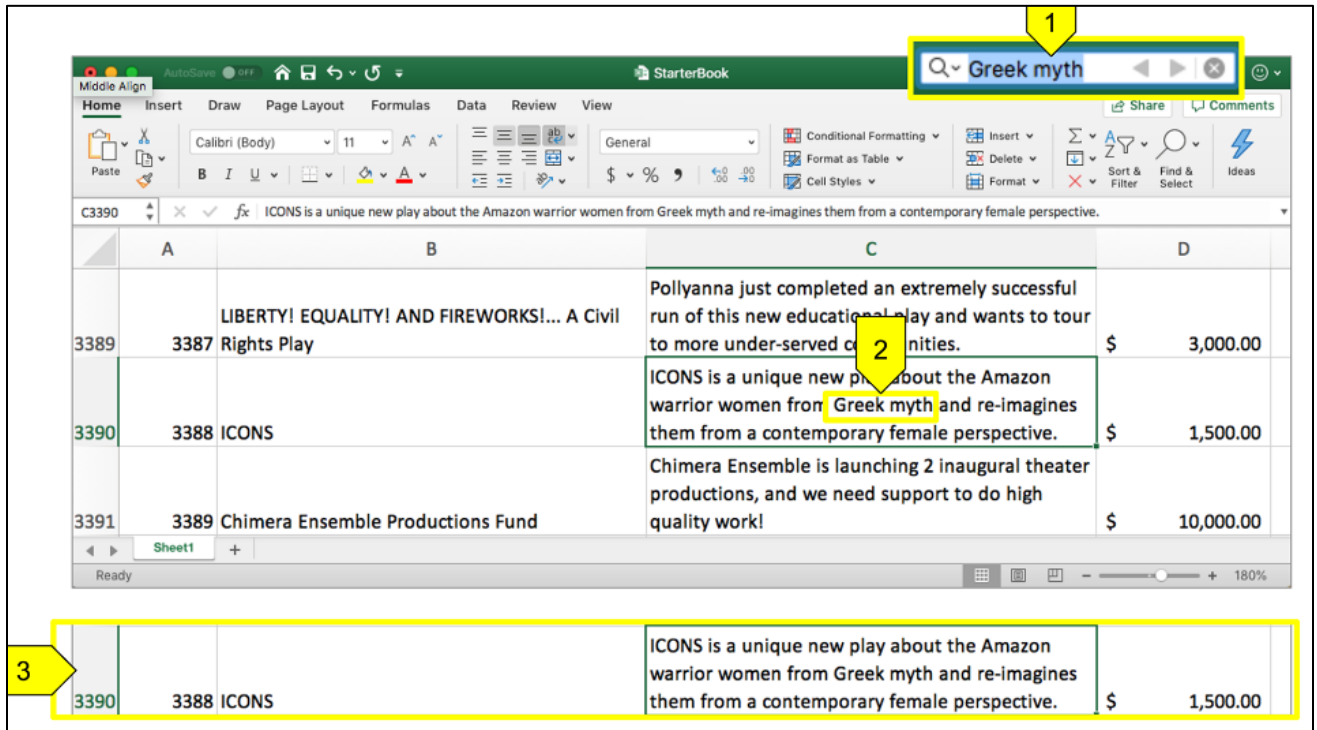


Multiple Search:

## VLOOKUP

lets us pull specific columns from our main dataset into a new sheet without having to search for each column and then copy and paste the data. This way, we can pull only the data points we're interested in.

Ex: if we only want to see the blurb of the play, we can tell Excel to pull only the data in the blurb cell for that play.



