BNP Paribas Cardif x GA

Data Analytics
17th to 21st March 2025

GENERAL ASSEMBLY

WELCOME!

We're excited you're here.

GENERAL ASSEMBLY AT A GLANCE

General Assembly is at the forefront of tech talent development specializing in today's most in-demand skills. The leading source for upskilling, reskilling, and revolutionizing teams, GA programs offer proven impact on the job, both for the employee and the employer.

CONTENT AREAS

- Engineering & Development
- Data & AI
- Product & UX
- Digital Marketing

FLAGSHIP CAMPUSES

FAST FACTS

- 97,000+ alumni, 19,000+ hiring partners
- 25,000+ employees of Fortune 500 companies trained in 20+ countries
- Flagship campuses in NY and Singapore + global virtual learning

General Assembly is a brand of the **Adecco Group**, the world's foremost provider of staffing, career transition, and talent development solutions.

Introductions

Meet Your Instructors

Meet General Assembly

Your Instructor

Jack Tyler-Whittle

Data Instructor and Learning Experience Designer

- Teach Data Science, Data Analytics and Python courses.
- Design, create and update General Assembly's data courses.
- Was a Data Scientist at a startup, designing and building natural language processing (NLP) data products, creating the APIs to make them accessible, and maintaining the database to store the data.
- Background in digital transformation and data analytics in Telecoms (Vodafone) and Financial Services (Capital One, HSBC).
- MBA from London Business School.

Meet General Assembly

Your Instructor

Harry Long

Senior Data Scientist, UKHSA

- MPhil Philosophy, UCL, MSc Computer Science, UCL, BA Philosophy and French, University of Oxford.
- PhD candidate and Keeling scholar in Philosophy at UCL, focussing on formal epistemology and philosophical methodology.
- French speaker and francophile! Also enjoys riding motorcycles.

Solo Exercise:

Getting to know you

Find an object close to you that would give your colleagues a hint or clue about you!

*If your video doesn't work, type in the Chat Box what you are holding

Example:
My knitting needles

Discussion:

Our Working Agreement

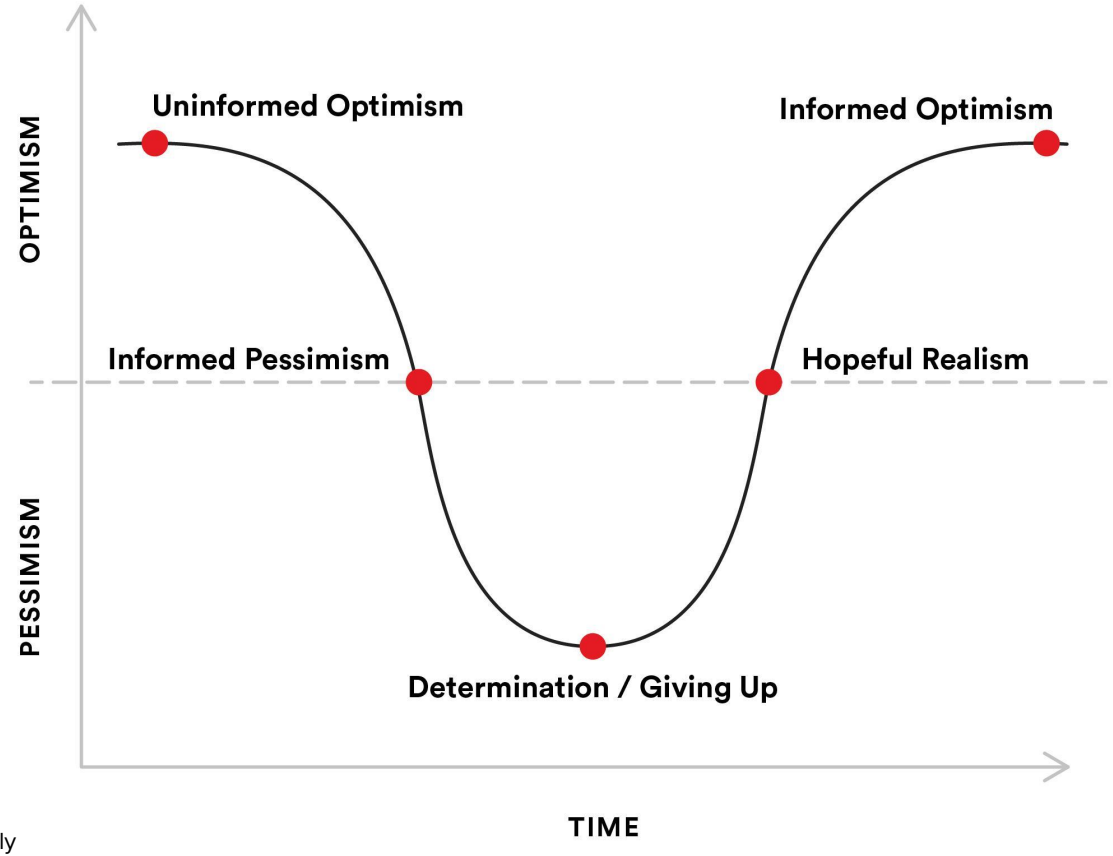
At GA, we create norms for how we'll work together during the course.

Check out the working norms.

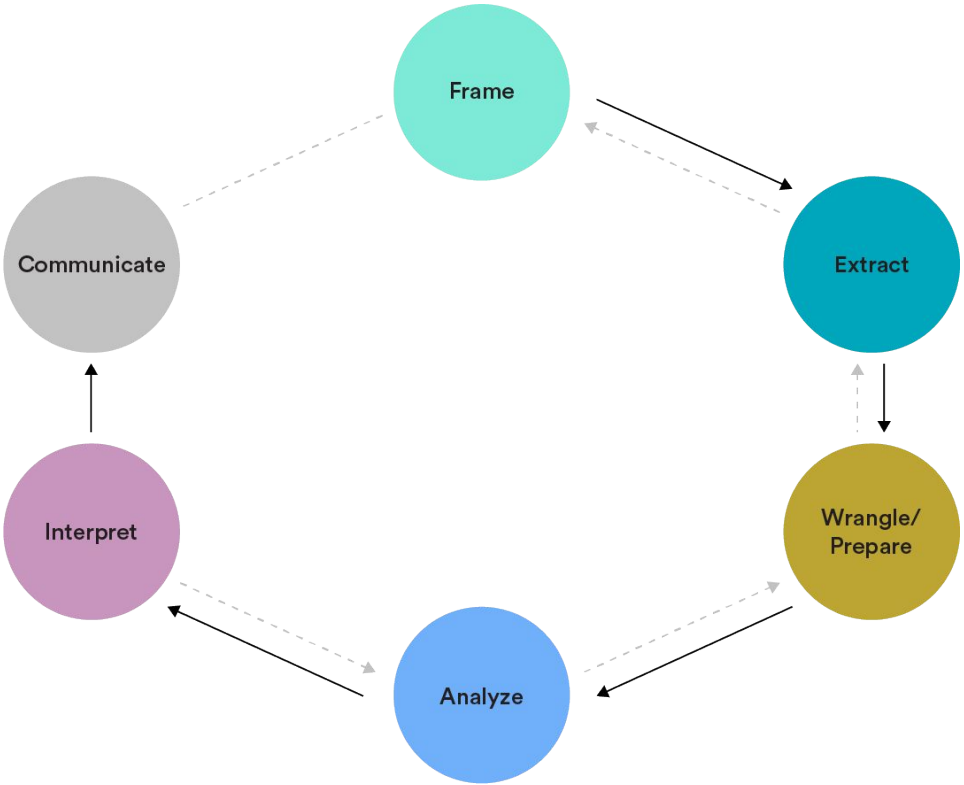
Is there anything else we should add to the list?

Be Present	**Contribute Constructively**	**Work Hard**
Ask Questions	**Be Supportive**	**Talk to Us!**

Buckle Up for the Journey Ahead!



The Data Analytics Workflow



Frame: Develop hypothesis-driven questions for your analysis.

Extract: Select and import relevant data.

Wrangle/Prepare: Clean and prepare relevant data.

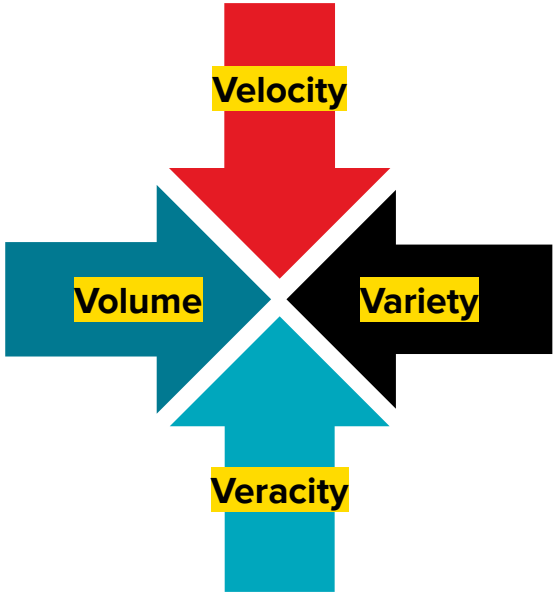
Analyze: Structure, comprehend, and visualize data.

Interpret: Leverage your analysis to make decisions and recommendations.

Communicate: Present data-driven findings and insights in a compelling manner.

The Four Vs of Big Data

High frequency, real-time



Enough data to recognize patterns.

