

Data Analytics

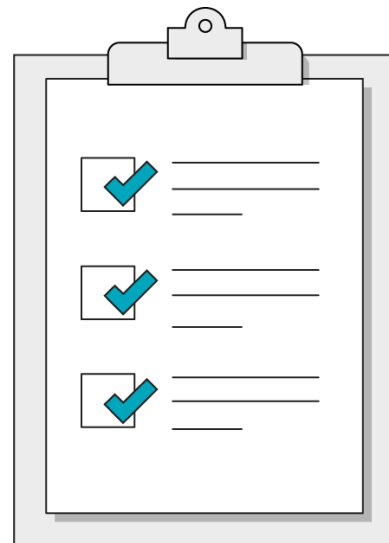


# Data Cleaning and Formulas

# Our Learning Goals

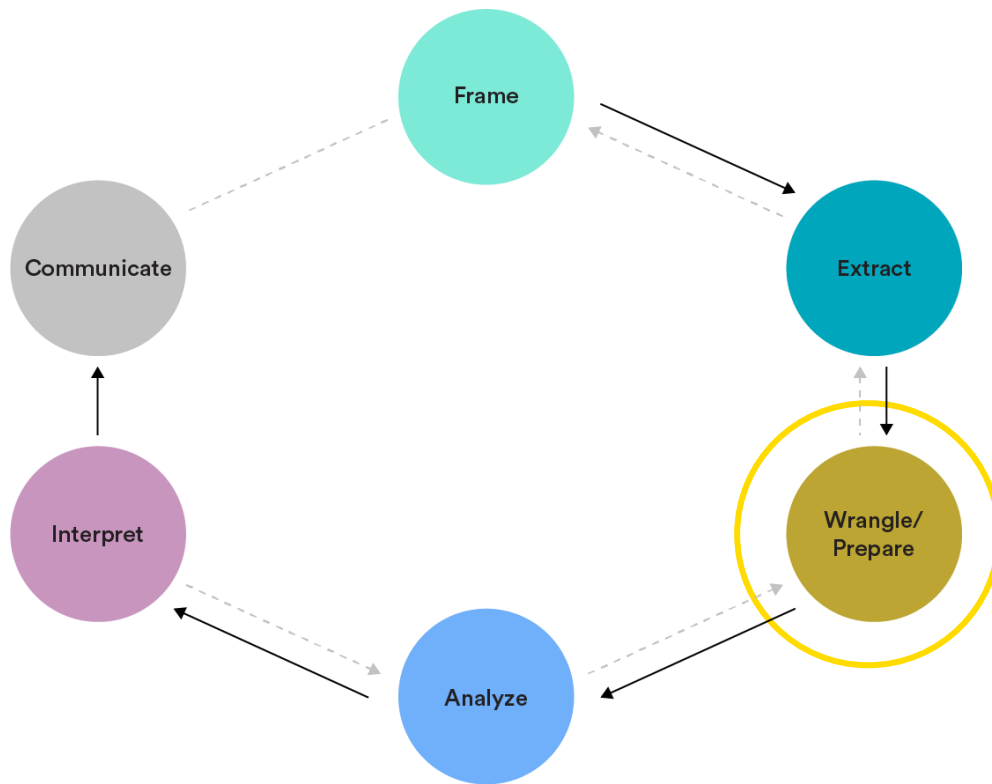
In this lesson, we'll:

- Apply data cleaning best practices, including working with NULLs.
- Conduct exploratory analyses.
- Experiment with common Excel formulas.



# Where We Are in the DA Workflow

**Wrangle/Prepare:**  
Clean and prepare  
relevant data.





## Discussion:

# Getting to Know Your Data

The Superstore regional sales director from the central U.S. region has reached out to you with a request:

**We are seeing a high volume of returns. Can you dig into what might be causing this?**

What should we look into first?



1. Download Lesson 06\_Superstore workbook and examine the “Orders” and “Returns” sheets.
2. Discuss with a partner which data points we should examine to determine why return volume has increased.
3. Then, discuss where we'll need to dig in to explain the higher volume of returns.