1. Search for 'Greeks myth' in the formula bar.

2. Select the cell containing the search result (C3390).

3. Select the cell containing the search result (C3390).

	A	B	C	D
3389		LIBERTY! EQUALITY! AND FIREWORKS!... A Civil Rights Play	Pollyanna just completed an extremely successful run of this new educational play and wants to tour to more under-served communities.	\$ 3,000.00
3390		3388 ICONS	ICONS is a unique new play about the Amazon warrior women from Greek myth and re-imagines them from a contemporary female perspective.	\$ 1,500.00
3391		3389 Chimera Ensemble Productions Fund	Chimera Ensemble is launching 2 inaugural theater productions, and we need support to do high quality work!	\$ 10,000.00

	A	B	C	D
3390		3388 ICONS	ICONS is a unique new play about the Amazon warrior women from Greek myth and re-imagines them from a contemporary female perspective.	\$ 1,500.00

## SIMPLE VLOOKUP

Get blurb for 5 specific plays

1. Take note of the original worksheet's columns & order:

B	C	D	E
name	blurb	goal	pledged
Darktales The Play	Tim Arthur's 21st anniversary sell-out produ	\$ 5,000	\$ 5,105
Zero Down	Angel on the Corner need YOUR help to rais	\$ 3,500	\$ 3,804

2. Make new worksheet
3. Name the columns
4. Put the lookup values in column A, must be exact match to original data:
5. In column B start the formula: ,



blurb                    =VLOOKUP(A2, Kickstarter!B:C, 2, FALSE)

“Look for [value of A2],  
in [Kickstarter worksheet column B, ending with column C],  
and grab corresponding data from the [2<sup>nd</sup> referenced column, C in our case],  
with [FALSE: only exact matches]”

- Copy the formula into B3 through B6 to get the blurbs for all five plays.

	A	B
1	Name	blurb
2	Be Prepared	Help us get actor-writer Ian Bonar's debut play - a hilarious, heartbreaking story of grief and loss - to the 2016 Edinburgh Fringe.
3	Checkpoint 22	The play yet to be described as "A surefire Edinburgh Fringe Festival Cult Hit". Coming to the Underbelly, Edinburgh, 5th-30th August.
4	Cutting Off Kate Bush	Cutting Off Kate Bush is a one-woman show written & performed by Lucy Benson-Brown, premiering at the Edinburgh Fringe Festival 2014
5	Jestia and Raedon	Jestia and Raedon is a brand new romantic comedy play going to the Edinburgh Fringe Festival this summer.
6	The Hitchhiker's Guide to the Family	A one-man show about love, loss, and motorways, written & performed by Ben Norris. Help us get to the 2015 Edinburgh Fringe and beyond!
7		
8		

Edinburgh Research    Theatre Outcomes    Subcategory Outcomes    Outcomes by Launch Date    Kickstart

## MULTIPLE VLOOKUP

Add what each play's goal was and the amount pledged.

goal                    =VLOOKUP(A2, Kickstarter!B:E, 3, FALSE)

“Look for [value of A2],  
in [Kickstarter worksheet column B, ending with column E],  
and grab corresponding data from the [3<sup>rd</sup> referenced column, D in our case],  
with [FALSE: only exact matches]”

pledged                =VLOOKUP(A2, Kickstarter!B:E, 4, FALSE)

“Look for [value of A2],  
in [Kickstarter worksheet column B, ending with column E],  
and grab corresponding data from the [4<sup>th</sup> referenced column, E in our case],  
with [FALSE: only exact matches]”

	A	B	C	D
1	name	blurb	goal	pledged
2	Be Prepared	Help us get actor-writer Ian Bonar's debut play - a hilarious, heartbreaking story of grief and loss - to the 2016 Edinburgh Fringe.	\$ 2,000	\$ 2,020
3	Checkpoint 22	The play yet to be described as "A surefire Edinburgh Fringe Festival Cult Hit". Coming to the Underbelly, Edinburgh, 5th-30th August.	\$ 2,000	\$ 2,020
4	Cutting Off Kate Bush	Cutting Off Kate Bush is a one-woman show written & performed by Lucy Benson-Brown, premiering at the Edinburgh Fringe Festival 2014	\$ 1,500	\$ 2,576
5	Jestia and Raedon	Jestia and Raedon is a brand new romantic comedy play going to the Edinburgh Fringe Festival this summer.	\$ 1,000	\$ 1,168
6	The Hitchhiker's Guide to the Family	A one-man show about love, loss, and motorways, written & performed by Ben Norris. Help us get to the 2015 Edinburgh Fringe and beyond!	\$ 4,000	\$ 4,137

## MEASURES OF CENTRAL TENDENCY

Statistics provide an unbiased view of the data and make conclusions based on calculations rather than gut feelings.