Images, text, numerical data

Good quality

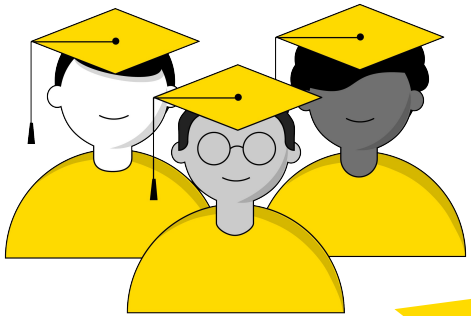
Obtaining good quality data is important

1. Is the data from a trustworthy source?
2. How did the source organisation collect the data?
3. How easy is it to obtain? Is it free to access?
4. Do you have permission to use it?
5. Does the data contain personal or sensitive information?
6. Do we know what the data means? ***Data Dictionary***
7. What are the limitations of the data?

You'll Leave This Course Saying...

"I am no longer intimidated by rows and rows of data! I can combine, clean, and visualize data to gain insights into important business trends."

"I feel empowered to continue learning new techniques and acquiring new ways of working with data."



"Converting numbers into visually appealing, easy-to-understand visualizations is such a creative process. It's a lot of fun!"

Cleaning and Aggregating Data With Excel

Course Logistics

Staying Connected | Teams

- Keep as much course communication as possible within Teams.
- Feel free to use emojis!
- Don’t be afraid to send direct messages.
- Respond to instructor, IA polls, and check-ins (you can do this with an emoji).

Important: *when you connect to the MS Teams space, it’s important to join* ***without switching organisations.*** *This will allow you to access and fully participate in the training session.*

Our (approximate) schedule

Times are in CET	Monday	Tuesday	Wednesday	Thursday	Friday
9 - 10:30 AM	Introduction to GA and Data Analytics	Getting Started With SQL	Advanced JOINs and NULLs	Getting Started with Tableau	Dashboard Design & Stories in Tableau
10:30 - 10:45 AM	Break	Break	Break	Break	Break
10:45 AM - 12 PM	Data Cleaning and Formulas	From Stakeholder Questions to Efficient Queries	SQL Functions	Marks Cards Working with Dates	Practice Application
12 - 1 PM	Lunch	Lunch	Lunch	Lunch	Lunch
1- 2:30 PM	Referencing and Lookups	Combining Data with JOINs and UNIONs	Open Workshop	Connecting to PostgreSQL	Open Workshop
2:30- 2:45 PM	Break	Break	Break	Break	Break
2:45 - 4 PM	PivotTables	Combining Data with JOINs and UNIONs (cont.)	Open Workshop	Show Me Filters Starting a Dashboard	Open Workshop
4 - 4.45 PM	Data Visualization	Practice Application		Visual Analytics	Project Presentation
4.45 - 5 PM	Review & Recap	Review & Recap	Review & Recap	Review & Recap	Happy Half Hour

Notes

- **Break times** are fixed
- **Timeslots for topics** are approximate and will be flexed depending on the needs of the class
- We will go at the pace of the majority of the class
- We are staring with the assumption that you know nothing about Excel, SQL or Tableau

Solo Exercise:

Getting excited!

Post your answer to one or more of these questions in **Teams**

I'm **most** excited to learn about...

The thing I **most** want to get out of this week is...

I'm **most** concerned about...

Tech Check

- Do you have **Excel** installed and ready to go, for today?

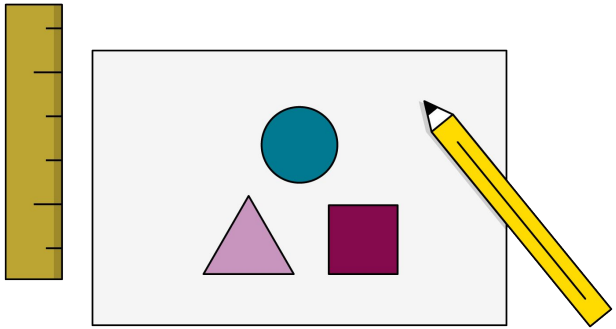
- Can you access this link okay, for **SQL**?
  https://analyticsga-global.generalassemb.ly

- Do you have **Tableau** installed and ready to go for Thursday?

Project Work

In this course, you'll complete **a final project**:

- This will be a group project.
- You will have time on Wednesday and Friday to complete the project.
- Your final project is followed by a brief (up to 10 minute) presentation to your class on Day 5.
- We will give you the data for the project.



What You'll Learn Today

- Apply data cleaning best practices, including working with **NULLs**.
- Conduct exploratory analyses.
- Manipulate data sets using **VLOOKUP and XLOOKUP**.
- Summarize data with **PivotTables**.
- Identify the appropriate visualization types for a given data set.
- Create analytics visuals such as bar charts, pie charts, line graphs, histograms, and scatterplots.

Cleaning and Aggregating Data With Excel

Data Formats

Warning: there are some differences between Excel in Mac and Windows

You may encounter a few today

Let’s Talk About Data Formats

You’ll be looking at a lot of data throughout this course.

The formatting of data can make a real difference in your work as a DA!

Discussion:

Data Formats | Columns

Take a look at this example:

Street Address
1234 Main Street, Sacramento, CA, 95822
5678 Broadway Avenue, Denver, CO, 80122
9810 Poplar Street, Philadelphia, PA 19108

- **What do you notice?**
- **What can we do to improve it?**

Discussion:

Data Formats | Columns

How about now?

- What's changed?
- What makes this version better?

Street Address	City	State	Zip
1234 Main Street	Sacramento	CA	95822
5678 Broadway Avenue	Denver	CO	80122
9810 Poplar Street	Philadelphia	PA	19108

Partner Exercise:

Data Formats | Columns and Rows

Pair up with a classmate and take a look at the example below.

- **What do you notice? What does this data set tell you?**
- **What can we do to improve it?**

Country Name	Country Cod	Indicator Name	1960	1961	1962	1963
Aruba	ABW	Life expectancy at birth, total (years)	65.5693658	65.9880243	66.3655365	66.7139756
Andorra	AND	Life expectancy at birth, total (years)				
Afghanistan	AFG	Life expectancy at birth, total (years)	31.5800487	32.0959756	32.6118780	33.1273170
Angola	AGO	Life expectancy at birth, total (years)	32.9848292	33.3862195	33.7875853	34.1884634
Albania	ALB	Life expectancy at birth, total (years)	62.2543658	63.2734634	64.1628536	64.8870975
Arab World	ARB	Life expectancy at birth, total (years)	46.7626948	47.3886012	48.0024362	48.6075914
United Arab Emirates	ARE	Life expectancy at birth, total (years)	52.2432195	53.2865609	54.327	55.3635122
Argentina	ARG	Life expectancy at birth, total (years)	65.2155365	65.3385122	65.4326097	65.5093902
Armenia	ARM	Life expectancy at birth, total (years)	65.8634634	66.2843902	66.7098536	67.1378536

Data Formats | Columns and Rows

Here's a better way!

One row for each *variable*: country name, country code, year, and life expectancy.

It's OK if some data are repeated!

Country Name	Country Code	Year	Life Expectancy
Aruba	ABW	1960	65.56936585
Aruba	ABW	1961	65.98802439
Aruba	ABW	1962	66.36553659
Aruba	ABW	1963	66.71397561
Afghanistan	AFG	1960	31.58004878
Afghanistan	AFG	1961	32.09597561
Afghanistan	AFG	1962	32.61187805
Afghanistan	AFG	1963	33.12731707
Angola	AGO	1960	32.98482927
Angola	AGO	1961	33.38621951
Angola	AGO	1962	33.78758537
Angola	AGO	1963	34.18846341

Cleaning and Aggregating Data With Excel

Data Cleaning and Formulas

Superstore data

Your work for a chain of US supermarkets.

Their regional sales director from the central U.S. region has reached out with a request:

We want to reduce our returns. Can you dig into what might be causing them?

Let's explore the dataset using Excel!

Image source: Superstore Season 5: Release Date & Story Details (screenrant.com)

Discussion:

Getting to Know Your Data

The request:

> *We want to reduce our returns. Can you dig into what might be causing them?*

What should we look into first?

Partner Exercise:

Let's Find Out!

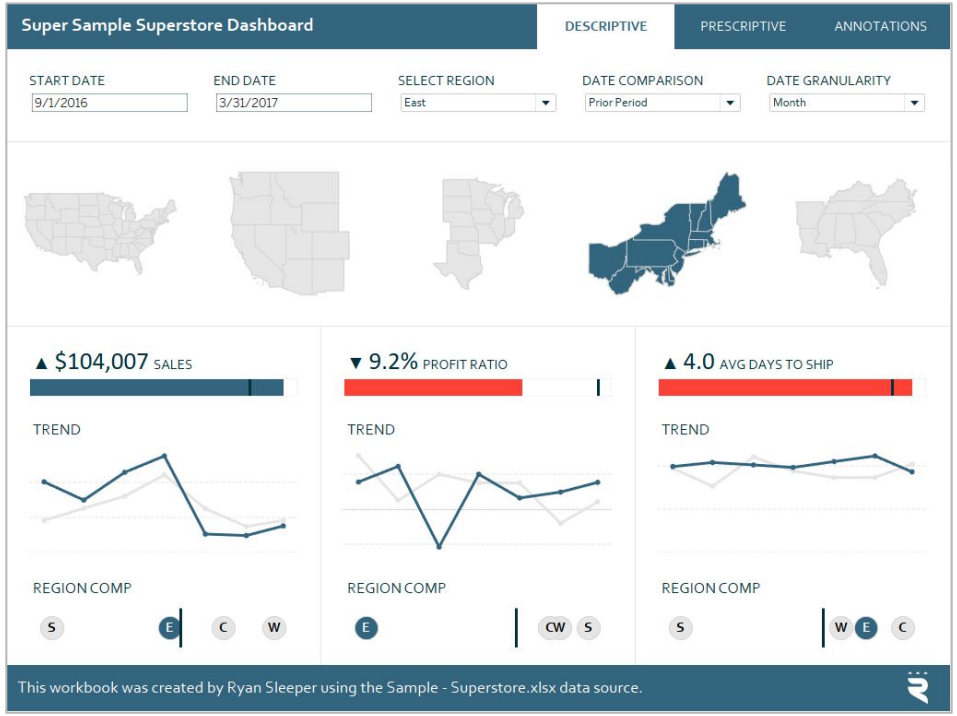
Open the workbook for today (**Day 01_Superstore Workbook.xlsx**), and with your partner:

1. Examine the **orders**, **returns** and **customers** sheets. Also look at the **superstore_dictionary.csv** which gives an explanation of the **orders** fields
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.