Data Cleaning and Formulas

---

# Importing Data for Excel Best Practices



# Importing Data | Getting Your Sandbox Ready

iGPGGA,150104.000,3957.1213,N,07511.4297,W,1,5,1.63,37.5,M,-33.9,M,,\*4E\$GPGGA,150105.000,3957.1299,N,07511.4111,W,1,5,1.63,37.5,M,-33.9,M,,\*56\$GPGGA,150118.000,3957.1296,N,07511.4123,4138,W,1,5,1.63,37.5,M,-33.9,M,,\*5B\$GPGGA,150132.000,3957.1298,N,07511.4138,W,1,5,1.63,37.5,M,-33.9,M,,\*5E\$GPGGA,150146.000,3957.1298,N,07511.4138,W,1,5,1.63,37.5,M,-33.9,M,,\*5C\$GPGGA,150200.000,3957.1293,N,07511.4137,W,1,5,1.63,37.5,M,-33.9,M,,\*5A\$GPGGA,150201.000,3957.1288,N,07511.4126,W,1,5,1.63,37.5,M,-33.9,M,,\*52\$GPGGA,150215.000,3957.1289,N,07511.4132,W,1,5,1.63,37.5,M,-33.9,M,,\*51\$GPGGA,150229.000,3957.1295,N,07511.4132,W,1,5,1.63,37.5,M,-33.9,M,,\*5B\$GPGGA,150243.000,3957.1242,N,07511.4135,W,1,5,1.63,37.5,M,-33.9,M,,\*5E\$GPGGA,150257.000,3957.1152,N,07511.4124,W,1,5,1.63,37.5,M,-33.9,M,,\*5E\$GPGGA,150311.000,3957.1052,N,07511.4138,W,1,5,1.63,37.5,M,-33.9,M,,\*59\$GPGGA,150312.000,3957.1052,N,07511.4163,W,1,5,1.63,37.5,M,-33.9,M,,\*5A\$GPGGA,150340.000,3957.0935,N,07511.4218,W,1,5,1.63,37.5,M,-33.9,M,,\*51\$GPGGA,150354.000,3957.0891,N,07511.4162,W,1,5,1.63,37.5,M,-33.9,M,,\*50\$GPGGA,150408.000,3957.0810,N,07511.4024,W,1,5,1.63,37.5,M,-33.9,M,,\*50\$GPGGA,150409.000,3957.0707,N,07511.3940,W,1,5,1.63,37.5,M,-33.9,M,,\*5D\$GPGGA,150423.000,3957.0702,N,07511.3856,W,1,5,1.63,37.5,M,-33.9,M,,\*59\$GPGGA,150437.000,3957.0643,N,07511.3851,W,1,5,1.63,37.5,M,-33.9,M,,\*59\$GPGGA,150451.000,3957.0632,N,07511.3849,W,1,5,1.63,37.5,M,-33.9,M,,\*55\$GPGGA,150505.000,3957.0608,N,07511.3742,W,1,5,1.63,37.5,M,-33.9,M,,\*5E\$GPGGA,150519.000,3957.0625,N,07511.3618,W,1,5,1.63,37.5,M,-33.9,M,,\*57\$GPGGA,150520.000,3957.0637,N,07511.3457,W,1,5,1.63,37.5,M,-33.9,M,,\*5B\$GPGGA,150534.000,3957.0639,N,07511.3453,3459,W,1,5,1.63,37.5,M,-33.9,M,,\*52\$GPGGA,150548.000,3957.0680,N,07511.3460,W,1,5,1.63,37.5,M,-33.9,M,,\*A\$GPGGA,150559.000,3957.0730,N,07511.3567,W,1,5,1.63,37.5,M,-33.9,M,,\*57\$GPGGA,150609.000,3957.0799,N,07511.3612,W,1,5,1.63,37.5,M,-33.9,M,,\*5A\$GPGGA,150610.000,3957.08



# Data Set Best Practices | Resave

If you plan to analyze data in Excel,  
always and immediately

**convert .CSV files to .XLSX**

- Go to File >> Save As

But why?

**CSV** (comma-separated values) is plain text,  
while **XLSX** is a binary file format that holds  
information — including both content and  
formatting — on all the worksheets.

## Excel Workbook (.xlsx)

### Common Formats

- Excel 97-2004 Workbook (.xls)
- ✓ CSV UTF-8 (Comma delimited) (.csv)
- Web Page (.htm)
- Excel Template (.xltx)
- Excel 97-2004 Template (.xlt)
- PDF

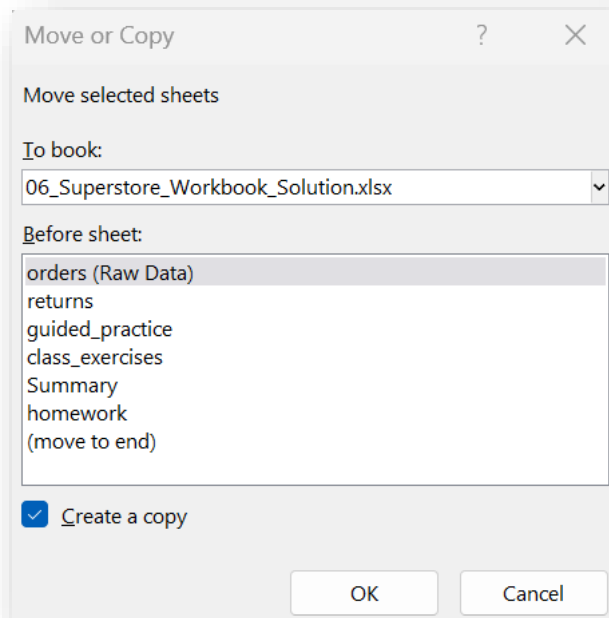
### Specialty Formats

- Excel Macro-Enabled Workbook (.xlsm)
- Excel Binary Workbook (.xlsb)
- Single File Web Page (.mht)
- Excel Macro-Enabled Template (.xltm)
- Tab delimited Text (.txt)
- UTF-16 Unicode Text (.txt)
- Excel 2004 XML Spreadsheet (.xml)
- Microsoft Excel 5.0/95 Workbook (.xls)
- Comma Separated Values (.csv)
- Space Delimited Text (.prn)
- Macintosh Formatted Text (.txt)
- MS-DOS Formatted Text (.txt)
- Macintosh Comma Separated (.csv)
- MS-DOS Comma Separated (.csv)
- Data Interchange Format (.dif)
- Symbolic Link (.slk)
- Excel Add-in (.xlam)
- Excel 97-2004 Add-in (.xla)
- Strict Open XML Spreadsheet (.xlsx)
- OpenDocument Spreadsheet (.ods)



# Data Set Best Practices | Rename

- Rename the sheet that contains the orders data “**orders (Raw Data).**”  
Make a copy of that sheet by right clicking on the sheet’s tab, and choosing **Move or Copy.**
- In the window that appears, check off the box next to “**Create a copy.**”
- Hit “OK” and rename the copied sheet “**orders (Clean Data).**”





## Computers Out: Data Set Best Practices

Let's do this together! Open up the lesson workbook and...