**CENTRAL TENDENCY** refers to the tendency of data to be toward the middle of the dataset. The three key measures:

### MEAN (average)

- sum of the data divided by the number of data points.
- "If every data point contributed the same amount, what would that amount be?"
- Ex, if you and two friends all chipped in to buy a pizza, and you put in \$12, one friend put in \$7, and the other friend put in \$5, the mean cost would be calculated this way:

$$(12 + 7 + 5) / 3 = \\ 24 / 3 = 8$$

### MEDIAN (middle)

- "Where is the midpoint of the data?"
- aka the 50th percentile
- the value that splits the data into two equal halves: 50% of the data is lower than the median, and 50% of the data is higher.
- To calculate the median,
  1. sort the data points in order, and
  2. then locate the point in the middle.
- Ex: if the grades on a quiz are 82, 79, 79, 77, 70, 90, 71, 86, 83, first we would put them in order:
  1. 

70	71	77	79	79	82	83	86	90
1	2	3	4	5	6	7	8	9

 scores sorted in order  
count of scores = 9

midpoint

### SKILL DRILL

	A	B	C	D
1	make a dataset with...	...median = 50	...mean = 50	...mode = 50
2	5	50	50	50
3	17	=MEDIAN(A2:A10)	=AVERAGE(A2:A10)	=MODE(A2:A10)
4	33			
5	45			
6	50			
7	50			
8	58			
9	92			
10	100			

### MODE (most)

- "What value shows up the most?"
- There can be more than 1 mode, or no modes!
- For Ex: 70, 71, 77, 79, 79, 82, 82, 83, 86, 90 scores from above example... there are 2 "mosts"

**SKEWNESS** is a statistic that quantifies how skewed, or asymmetrical, a distribution is.

If the mean is significantly different than the median, the data is **skewed**, meaning that

When the mean and median are close to each other, the data is roughly **symmetric**: half

If the mean is significantly different than the median, the data is **skewed**, meaning

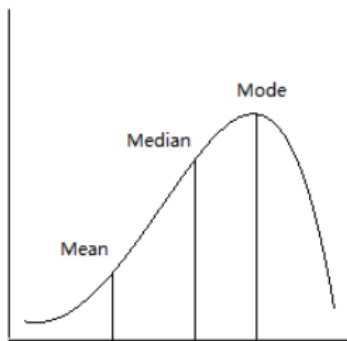
some number of extreme values are pulling the mean higher or lower.

If the mean is much lower than the median, the data is **skewed to the left**.

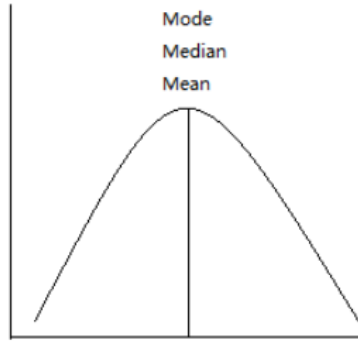
the data is above the mean, half the data is below.

that some number of extreme values are pulling the mean higher or lower.

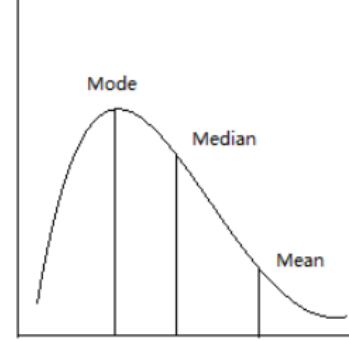
If the mean is much higher than the median, the data is **skewed to the right**.