Be ready to share your thoughts with the class!

Importing Data for Excel Best Practices

The Superstore Data Set



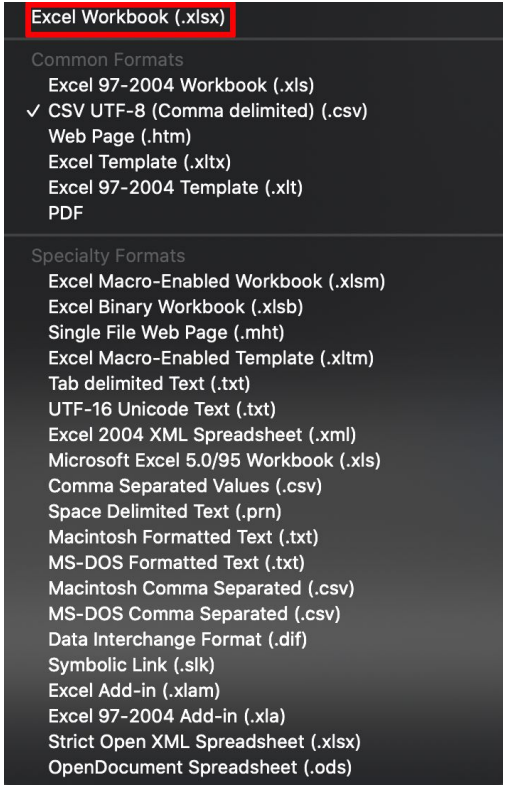
Data Set Best Practices | Resave

If you plan to analyze data in Excel, always and immediately **convert .CSV files to .XLSX**

- Go to File >> Save As

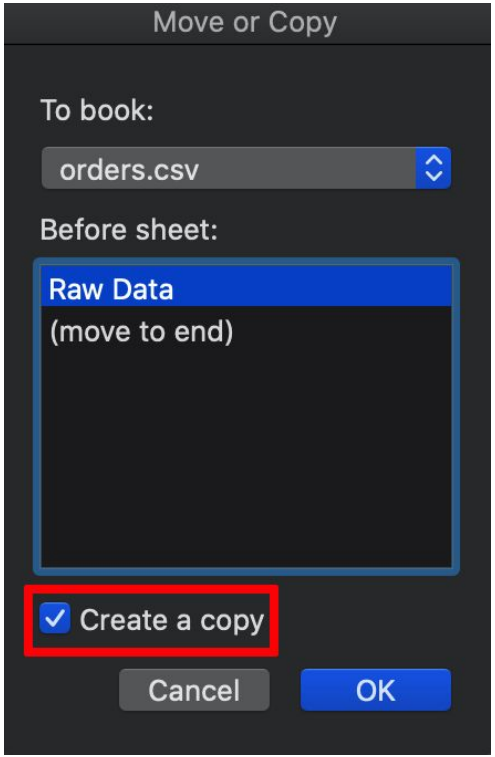
But why?

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.



Data Set Best Practices | Rename

- If you are working with one sheet of data, rename the sheet that contains the data to **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 the box next to **Create a copy**
- Hit **OK** and rename the copied sheet **Clean Data**
- When you have lots of sheet, save a copy of the entire workbook. **Why?**



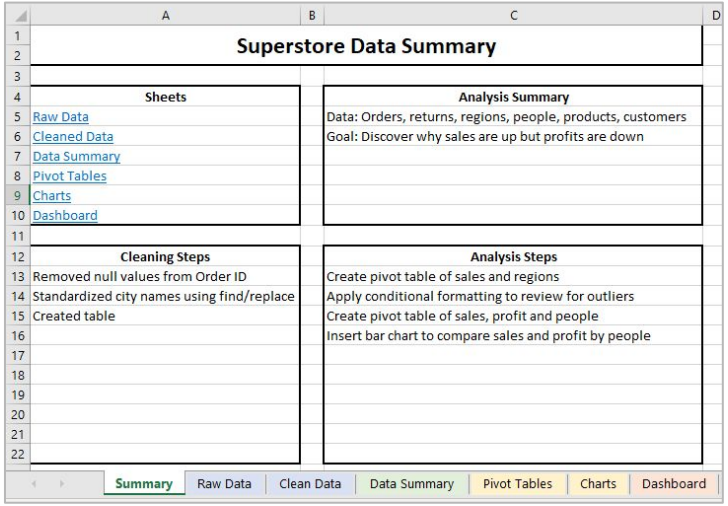
Computers Out:

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

Summary | Raw Data | Clean Data | Data Summary | Pivot Tables | Charts | Dashboard

Strategies for Cleaning and Preparing Your Data

Data Cleaning

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

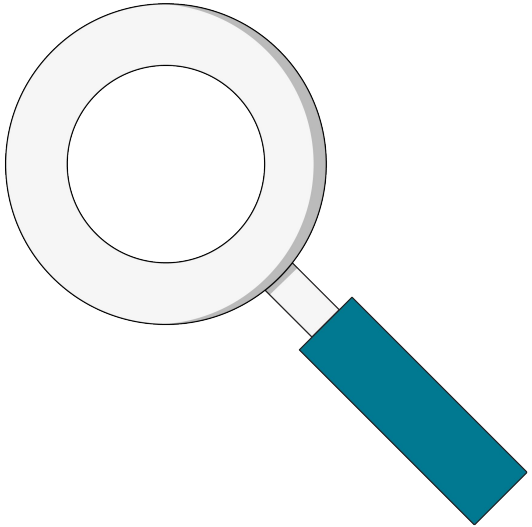
Common data cleaning actions include:

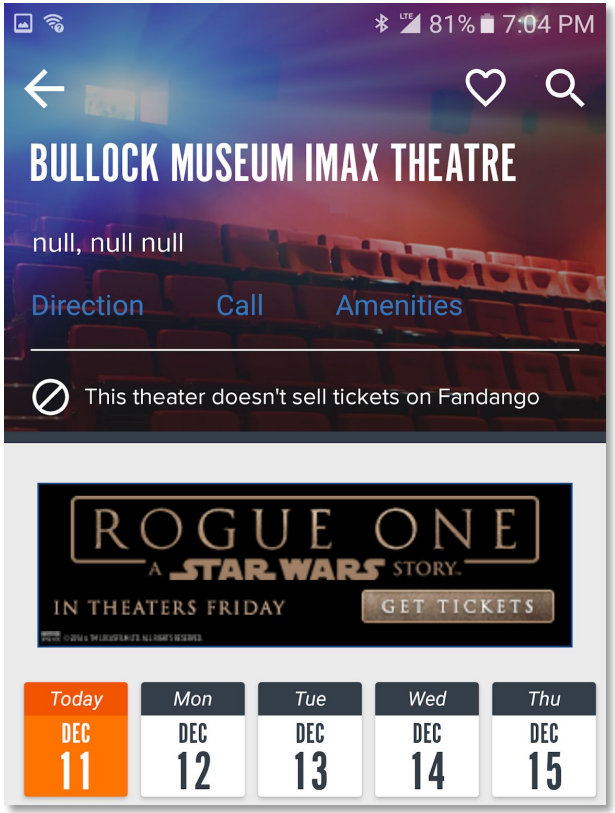
- **Reformatting dates** so that Excel recognizes them as dates.
- **Extracting day/hour/month/year from a date** to aggregate by those categories.
- **Removing duplicate values** or rows.
- **Combining data sources** into one table.
- **Concatenating or separating data.**

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!

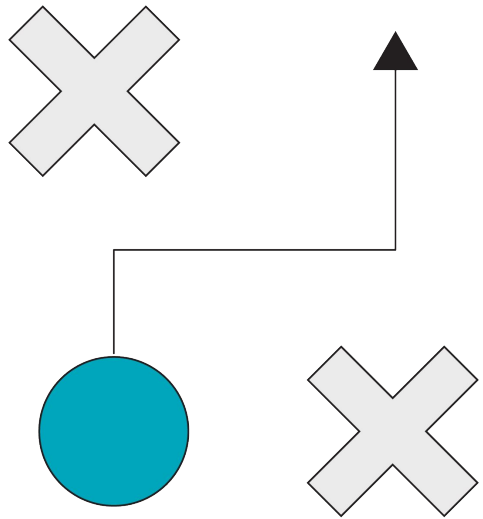




[How a 'NULL' License Plate Landed One Hacker in Ticket Hell | WIRED]

Four Primary Strategies for Handling NULLs

1. **Find missing values** (using other data sources, if available).
2. **Ignore them** (but note: some may have meaning).
3. **Impute values** (e.g. use the median or zeros).
4. **Delete them** (only with caution).