1. Document **ALL of the steps** you take in your analysis.
2. Create a **working summary sheet** that includes the following:
  - a. A directory of other sheets.
  - b. An explanation of analysis.
  - c. A short summary of your results.

Be sure to update this sheet regularly!

	A	B	C	D
1	<b>Superstore Data Summary</b>			
2				
3	<b>Sheets</b>		<b>Analysis Summary</b>	
4	<a href="#">Raw Data</a>		Data: Orders, returns, regions, people, products, customers	
5	<a href="#">Cleaned Data</a>		Goal: Discover why sales are up but profits are down	
6	<a href="#">Data Summary</a>			
7	<a href="#">Pivot Tables</a>			
8	<a href="#">Charts</a>			
9	<a href="#">Dashboard</a>			
10				
11	<b>Cleaning Steps</b>		<b>Analysis Steps</b>	
12	Removed null values from Order ID		Create pivot table of sales and regions	
13	Standardized city names using find/replace		Apply conditional formatting to review for outliers	
14	Created table		Create pivot table of sales, profit and people	
15			Insert bar chart to compare sales and profit by people	
16				
17				
18				
19				
20				
21				
22				

Summary Raw Data Clean Data Data Summary Pivot Tables Charts Dashboard

Data Cleaning and Formulas

---

# Strategies for Cleaning & Preparing Your Data



# Data Cleaning

Data cleaning is the process of assembling data into a **usable format for analysis**.

Common data cleaning actions include:

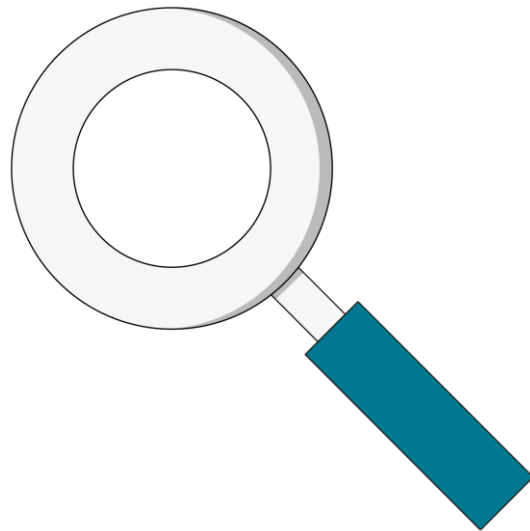
- **Reformatting dates** so that Excel recognizes them as dates.
- **Extracting day/hour/month/year from a date** to aggregate by those categories.
- **Removing duplicate values** or rows.
- **Combining data sources** into one table.
- **Concatenating or separating data**.



# NULLs

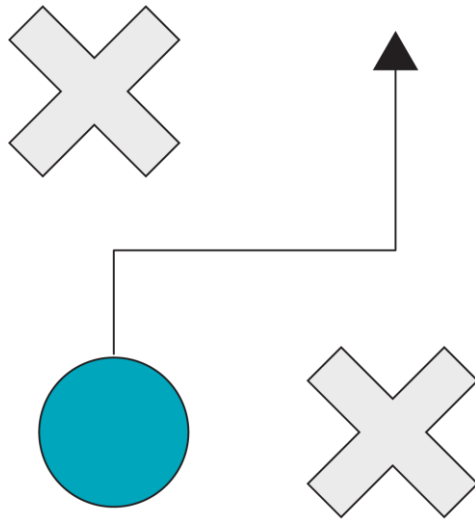
A **NULL value** is **any missing value** in your data.

One common way of conceptualizing a NULL value is thinking of it as “**empty**” — not 0, not the word “NULL,” just empty!



# Four Primary Strategies for Handling NULLs

1. **Find missing values** (using reference resources).
2. **Ignore them** (some may have meaning).
3. **Impute values** (e.g., median or zeros).
4. **Delete them** (only with caution).





Take a look at your profit value for

**Row 2.**

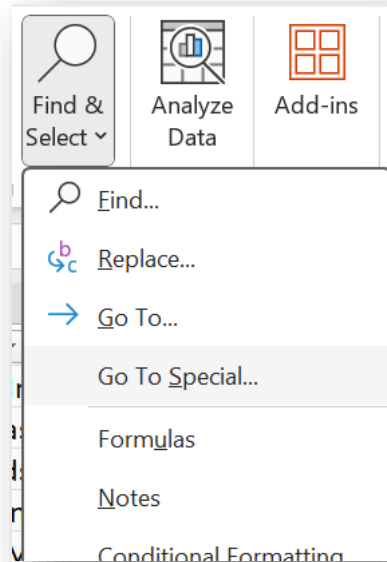
Should this be 0?

Share your answer and reasoning with the class.

	K	L	M	N	O	P
	sales	profit	quantity	discount	profit_ma	region_id
	\$ 95.80	\$ 0.33	4	0.4		1488
	\$ 3.43	\$ 0.33	1	0.4	9.62%	1488
	\$ 95.80	\$ 0.33	4	0.4	0.34%	1488
	\$ 95.80	\$ 0.33	4	0.4	0.34%	1488
	\$ 42.48	\$ 0.28	1	0.35	0.66%	1488
	\$ 95.80	\$ 0.33	4	0.4	0.34%	1488
	\$ 59.37	\$ 0.28	1	0.35	0.47%	1488
	\$ 59.37	\$ 0.28	1	0.35	0.47%	1488
	\$ 6.43	\$ 0.24	2	0.3	3.73%	1488
	\$ 6.97	\$ 0.28	1	0.35	4.02%	1488
	\$ 6.43	\$ 0.24	2	0.3	3.73%	1488

# Finding and Replacing Blanks

- In the “Home” menu, choose the “**Find & Select**” button.
- Click “**Go to Special...**”
- Select the “**Blanks**” radio button and hit OK.
- **Don't click anything!** Just type a “-”.
- Then hold down the control key (same for Mac users) and tap “**enter**,” and all of the blank cells should now be filled in with dashes.