Left skew



Normal Distribution



Right skew

MCD Example with Kickstarters:

1. Filter Kickstarter data on: US only, Plays only, successful / unsuccessful, &

2. make 2 new sheets with just that data:

**Success US Play Kickstarters**

**Failed US Play Kickstarters**

3. Make a blank chart to hold results:

	A	B	C
1		Successful	Failed
2	Mean Goal		
3	Median Goal		
4			
5	Mean Pledged		
6	Median Pledged		

4. Enter formulas

- ★ "D:D" specifies the entire column.
- ★ When copying & pasting formulas [esc] is super helpful!
- ★ To display formulas (instead of answers): Formulas: Formula Auditing > Show Formulas

	A	B	C
1		Successful	Failed
2	Mean Goal	=AVERAGE('Success US Play Kickstarters'!D:D)	=AVERAGE('Failed US Play Kickstarters'!D:D)
3	Median Goal	=MEDIAN('Success US Play Kickstarters'!D:D)	=MEDIAN('Failed US Play Kickstarters'!D:D)
4			
5	Mean Pledged	=AVERAGE('Success US Play Kickstarters'!E:E)	=AVERAGE('Failed US Play Kickstarters'!E:E)
6	Median Pledged	=MEDIAN('Success US Play Kickstarters'!E:E)	=MEDIAN('Failed US Play Kickstarters'!E:E)

	A	B	C
1		Successful	Failed
2	Mean Goal	\$ 5,049	\$ 10,554
3	Median Goal	\$ 3,000	\$ 5,000
4			
5	Mean Pledged	\$ 5,602	\$ 559
6	Median Pledged	\$ 3,168	\$ 103

## 5. Analysis:

- a. Failed Kickstarter campaigns have much higher fundraising goals than successful Kickstarter campaigns. Louise is asking for more than twice the average successful Kickstarter goal, so this isn't great news for her campaign.
- b. The mean and median pledged amounts are much lower than the successful pledges, which indicates that failed Kickstarter campaigns are unsuccessful for reasons other than asking for too much money. In other words, if the failed projects were also getting a median pledge amount of around \$3,000, it's possible that those that failed just asked for too high of a price. Since the median is much lower, there must be another factor keeping people from pledging to those unsuccessful projects.

## MEASURES OF SPREAD

Measures of central tendency distill a lot of information about the distribution of a dataset down to one number. However, two datasets can have the same mean or median but still look very different—that is, the spread of data between the two datasets can vary quite a bit. When considering the distribution of a dataset, we also want to have measures of its spread. **MEASURES OF SPREAD** include:

### RANGE

- difference between the maximum value of the dataset and the minimum value of the dataset. For our purposes, the range does not capture as much information as we'd like. What we would really like to know is roughly how far each data point is from the center, or mean, or how much of the data is near the center...

### VARIANCE

- measure of how far data points are from the center, or mean.
- Calculation:
  1. Subtract the mean from each data point.
  2. Square the difference so that it's positive.
  3. Take the average of those squared differences.

Because we've taken the average of the *squared* differences, the unit of variance doesn't quite match our dataset. To get the unit to match, we take the square root of the variance to standardize it, or get the...

## STANDARD DEVIATION

- often represented with a lowercase sigma ( $\sigma$ ), or standard deviation squared ( $\sigma^2$ ).
- Ex: Imagine we have a dataset of five backers. We'll signify that this is a **DATASET** by placing the data within brackets: [1, 3, 6, 7, 8], which then makes it into a "set" of numbers. How do we find the standard deviation?
- Let's begin working through the standard deviation equation:
  - $\sigma$ , lowercase sigma, is the symbol for standard deviation.
  - $\sum$ , uppercase sigma, is the symbol for summation or the sum.
  - $X$ , represents each point of data in the dataset.
  - $\bar{X}$ , represents the mean of the dataset and is pronounced "x-bar."
  - $n$ , represents the total number of points in the dataset.

$$\sigma = \sqrt{\frac{\sum (X - \bar{X})^2}{n - 1}}$$

The diagram shows the formula with several annotations: a light blue box around the entire formula, a light green box around the summation symbol and the term (X - X-bar), a light orange box around X, a light purple box around X-bar, and a light blue box around n-1.