Discussion:

What to Do With Blank Cells

Take a look at the value of **Profit Margin** in **Row 3** on the **orders** sheet.

Should this be 0?

Share your answer and reasoning with the class.

	N	O	P
1	discou	profit_margin	region_
2	0.05	0.00%	1488
3	0.25		1488
4	0.1	34.00%	1488
5	0.3	29.00%	1488
6	0.35	2.54%	1488
7	0.35	1.97%	1488
8	0.15	27.00%	1488
9	0.15	0.04%	1488
10	0.15	0.42%	1488
11	0	25.00%	1488
12	0.5	0.18%	1488

Calculating Profit Margin

In the cell with the missing profit margin, enter:

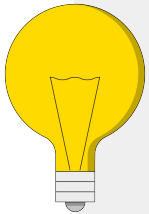
```
=K3/J3
```

You can do this by:

- Typing it in the cell exactly as written
- (If the formula for the other rows was in the spreadsheet, rather than just the values, you could copy or fill from the row above)

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!

Guided Walk-Through:

Text to Columns | Step by Step

Step 1: Right click on the column letter 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`.

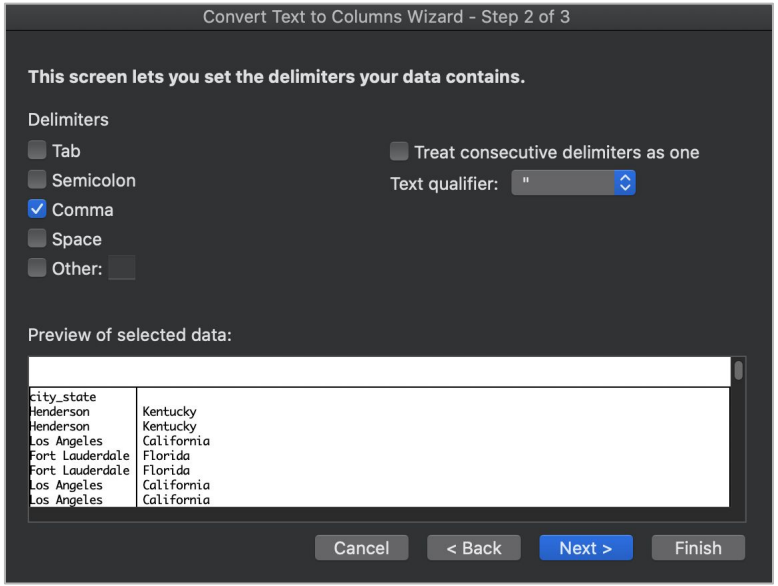
Step 2: Click on the column letter above `city_state` to select the entire `city_state` column.

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

Guided Walk-Through:
Text to Columns | Step by Step

Step 4: Choose **Delimited**. Then, click **Next** and check **Comma** only. Click **Finish**.

Step 5: Rename the `city_state` column to just `city`, and the second column to `state`.