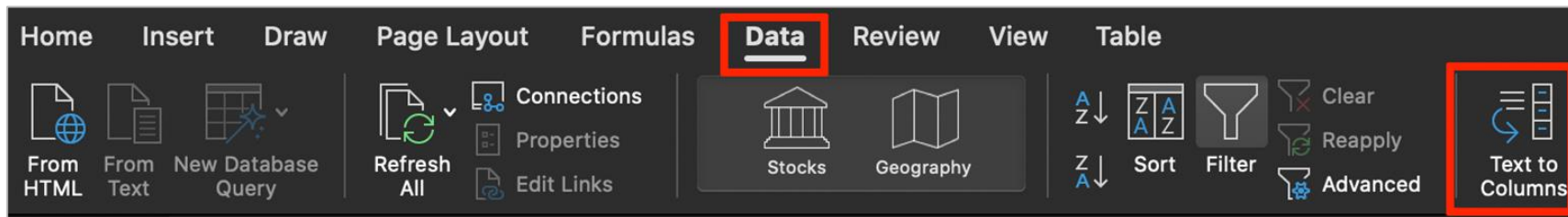
# Text to Columns

What if we hypothesized that there might be a difference in **sales or profit between states?**



Right now, we can't complete that analysis because **city and state** are lumped together. We can fix this, however, using Excel's "**Text to Columns**" feature!

# Text to Columns | Step by Step



**Step 1:** Right click on the column to the right of the “**city\_state**” column (it should be “**sub\_region**”) and choose “Insert” to insert a new blank column to the right of “city\_state.”

**Step 2:** Click on the “city\_state” header to select the entire “city\_state” column.

**Step 3:** Select the “Text to Columns” button in the “Data” menu on the ribbon.

# Text to Columns | Step by Step

**Step 4:** Choose delimited. Then, click “Next” and check off “Comma.” Click “Finish.”

**Step 5:** If it gives you an error saying it will replace data, hit “OK.”

**Step 6:** Rename the “city\_state” column to just “city,” and the second column to “state.”

Convert Text to Columns Wizard - Step 2 of 3

This screen lets you set the delimiters your data contains.

Delimiters

☐ Tab ☐ Treat consecutive delimiters as one

☐ Semicolon

☒ Comma

☐ Space

☐ Other:

Text qualifier: "

Preview of selected data:

city_state	
Henderson	Kentucky
Henderson	Kentucky
Los Angeles	California
Fort Lauderdale	Florida
Fort Lauderdale	Florida
Los Angeles	California
Los Angeles	California

Cancel < Back Next > Finish

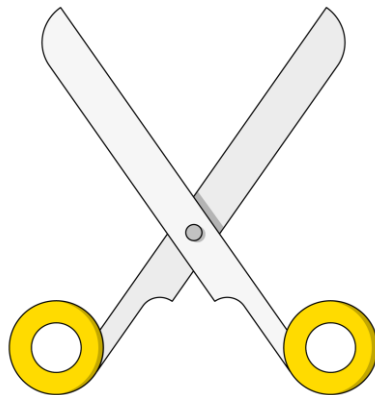
## Text to Columns | Trimming the Spaces

Oh no! The space transferred over with the state name.  
Let's clean this up:

**Step 1:** Insert another column to the right of the state column;  
name this new column “state\_trimmed.”

**Step 2:** Use the TRIM function to take out the extra space in front.

**=TRIM(text)**





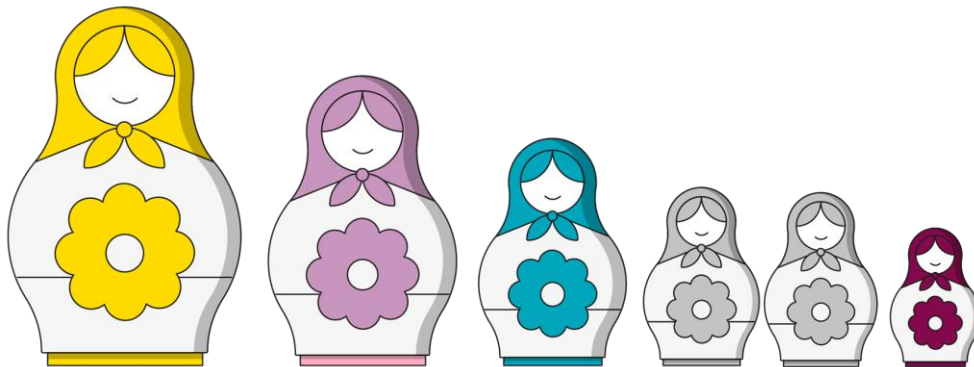
## Discussion: Checking for Duplicates

2 minutes



Finally, let's check for duplicates!