This equation looks fairly complicated, so let's talk through what's happening.

1. Find the mean.
2. For each number in the dataset, subtract the mean and square the result.
3. Find the mean of these new numbers.
4. Take the square root of the mean.

Let's apply the equation to our set:

$$1. \text{ Find the mean: } (1 + 3 + 6 + 7 + 8) / 5 = 25 / 5 = 5$$

sum the values / number of values

2. Next, we find the deviations. The deviations from the *sample* mean are

- $(1 - 5) = -4$
- $(3 - 5) = -2$
- $(6 - 5) = 1$
- $(7 - 5) = 2$
- $(8 - 5) = 3$

Ideally, we'd like to know the deviations from the actual population mean, but because we don't know the actual population mean, these deviations have a subtle and slight bias to them. We'll correct that bias in the next step.

3. Find the variance.

1. Square all of those deviations; this way, we will always be working with positive numbers.

16, 4, 1, 4, 9

Square all the deviations

If we take the average of these values, we'll get a slightly smaller variance than the actual population. To correct for this bias, we instead divide by the number of samples minus 1 (if you're curious, this is known as Bessel's correction). Thus, the unbiased average of these

$$\text{values is } (16 + 4 + 1 + 4 + 9) / (5-1) = 34 / 4 = 8.5$$

Std deviations / # of samples - 1

4.  $\sqrt{8.5} \approx 2.92$  = the standard deviation

## INTERQUARTILE RANGE

- Another method for measuring the spread.
- 1. Sort from lowest value to highest value, we can
- 2. break it into four separate parts known as quartiles.

## QUARTILES

- percentiles.
- The lower / 1<sup>st</sup> quartile is the 25th percentile, that is, 25% of the data is less than the lower quartile.
- the upper / 3<sup>rd</sup> quartile is the 75th percentile, so 75% of the data is less than the upper quartile.
- The 2<sup>nd</sup> quartile is the median

## INTERQUARTILE RANGE (IQR)

- The difference between the upper and lower quartiles.
- Gives us a sense of how far out you can go from the mean to get 50% of the data.
- Calculation:  $Q3 - Q1$
- tells us how the data is spread around the median.
- [this article that provides additional examples and explanations for the uses of IQR](#)

Example with Kickstarters, added to last D.S. sheet

The function to calculate the standard deviation of a population in Excel is **STDEV.P**. (The other option is **STDEV.S**, which calculates the standard deviation based on a sample of the whole population. There's a subtle difference between these formulas (one is for the entire population of a dataset while the other is for a sample of the whole) that statisticians care about, but we're going to ignore it. Don't tell any of your statistician friends.

## OUTLIERS

How do we define an outlier? We can use the tools we've just learned along with some guidelines generally accepted by statisticians. There are two main techniques for determining outliers, and each technique uses a measure of central tendency and a measure of spread. 2 options:

- mean + standard deviation
  - ⇒ Any value that is more than 3 standard deviations +/- than the mean
- median + interquartile range (IQR)
  - PREFERRED OPTION b/c Medians and quartiles are **robust statistics** (less sensitive to outliers)
  - ⇒ Any value greater than the upper quartile + (1.5 x IQR)
  - ⇒ Any value less than the lower quartile - (1.5 x IQR)

Why don't we use variance to determine outliers? We use standard deviation because taking the square root of the variance standardizes the "units" of the variance to match the "units" of the dataset. (This is also why it's called "standard" deviation.)