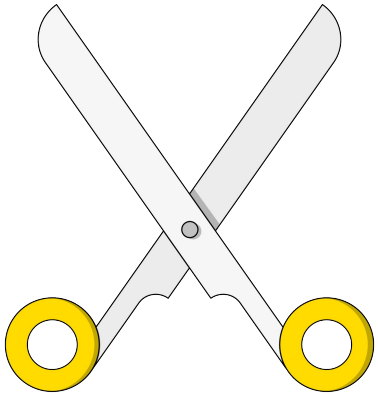
Text to Columns | Trimming the Spaces

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)
```

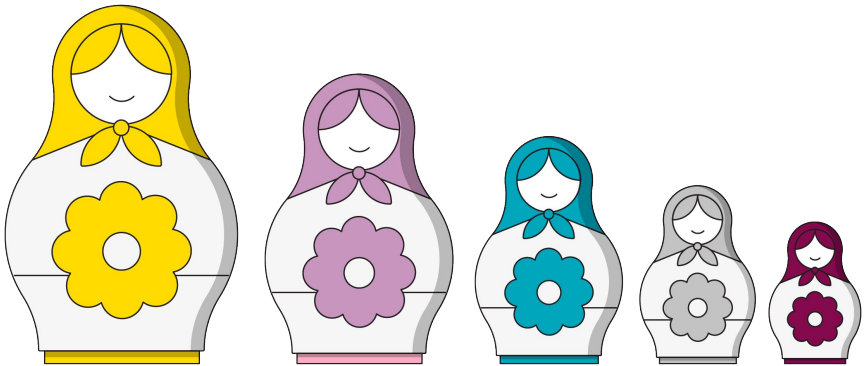


Discussion:

Checking for Duplicates

Finally, let's check for duplicates!

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



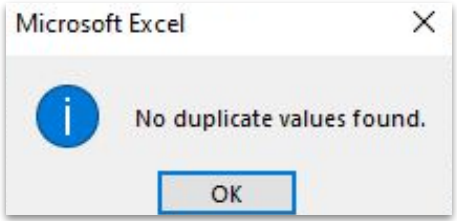
Guided Walk-Through:

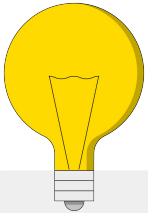
Checking for Duplicates | Step by Step

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

Step 2: Unselect All.

Step 3: Then, check ONLY `order_id` (Column A) and `product_id` (Column E) before clicking **OK**.





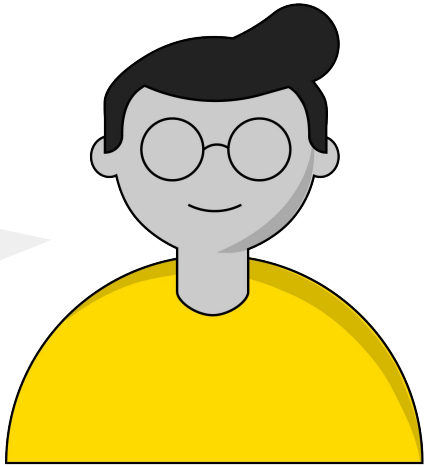
If duplicates are removed by this method, we aren't told which rows they are. We also can't check for duplicates without removing them. Later we will cover **PivotTables**, which can help us do this.

Asking the Right Questions (of Your Data)

Asking the Right Questions

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

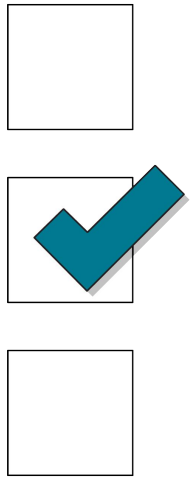
Hmm, this question is really broad. Let me explore the data set first.



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

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

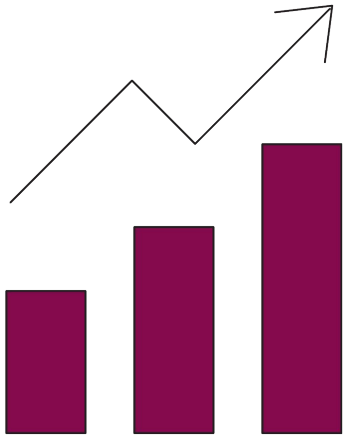
- **How was the data for each column collected?**
- **What are the units of each column?**
- **According to the data dictionary, what does each column represent?**

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 | Parts of a Hypothesis

A hypothesis is an idea that is yet to be proven. In Data Analysis, we use hypotheses to help us ask questions where we can provide insights and solutions to a problem or to enable us to make data driven decisions. A hypothesis typically contains 3 parts:

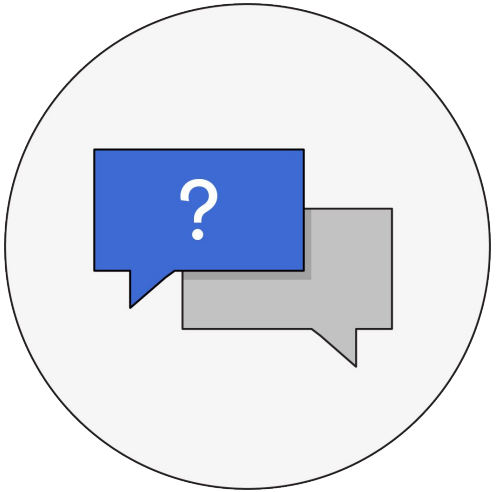
- **Problem:** What question are we trying to ask now that we have explored the data? Can we validate it with qualitative or quantitative data?
- **Solution:** Can we derive a proposed solution or state a rationale? Will this solve our problem questions? As we dig into the data deeper, this helps to solidify the hypothesis/problem statement we initially asked.
- **Result:** What are the suggested results (metrics/measures)? What will be my criteria to determine failure/success?

Source: https://help.optimizely.com/Ideate_and_Hypothesize/Design_an_effective_hypothesis

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 (which we need to prioritise)

Questions	Hypotheses
Does shipping have something to do with the profit margin?	If we look at profit and ship mode together, we might discover that certain ship modes are consistently associated with lower profits.
Which region has the highest number of returns?	If we look at the the overall returns, we might discover that specific regions have more returns than others.
Which salesperson has the highest number of returns?	If we look at the regions and salespersons, we might discover that not only do specific regions have higher profits, but they are associated with specific salespersons.
Which salesperson has the highest profit in 2019?	If we look at profit together with sales person together for only 2019 do specific sales persons stand out as having the highest profits.

Discussion:

Hypotheses | Good or Bad?

Take a moment to review the hypotheses below. Keeping in mind that a good hypothesis is actionable, **which one of the following is good, and which one can use some improvement?**

A. If we look at profit and ship mode together, we might discover that certain ship modes are consistently associated with lower profits. Therefore, we might recommend that Superstore stop offering those ship modes to customers in order to boost profits.

B. 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.

Discussion:

Hypotheses | Good or Bad?

Which one of the following is good, and which one can use some improvement?

A. Number of orders with returns and salespeople. We can get the list of order_ids with a return. For example, order 123 was returned on 2/23/2019. But that doesn't really lead to many useful insights for the store. An aggregate of the count of the returned orders and the salespeople who are in that region across all returns might be more useful.

B. If we look at orders where there was a return for salespeople, we might discover that some salespeople have higher numbers of returns compared to other salespeople. Therefore, we might recommend that Superstore dig into why some salespeople have these higher return rates and the reason why the order was returned and look at setting a limit on the amount of returns accepted.