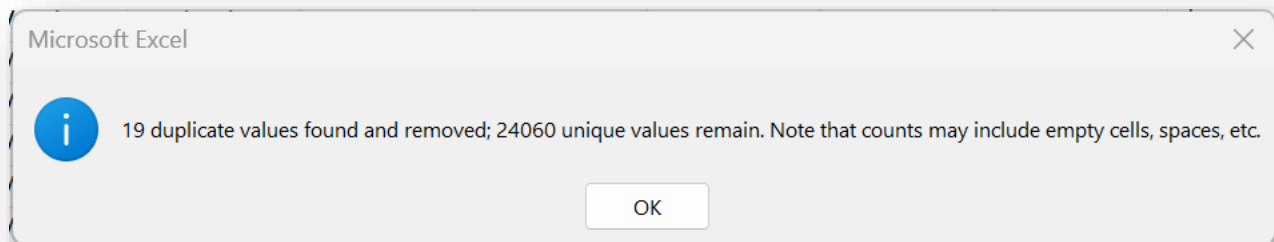
What would be an indicator of a duplicate in our data set?



# Checking for Duplicates | Step by Step

**Step 1:** Click on “Remove Duplicates” from the “Data” menu in the ribbon.

**Step 2:** Uncheck “Select All.” Then, check off ONLY “order\_id\_number” (Column B) and “product\_id” (Column F) before clicking “OK.”



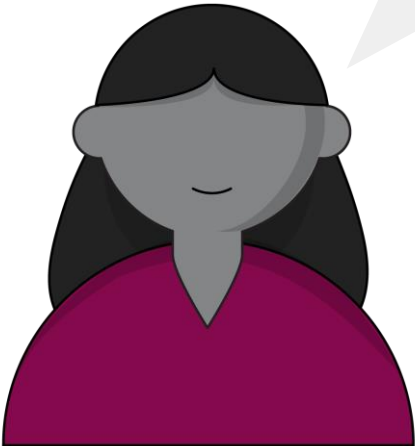
Data Cleaning and Formulas

---

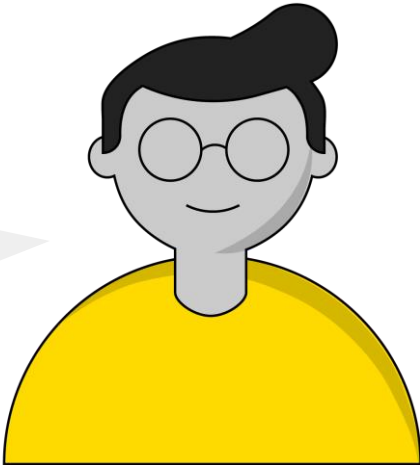
# Asking the Right Questions (of Your Data)



# Asking the Right Questions



What insights about returns can be gained from the Superstore data set?



Hmm, this question is really broad. Let me explore the data set first.



# Exploratory Analysis | Best Practices

As part of an exploratory analysis, you should ALWAYS determine:

- **The number of rows** in the data set.
  - **What each row represents** in the data set — a unique *what*.
- **The number of columns** in the data set.
  - What **each column represents** and **how that data was collected**. *Try getting a data dictionary!*





Computers Out:

# Getting to Know the Superstore Data Set

5 minutes



Take five minutes to explore the columns in the Superstore data set and consider the following:

- How was the data for each column collected?
- What are the units of each column?
- According to the data dictionary, what does this column represent?



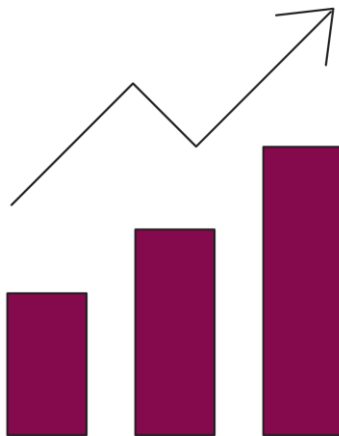
Be prepared to share some insights with your class!



# Exploratory Analysis | Definition

In a nutshell, exploratory analysis means “**getting to know**” a **data set**, which can include:

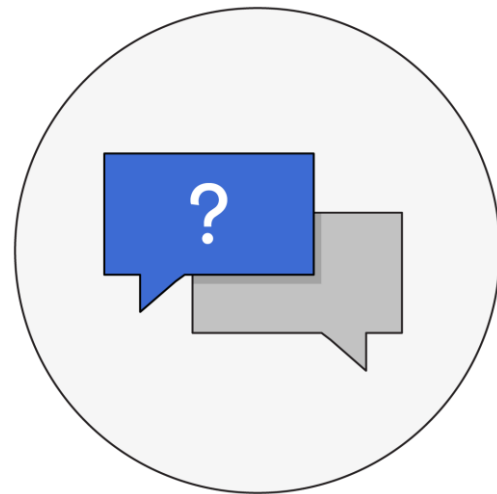
- **Reviewing columns’ names.**
- **Obtaining aggregate metrics** for number columns (average, sum, min, max, etc.).
- **Creating PivotTables** to view the unique values that can appear in a given text column.
- **Crafting preliminary visualization.**



# From Questions to Hypotheses

Start by asking yourself...

- What **fields** can I **COMBINE** to find interesting insights?
- What **ACTIONS** can **someone take** as a result of my charts and analyses?



# From Questions to Hypotheses | Examples



**Good example:** If we look at profit and ship mode together, we might discover that certain ship modes are consistently associated with lower profits. Result/action: We might recommend that Superstore stop offering those ship modes to customers in order to boost profits.

**Bad example:** Sales and order\_id. We can get the average dollar amount per item in an order\_id; for example, the average cost of a product in order 123 was \$15. But that doesn't really lead to many useful insights for the store. An aggregate of the average order amount across all orders or particular categories might be more useful.





Let's brainstorm questions we can ask about the Superstore data set together.

What might be some interesting variables to combine to gain meaningful insights?