Consider a county with a small population of people making a modest living. Now, imagine a billionaire moves into the county. The median income would barely change, if at all, but the mean would catapult to a much higher value. In fact, if the county is small enough, everyone but the billionaire could end up being "below average" based on the mean.

So if the IQR rule is preferred, why is there a method that uses mean and standard deviation to determine outliers? For one thing, mean and standard deviation can be calculated more quickly. Finding percentiles requires sorting the data, which can be time-consuming with large datasets. The mean and standard deviation can be calculated without sorting data, which means that our computers won't need to work as hard to perform the calculations.

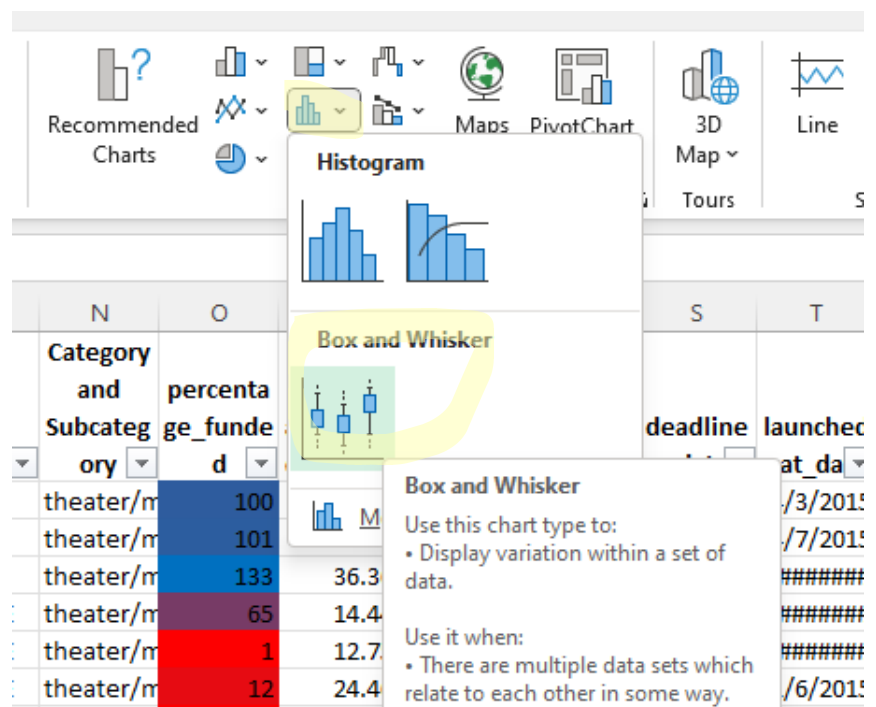
Now that we can identify outliers, what do we do with them? This is a tricky question. Changing or removing data points changes the story you're trying to tell with your data. If the identified outliers are a mistake (say, the data was entered with a typo), ideally, we would just want to correct the mistake and leave the data point in the dataset; if that's not possible, we would have to throw out the data point. However, if an outlier is a legitimate member of the dataset, it's better to leave it in and tell the full story of the data.

## BOX & WHISKER PLOTS

an effective way to show large amounts of information about a distribution in a small amount of space.