Discussion:

Hypotheses | Good or Bad?

Which one of the following is good, and which one can use some improvement?

A. If we look at profits for orders in 2019 for salespeople, we might discover that some salespeople produce higher profits and some produce lower profits. Therefore, we might recommend that Superstore investigate what the top profiting salespeople are doing to be profitable and look at ways to help the lower performing salespeople improve their performance.

B. Look at orders with a profit and salespeople. We can get the list of order_ids and their profit and a list of salespeople for those orders. For example, order 123 was from Annelise Williams and has a profit of $0.33. But that doesn't really lead to an true insights to how profitable Annelise Williams is as a salesperson. An aggregate of the average profit by the salespeople sorted descending to see the top and bottom performers be more useful.

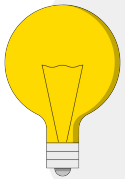
Discussion:

Formulating Superstore Hypotheses

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 to share it with your class.



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

Discussion:

Creating Hypotheses to Analysis

We can ask some interesting hypothesis-driven questions with our dataset:

1. Do different types of customer segments make more returns?
2. Do different categories get returned more often than others?
3. Does the cost of an item impact how likely it is to be returned?

How would we answer these?

Partner Exercise:

Formulating Superstore Hypotheses

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

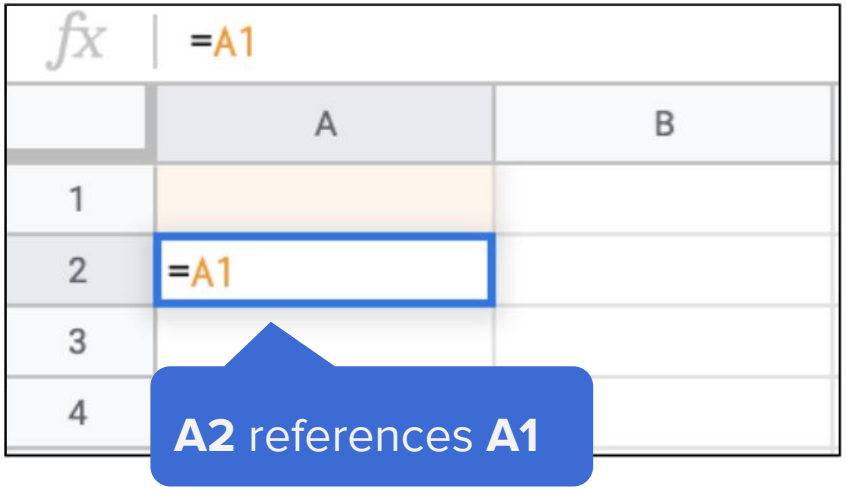
1. Identify the questions you can ask to help gain interesting insights from the data.
2. Then, formulate your questions into hypotheses. Here's an example:
   "If we compare the shipping cost and the order priority, we might find that high shipping costs for low-priority orders frequently lead to returns."
3. List it out in your worksheet.

Be prepared to share your work with your class.

Excel Functions

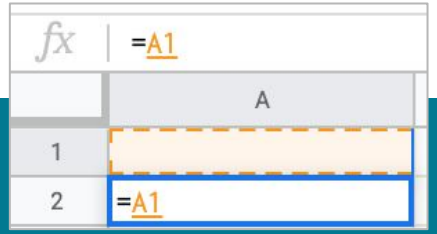
Data Referencing

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

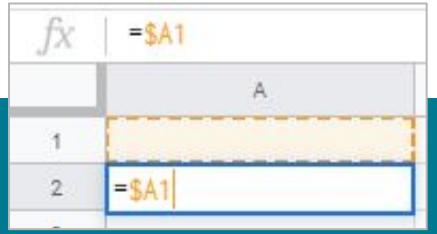


Cell Referencing

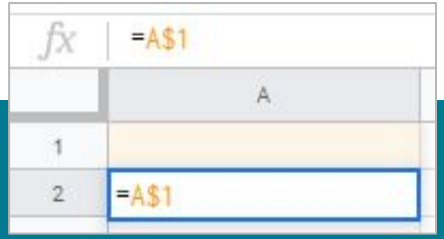
An **absolute** reference is a **fixed (locked) location** in a worksheet.



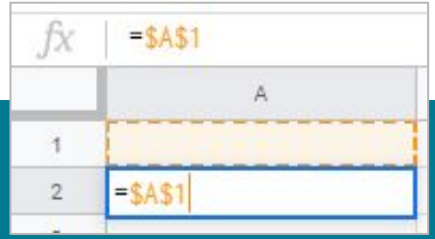
Relative



Mixed
Only Column is Fixed



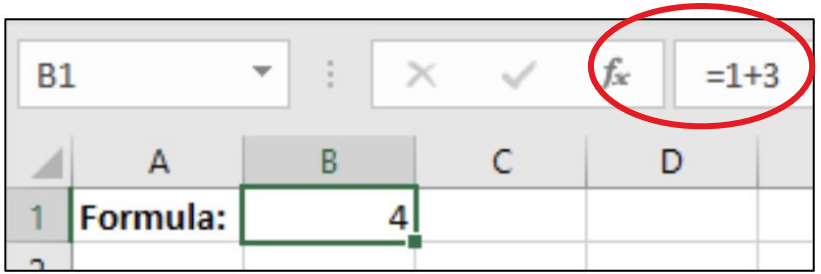
Mixed
Only Row is Fixed



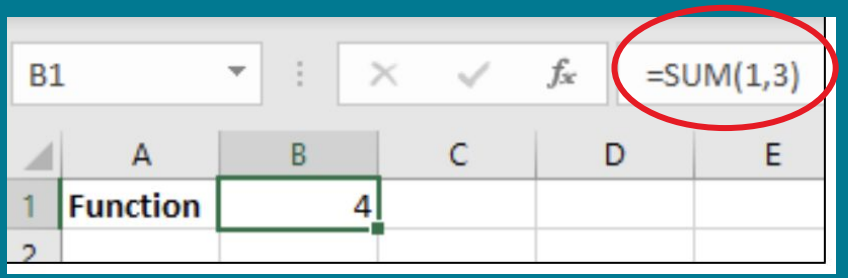
Absolute

What Is a Formula in Excel? And what is a Function?

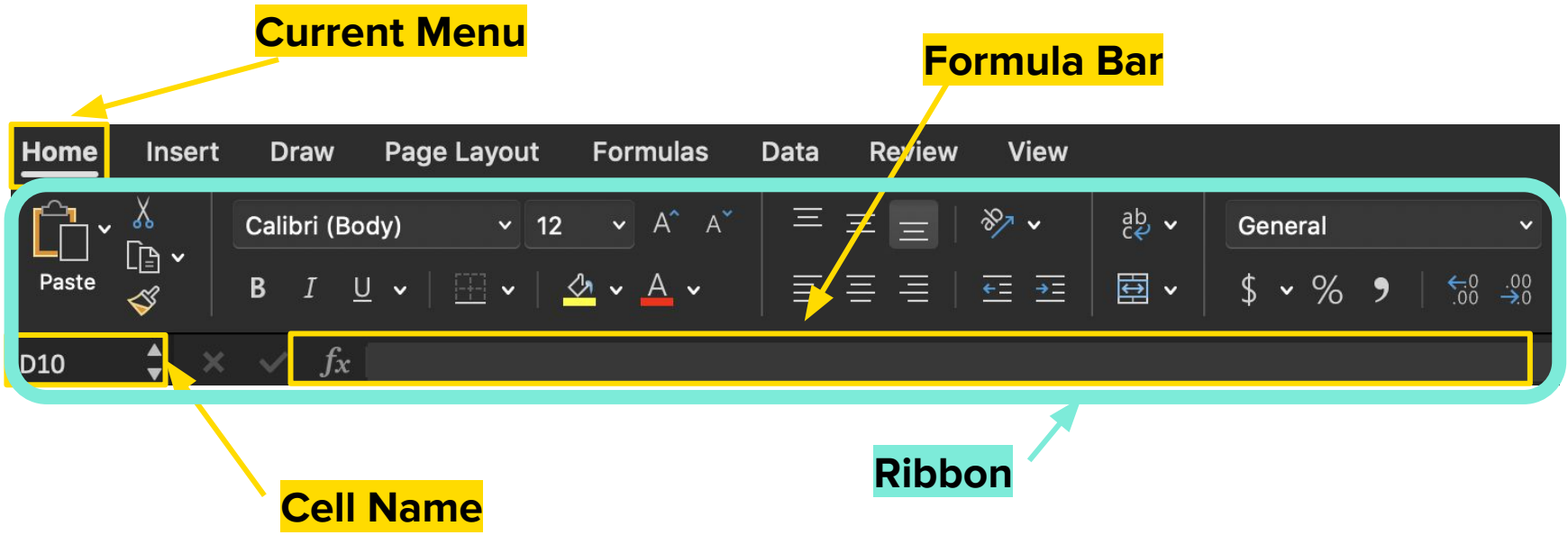
A **formula** is an expression which calculates the value of a cell.



Functions are predefined formulas that are already available in Excel.

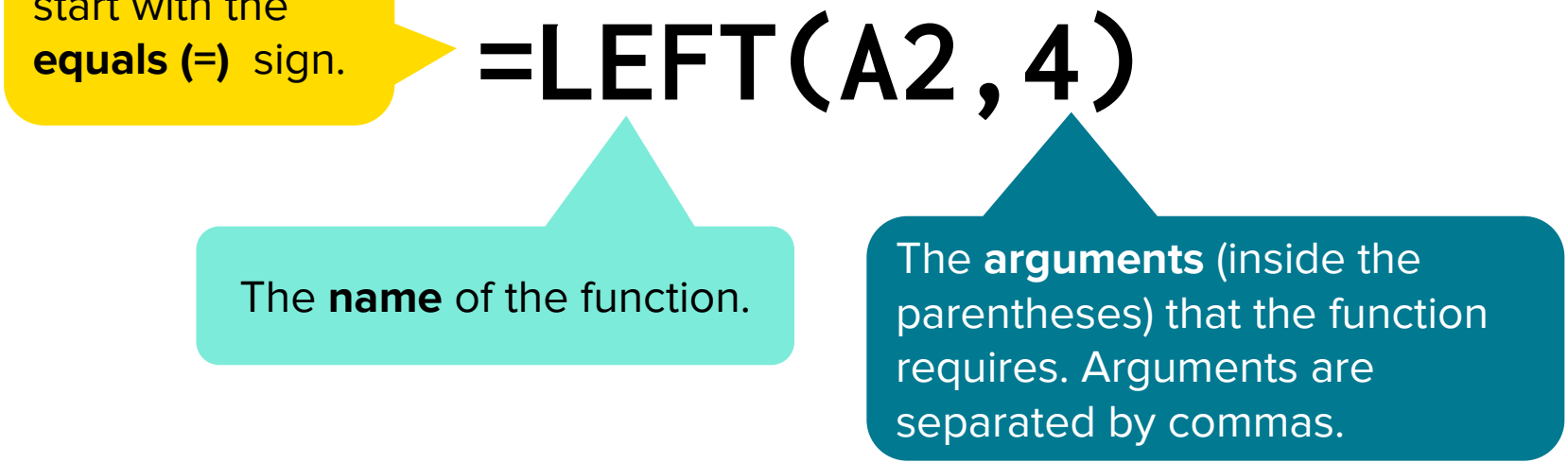


Navigating Formulas and Functions in Excel



The Anatomy of an Excel Function

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



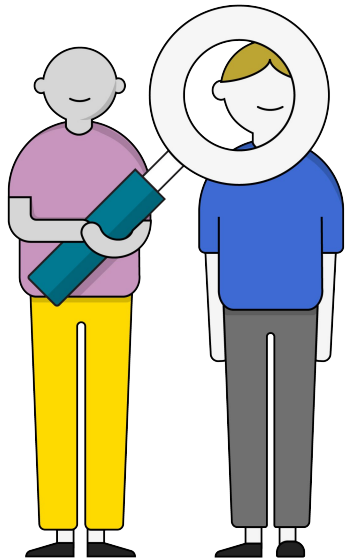
Finding the Right Function

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.

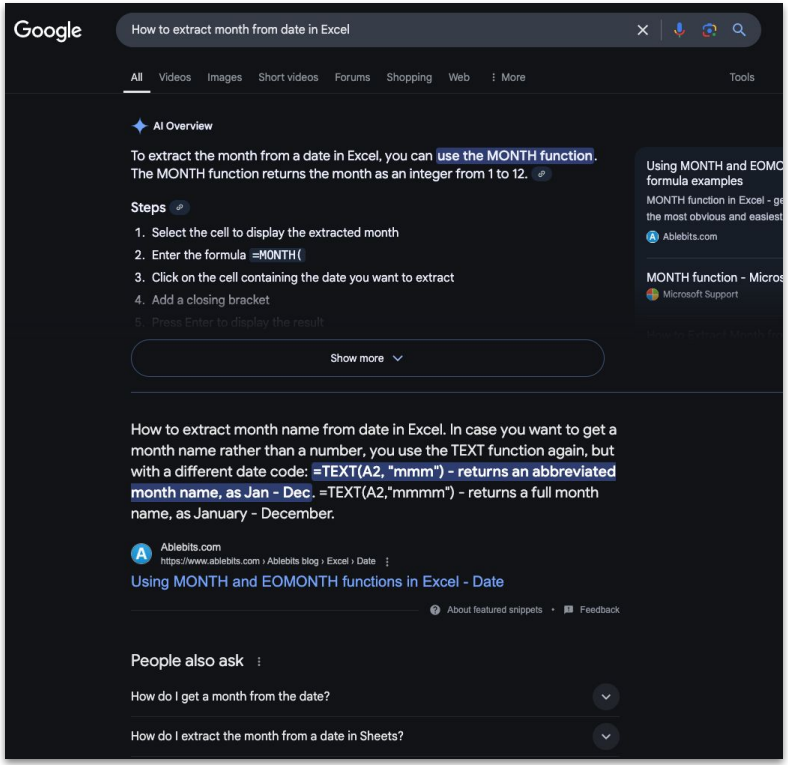


Finding the Right Function | Google It

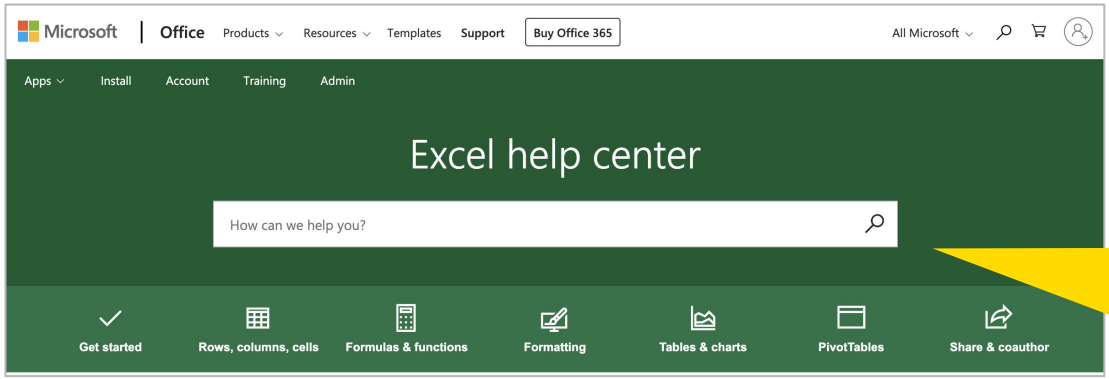
We want to examine whether returns happen more in certain months. But our dates are by day.

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:

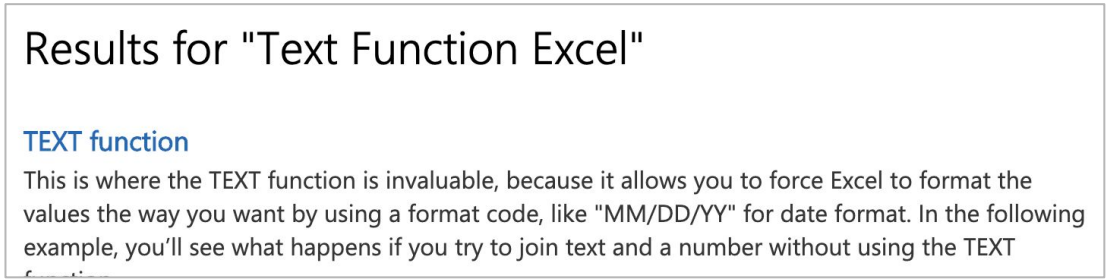
"How to extract month from date in Excel."



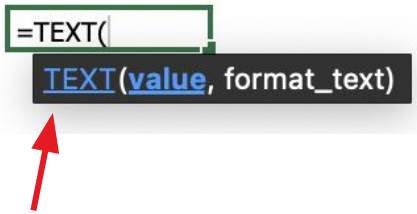
Finding the Right Function | MS Documentation



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



Finding the Right Function: MS Documentation



You can also type the function's name into a cell in Excel together with an opening parenthesis, and **click on the link** to go to the documentation for the function.

Discussion:

Finding the Right Function | Arguments

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.”**

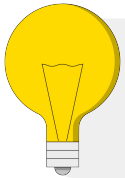
5	To display	As	Format	Formula
6	Months	1–12	"m"	=TEXT(B3,"m")
7	Months	01–12	"mm"	=TEXT(B3,"mm")
8	Months	Jan–Dec	"mmm"	=TEXT(B3,"mmm")
9	Months	January–December	"mmmm"	=TEXT(B3,"mmmm"
10	Months	J–D	"mmmmm"	=TEXT(B3,"mmmmm
11	Days	1–31	"d"	=TEXT(B3,"d")
12	Days	01–31	"dd"	=TEXT(B3,"dd")
13	Days	Sun–Sat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

Computers Out:

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 this function to the Superstore data set:
   - `=TEXT(B2, "mmmm")`



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

Guided Walk-Through:

Data Cleaning With COUNTIF

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)
```