Formulate them into a hypothesis and call out your response over the microphone or share it with your class.



Keep in mind that while there are instructions for the homework, some of the prompts are **intentionally vague**. You'll have to complete this exercise for your first project data set!



Let's revisit the business problem from earlier: **We are seeing a high volume of returns.** Now that you've identified the data points you need, open the lesson worksheet and work with your partner to:

1. Identify the questions you can ask to help gain interesting insights from the data.
2. Then, formulate your questions into a hypothesis. Here's an example:  
"If we compare the shipping cost and the order priority, we might find that high shipping costs for low-priority orders frequently lead to returns."
  1. List it out in your worksheet.
  2. Be prepared to share your work with your class.
3. Get month from order date, and day from ship date

Data Cleaning and Formulas

---

# Introduction to Excel Functions





# Data Referencing

**Referencing**, in its basic form, means pulling the value of one cell into another cell.

fx	=A1	
	A	B
1		
2	=A1	
3		
4		

A2 references A1

# Cell Referencing

An absolute reference is a **fixed (locked) location** in a worksheet.

fx	=A1
	A
1	
2	=A1

Relative

fx	=\$A1
	A
1	
2	=\$A1

Mixed  
Column Only

fx	=A\$1
	A
1	
2	=A\$1

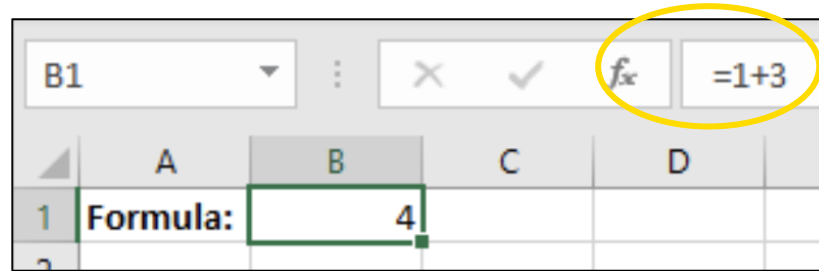
Mixed  
Row Only

fx	=\$A\$1
	A
1	
2	=\$A\$1

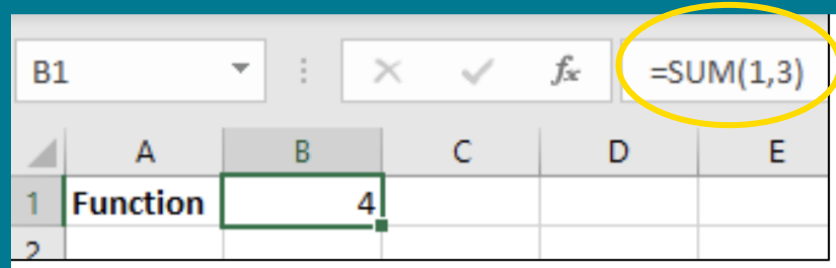
Absolute

# What Is a Formula in Excel?

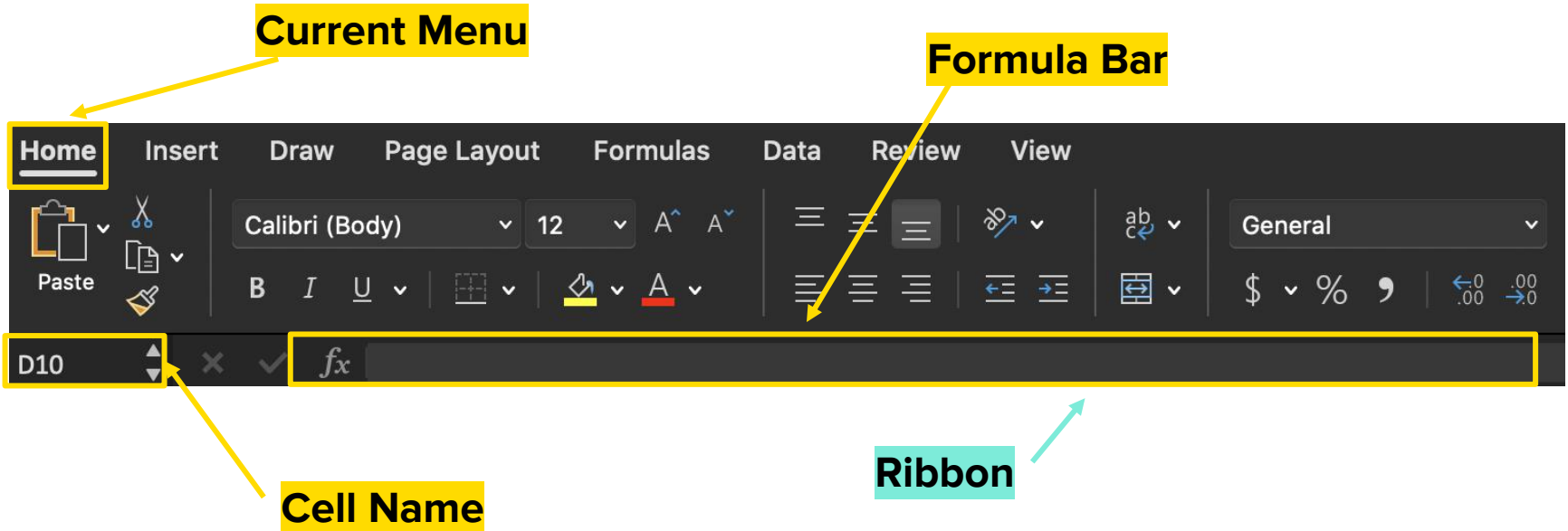
A **formula** is an expression which calculates the value of a cell.