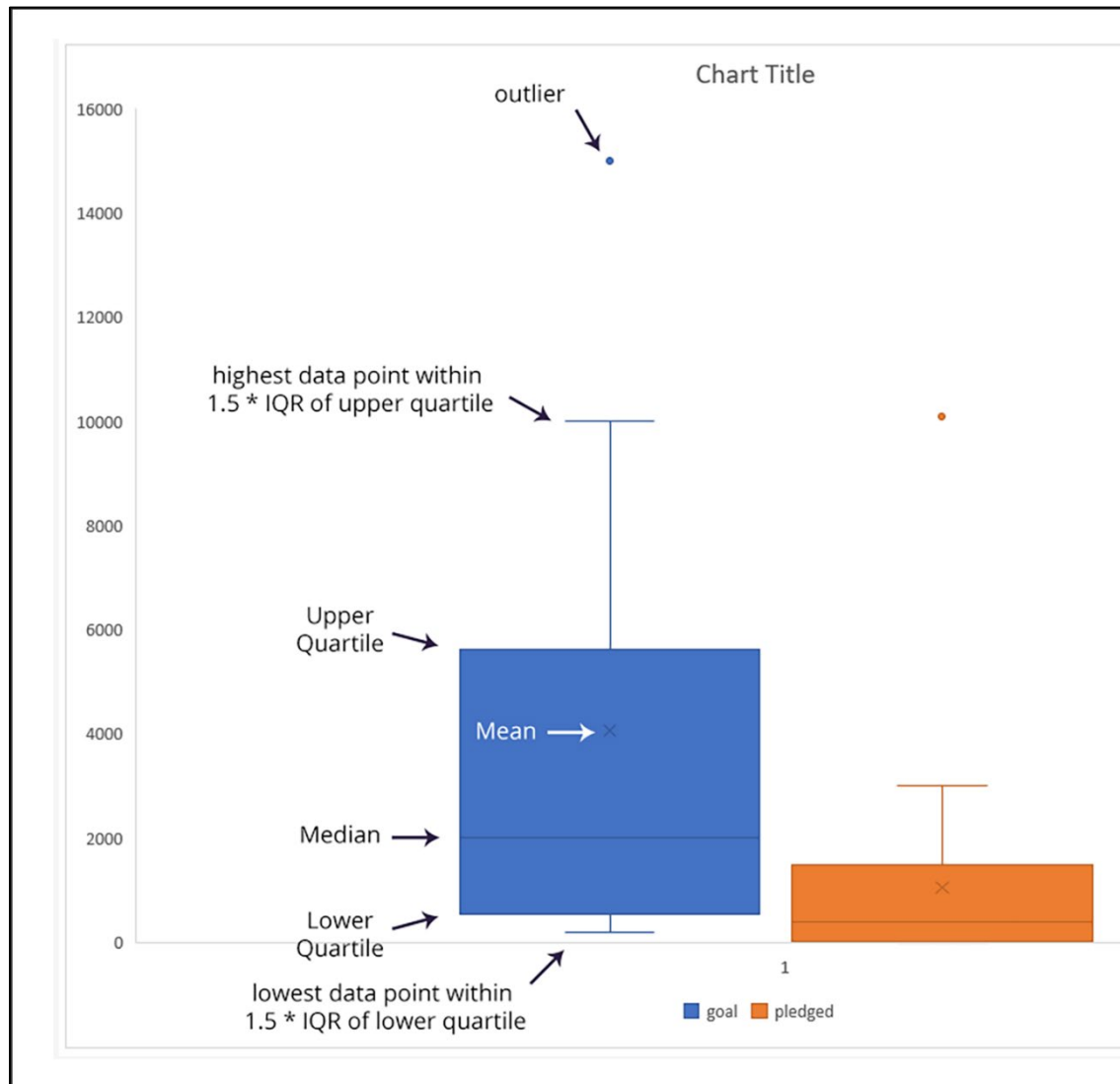
EX: compare the distribution of campaign goals and the distribution of total amounts pledged for plays in Great Britain. Remember that Louise estimates she'll need to raise £4,000 for her future project.

1. Filter for GB & musicals
2. Select Goal & Pledge columns
3. Insert: Charts >
4. Move chart to it's own sheet: click on chart > Chart Design: Location > Move Chart



How do we read a box and whisker plot? The

- box shows the interquartile range with a line for the median and an "X" to indicate the mean.
- whiskers show the extreme values within 1.5 times the interquartile range
- Outliers are represented by labeled dots



From these plots, we can see that the mean campaign goal is around £4,000. This is outside of the range of outliers for amount pledged, so Louise should probably try to get her play produced for less than £4,000. Half of the campaign goals are less than £2,000, which is just over the 3rd quartile for amounts pledged.