Guided Walk-Through:

Data Cleaning With COUNTIF | Let's Try It!

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(N2, ">=.3")`.

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

How many states are there in the dataset?

Suppose we want to know **which US states the Central United States sub-region has shipped to in 2019**.

Which column will give us this information?

Source of image

Guided Walk-Through:

Identifying the unique values in a category

We can use the `UNIQUE` function to give us a list of each unique state listed in our `state_trimmed` column.

Create a new sheet and call it `unique_states`, and in cell A1 type this:

```
=UNIQUE(orders!S:S)
```

What do you notice that's different from the formulas we've used so far?

Guided Walk-Through:
So, What's Really Going on With Returns? Part 1

To dive deeper into why Superstore is seeing a high volume of returns, we can take a closer look at **orders**, **profit**, and **sales**.

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`, what should we write out?

Guided Walk-Through:

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

2. Looking at profit, does profit margin impact whether or not something gets returned? We already have a **`profit_margin`** column but can recalculate it as a check on accuracy. To find out, insert a new column to the right of **`profit_margin`** and call it **`profit_margin_calc`**. Recalculate the profit margin (**`profit`** divided by **`sales`**) per row.

```
=K2/J2
```

Next, we will use IFERROR to *wrap* the formula. We do this to help us deal with any issues in the dataset that might prevent the formula from calculating (e.g. NULLs or 0s on the bottom of a fraction).

```
=IFERROR(formula,"")
```


Guided Walk-Through:

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

3. 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." Let's create a column called `sales_category`

```
=IF(J2>500, "High", "Low")
```

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

We'll find out soon... but before that let's add some more data to our sheets

Solo Exercise:

Optional Homework

Finish the **`homework (optional)`** 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_customer`** column that identifies customers in the given list.

You may need to use formulas we didn't cover in the lesson!

Cleaning and Aggregating Data With Excel

Referencing and Lookups

Discussion:

Looking Up Information

Referencing is all about looking something up elsewhere and comparing it with what's in front of you.

While we're still investigating the reason behind the increase in returns, let's open up all of your Superstore sheets and look into **those attributes you think will have the biggest impact on returns.**

Discussion:
Referencing Information

Based on the attributes we identified, we now know that we need to:

Reference customers with frequent returns in a region and match them up with attributes of their sale.

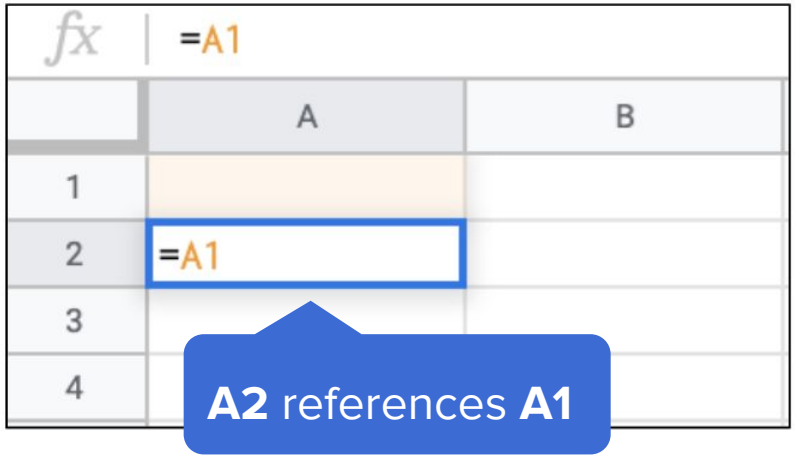
What's an efficient way to go about this?

Getting Started With VLOOKUP

Referencing and VLOOKUP

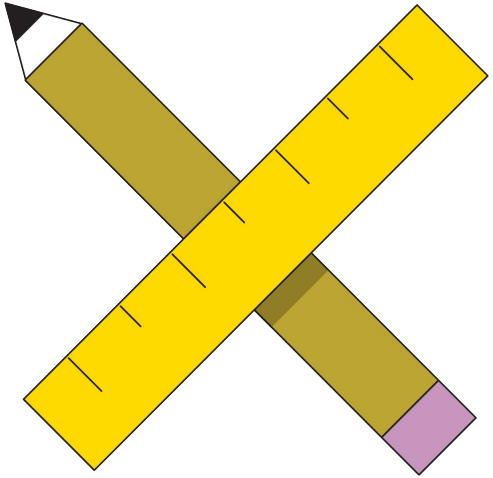
Referencing, in its basic form, means pulling the value of one cell into another cell.

With VLOOKUP, we will sometimes need to **reference across sheets, and even files,** and **lock references** to make the formula function properly.



Referencing With Advanced Excel Tools

VLOOKUP, **HLOOKUP**, **XLOOKUP**, and **INDEX/MATCH** are often considered advanced tools that **increase efficiency** while **reducing data integrity issues**.



What Is VLOOKUP?

V stands for “vertical.”

VLOOKUP is:

- A function that works with **data laid out in columns**.
- A function that finds or “looks up” the value in one column of data and **returns the corresponding value** from another column (and usually another table).

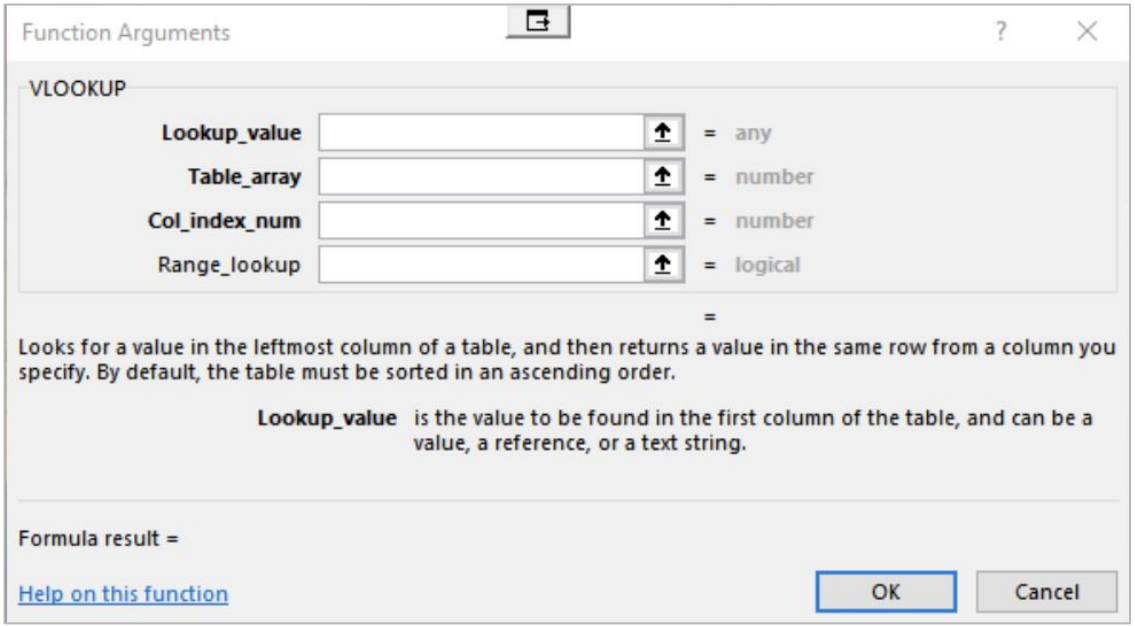
Building a VLOOKUP Statement

When building a **VLOOKUP** statement:

- In the cell, type "**=VL**."
- Click on the function name that's presented for syntax.

Building a VLOOKUP Statement

Alternatively, click ***fx*** in the menu ribbon and select **Function** > VLOOKUP.



VLOOKUP Syntax

Lookup_value is the value that will be used to match data. It's usually an identifier and it must exist in both worksheets.

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

Table_array is the table from which you want to retrieve data.

Col_index_num is the number of the column from the left side of the table_array from which you want to retrieve data.

VLOOKUP Syntax | Range_lookup

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

range_lookup defines whether or not the lookup_value is an approximate match or an exact match of the value you are comparing it to in the left-most column of the table_array.

TRUE: Approximate match is needed.

FALSE: An exact match is required.

In the work of a data analyst, it is rare to have all the data you need right in your data set.

Guided Walk-Through:

Pulling Data From Another Worksheet - Customer to Orders

Let's combine data from another worksheet into the Superstore data set and explore the concept of a "lookup table" as well as how to use one to categorize our quantitative data.

In order to combine the **orders** with the **customers**, we'll create two new columns: `customer_name` and `segment` in the `orders` sheet — we can use **VLOOKUP** to "look up" one value in another table and return another column in that row.

Guided Walk-Through:

Pulling Data From Another Worksheet - Customer to Orders

Let's try adding **customer names** and **segment** by selecting a range for our `VLOOKUP`.

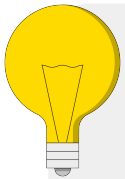
1. In cell **AC2** of `orders`, enter: `=VLOOKUP(Y2,customers!A:C,2,FALSE)`. The `customer name` should populate cell **AC2**.
2. Double click the bottom right corner of the cell — or click/drag — to replicate this function down the entire column.
3. In cell **AD2** of `orders`, enter: `=VLOOKUP(Y2,customers!A:C,3,FALSE)`. The `segment` should populate cell **AD2**.
4. Double click the bottom right corner of the cell — or click/drag — to replicate this function down the entire column.