**Functions** are predefined formulas that are already available in Excel.



# Navigating Formulas and Functions in Excel



# The Anatomy of an Excel Function

All functions start with the **equals (=)** sign.

**=LEFT(A2,4)**

The **name** of the function.

The **arguments** (inside the parentheses) that the function requires. Arguments are separated by commas.

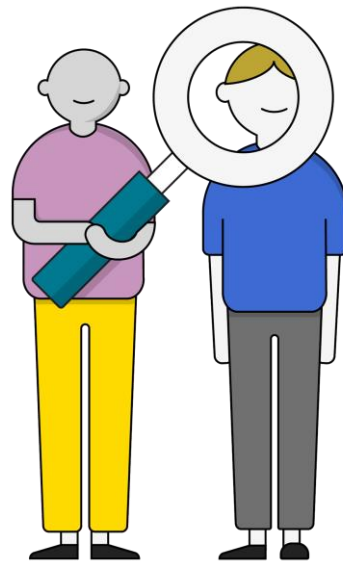
# Finding the Right Function

The typical workflow used by data analysts is:

**Step 1:** Google the task you are trying to accomplish.

**Step 2:** Find the name of the function (or functions!) you need in the search results.

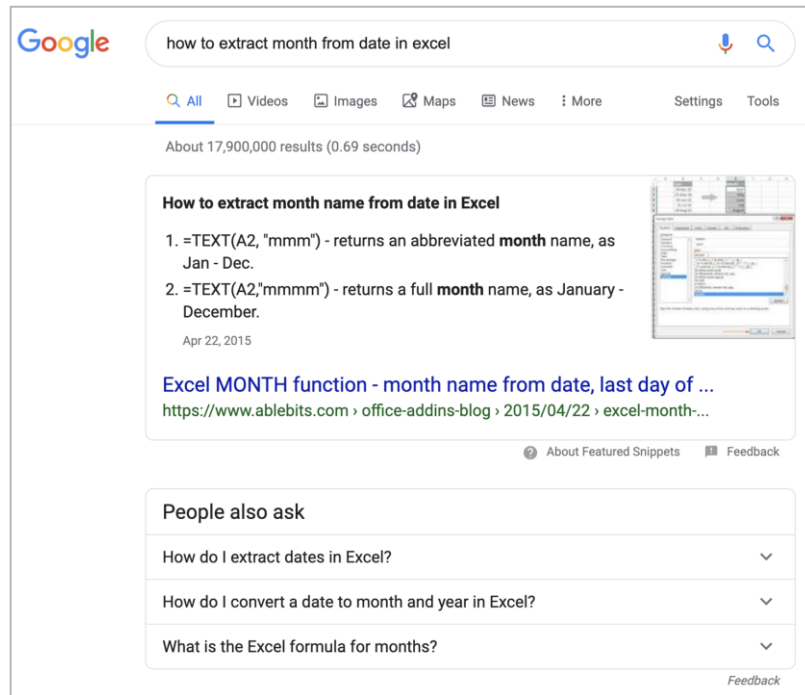
**Step 3:** Go to the [Microsoft Excel documentation](#) to learn how to implement the function and see examples.



# Finding the Right Function | Google It

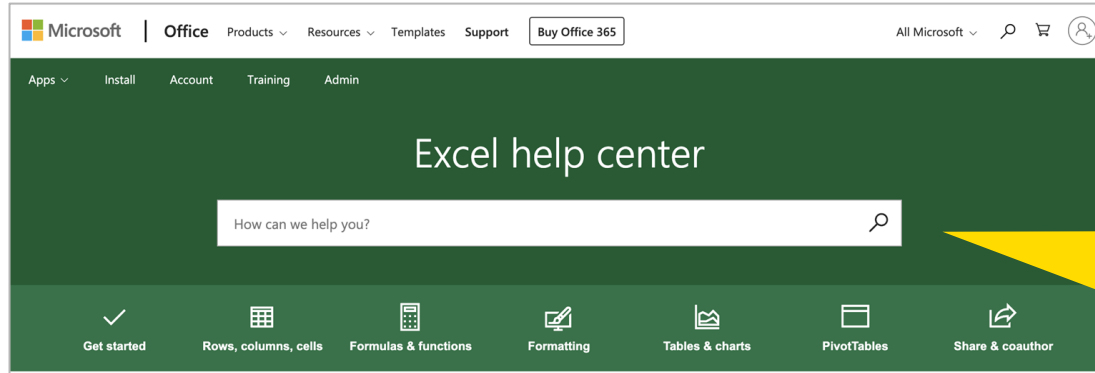
If you didn't already know the function for extracting months from dates in Excel, here is an example of how you'd phrase your Google search:

**“How to extract month from date in Excel.”**



# Finding the Right Function | MS Documentation

## [Excel help and learning on Microsoft Support](#)



Type **TEXT function** into the search box. One of the first results should be the page for the TEXT function.

## Results for "Text Function Excel"

### TEXT function

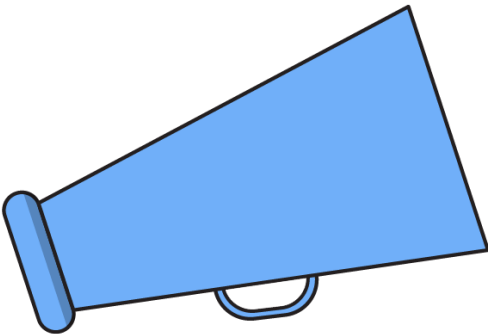
This is where the TEXT function is invaluable, because it allows you to force Excel to format the values the way you want by using a format code, like "MM/DD/YY" for date format. In the following example, you'll see what happens if you try to join text and a number without using the TEXT



How many arguments does the TEXT function require?

**Syntax**

**TEXT**(value, format\_text)



## Finding the Right Function | Arguments

Great! So we know that our function takes this form:

```
=TEXT(argument1, argument2)
```

Now, let's figure out what `argument1` and `argument2` are.

## TEXT Function | First Argument

**=TEXT(argument1, argument2)**

The MS documentation tells us that the first argument is “**Value you want to format.**”

So, what is it that we want to format?

We want to format each date in the **order\_date** column! To do so, we need to start with the **first order date.** Then, we can drag the formula down to calculate the rest. Thus, the first argument of our function will be **C2**.

## TEXT Function | Second Argument

**=TEXT(argument1, argument2)**

According to the MS documentation, the second argument is “**Format code you want to apply.**” We need to figure out what these format codes are.

Scroll down on the page. Do you see a section that might give us more details? Call out when you find it!

# Getting the Info We Need

## Format codes by category

Following are some examples of how you can apply different number formats to your values by using the **Format Cells** dialog, then use the **Custom** option to copy those **format codes** to your **TEXT** function.

Dates



Select “Dates” from the drop-down menu.

To get the full name of the month, we need to use **“mmmm.”**

	To display	As	Format	Formula
5				
6	Months	1–12	"m"	=TEXT(B3,"m")
7	Months	01–12	"mm"	=TEXT(B3,"mm")
8	Months	Jan–Dec	"mmm"	=TEXT(B3,"mmm")
9	Months	January–December	"mmmm"	=TEXT(B3,"mmmm")
10	Months	J–D	"mmmmm"	=TEXT(B3,"mmmmm")
11	Days	1–31	"d"	=TEXT(B3,"d")
12	Days	01–31	"dd"	=TEXT(B3,"dd")
13	Days	Sun–Sat	"ddd"	=TEXT(B3,"ddd")
14	Days	Sunday–Saturday	"dddd"	=TEXT(B3,"dddd")



# Our First Cleaning Function

Are you ready to clean some data? Let's get to it!

1. Open up your orders worksheet and add an "order\_month" column to the right of "order\_date."
2. Apply one of the functions to the Superstore data set:
  - `=TEXT(C2, "mmmm")`
  - `=TEXT([@[order_date]], "mmmm")`



**Best practice reminder:** Put all formulas to the right side of your data set; don't mix them in with the raw data.



# So, What's Really Going on With Returns? Part 1

To dive deeper into why Superstore is seeing a high volume of returns, we need to take a closer look at orders, profit, and sales as well as individual customers.

It's a lot to look at! But don't worry, we'll do this together, step by step. First, let's find out if some days of the week see higher volumes in sales and returns.

1. To extract the day of the week from the **order\_date**, write out:

**=TEXT(C2, "dddd")** OR **=TEXT([@[order\_date]], "dddd")**



COUNTIF is another useful function for data cleaning. It can be used to:

- Count the number of cells in a range that contain specific data.
- Tell us whether or not a single cell contains data based on a condition.

When there is a single cell in the COUNTIF range, the maximum that can be returned is 1 and the minimum that can be returned is 0.

Syntax:

**COUNTIF(range cell, condition)**





Let's use COUNTIF to return a 1 or 0 to help us figure out **whether or not a discount is more than our imposed limit of 30%.**

1. Open up the "Orders" sheet.
2. Insert a column to the right of the "**Discount**" column called "**discount\_over\_30.**"
3. Enter `=COUNTIF(P2, ">=.3")`.

We can now **SUM** this column to find out the number of orders that were discounted more than 30%.



## So, What's Really Going on With Returns? Part 2

2. Looking at profit, does profit margin impact whether or not something gets returned? To find out, recalculate the profit margin (**profit** divided by **sales**) per row. **Insert a new column next to profit in Column N.**

**=N2/M2 or ==[@profit]/[@sales]**

Next, we will use IFERROR to *wrap* the formula. We do this to help us deal with NULLs in the data set.

**=IFERROR(formula, "")**



## So, What's Really Going on With Returns? Part 3

3. Now, let's look at individual customers to see if some customers return more than others. You need to concatenate the `order_info_id` and the `order_id_number` with a dash in between them to create just a `order_id` column.

Write out:

```
=[@order_info_id]&"-"&[@order_id_number] OR =A2&"-"&B2
```



## So, What's Really Going on With Returns? Part 4

4. Finally, let's decipher sales volume! To help us categorize our sales without relying on the exact dollar amount, we'll categorize sale amounts above \$500 as "High" and below \$500 as "Low."

`=IF([@sales] > 500, "High", "Low")` OR

`=IF(M2>500, "High", "Low")`

Now that you have sales categorized, does it make a difference to returns?

Data Cleaning and Formulas

---

# Wrapping Up





Solo Exercise:

## Optional Homework

Finish the **homework** tab to practice your new formula skills.

1. Create a new “item\_size” column to categorize items as large or small.
2. Create a new “days\_to\_ship” column to see how many days it took to ship each item.
3. Create a new “top” customers column that identifies customers in the given list.



# Recap

## Today in class, we...

- Applied data cleaning best practices, including working with NULLs.
- Conducted exploratory analyses.
- Experimented with common Excel formulas.

# Looking Ahead

## Homework

- Refer to the "Homework" tab.

**Up next:** Referencing and Lookups



# Additional Resources

- [Excel File Setup for Analysis](#)