Guided Walk-Through:

Pulling Data From Another Worksheet - Creating a Key

Sometimes we need to concatenate fields to create a unique identifier for matching with other tables.

In order to combine orders and returns, we need to create a “key” that includes more than one column value. This is referred to as a “concatenated key”. In the orders tab, **insert a new column before Column A**. Name this column **key**.



Although it is usually better practice to add new columns to the right of the dataset, it will be easier to use **lookups** if we create this column to the left of the dataset. This column allows us to relate the orders data to data in other tables.

Guided Walk-Through:
Pulling Data From Another Worksheet - Creating a Key

Create the key we will use in the **VLOOKUP**.

1. In cell A2 of **orders**, enter: **=B2&F2**.

   The 2 values are brought together as a single value.

2. Double click the bottom right corner of the cell — or click/drag — to replicate this function down the entire column.

Guided Walk-Through:

Pulling Data From Another Worksheet - Creating a Key

Now create the same key in the **`returns`** sheet.

1. Insert a new column before **Column A** in the **`returns`** sheet, named **`key`**.
2. In cell A2 of **`returns`**, enter: **`=B2&F2`**.

   The 2 values are brought together as a single value.

2. Double click the bottom right corner of the cell — or click/drag — to replicate this function down the entire column.

We now have keys to perform our Lookups.

Guided Walk-Through:

Pulling Data From Another Worksheet - Returns to Orders

Let's try adding **return reason** to `orders` by selecting a range for our VLOOKUP.

1. In order to combine these two lists, we'll create a column in `orders` with `return_reason` — we can use **VLOOKUP** to "look up" one value in another table and return another column in that row.
2. In cell AD2 (or after the last column) of "**Orders**", enter: `=VLOOKUP(A2,returns!A:E,5,FALSE)`.

   The Return Reasons should populate cell **AD2**.
3. To remove the **#N/A,** use IFERROR to show blanks if no return reason. Enter: `=IFERROR(VLOOKUP(A2,returns!A:E,5,FALSE),"")`
4. Double click the bottom right corner of the cell — or click/drag — to replicate this function down the entire column.

Guided Walk-Through:

Pulling Data From Another Worksheet - Orders to Returns

Let's try adding customer names to "Returns" using VLOOKUP.

1. In order to combine these two lists, we'll create a column in `returns` with `customer_name` — we can use **VLOOKUP** to "look up" one value in another table and return another column in that row.
2. In cell G2 of `returns`, enter: `=VLOOKUP(A2,orders!A:AH,28,FALSE)`.

   The customer name should populate cell **G2**.
3. Double click the bottom right corner of the cell — or click/drag — to replicate this function down the entire column.

Guided Walk-Through:

Using a VLOOKUP - Orders to Returns

Now, we'll examine if certain **states** have more returns than others.

1. In our `returns` worksheet, go to Column **H**.
2. Name this `state` by typing this in Cell **H1**.
3. In Cell **H2**, enter:
   `=VLOOKUP(A2,orders!A:AH,21,FALSE)`
4. Expand this formula to all rows by double clicking the bottom-right corner of the cell.

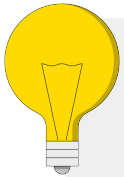
Now we have the states in our `returns` data set!

Solo Exercise:

Moving Data With VLOOKUP - Optional Homework

Now let's say we want to bring in more information from the **orders** data set. We don't need everything, just a few columns.

On your own, use **VLOOKUP** to bring the **category, sales,** and **profit** columns from the **orders** worksheet to the **returns** worksheet.



Remember to start from your "**key**" column!

Solo Exercise:

Moving Data With VLOOKUP - Optional Homework - Solutions

Category:

```
=VLOOKUP(A2,orders!A:AJ,7,FALSE)
```

Sales:

```
=VLOOKUP(A2,orders!A:AJ,11,FALSE)
```

Profit:

```
=VLOOKUP(A2,orders!A:AJ,12,FALSE)
```


Creating Categorical Variables

What Are Categorical Variables?

A categorical value, aka, a nominal variable, typically has **two or more non-ordinal categories** (values).

Hair color is a categorical variable that has a number of categories (e.g. blonde, brown, red...), but there is no inherent order to these categories.

Creating Categorical Values

Often, it's helpful to create categorical values from **numeric values**.

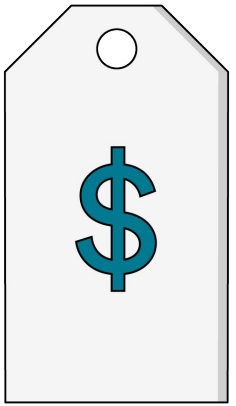
For example, a test that is scored 0–100 could be **classified** as A, B, C, D, or F, depending on the score.

Discussion:

Categorical Values in the Superstore Data Set

Knowing that it's helpful to create categorical values from **numeric values**, let's take a look at the **orders** data.

To help us solve the returns problem, which columns should we use to create categorical values that will help us analyze our data?

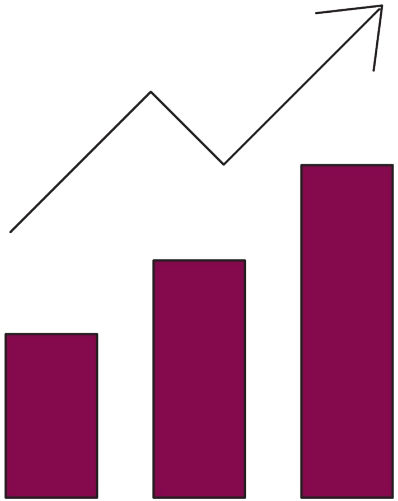


Computers Out:

Categorical Values in the Profit Margins

Let's practice assigning categorical values to the profit margins.

1. In the `returns` worksheet, in Cell **L1**, type **profit_margin**.
2. In Cell **L2**, enter the formula `=VLOOKUP(A2,orders!A:AD,15,FALSE)` and fill downwards.
3. Classify our profit margins as either *low*, *medium*, or *high*.



Guided Walk-Through:
Margin Lookup

Refer to the steps below:

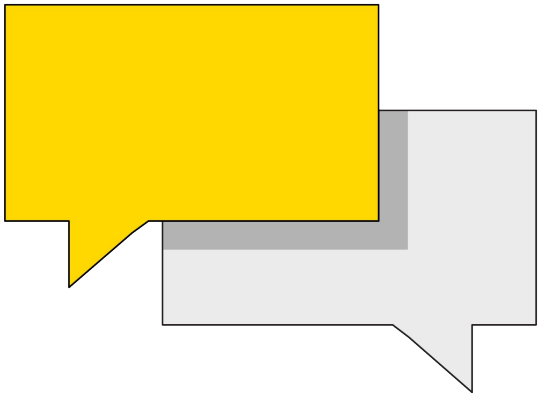
1. Create a new worksheet called **`margin_lookup`**
2. In Cells **A1**, **A2**, and **A3**, enter values -5, 0, and 0.3.
3. In Cells **B1**, **B2**, and **B3**, enter values low, medium, and high.
4. On the **`returns`** worksheet, type **`margin_category`** in Cell **M1**.
5. In **M2**, complete the lookup:
   `=VLOOKUP(L2,margin_lookup!A:B,2,TRUE)`.
6. Expand to all rows.

Discussion:

VLOOKUP and Absolute Cell References

Remember absolute cell references?

- What happens if we don't refer to entire columns?
- What would happen if we didn't use absolute cell references?



Share your answers with the class!

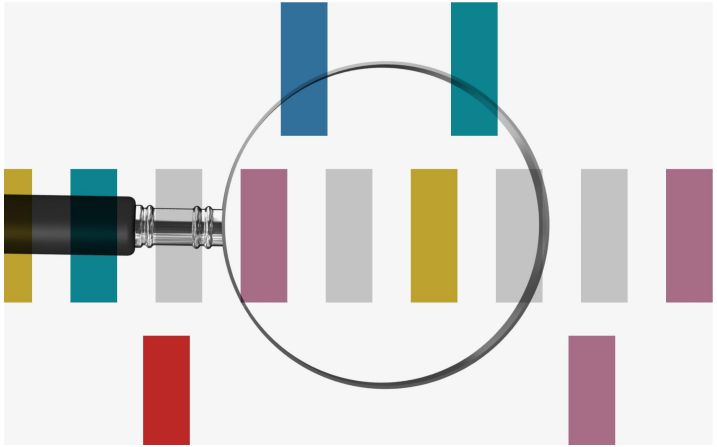
Other LOOKUPs

What Is HLOOKUP?

H stands for “horizontal.”

HLOOKUP is closely related to VLOOKUP, but instead of working with data sorted into columns, **HLOOKUP** work with **data sorted in rows**.

Because this is usually a difficult way to arrange data, it is *not* often used.



HLOOKUP Syntax

The **value** to look for in the first column of a table.

The **table** from which to retrieve a value.

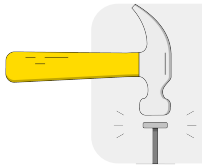
```
=HLOOKUP(lookup_value, table_array, row_index_num, [range_lookup])
```

The **row** in a table from which to retrieve a value.

Range_lookup: optional.
TRUE = Approximate match (default).
FALSE = Exact match.

Limitations of LOOKUPs

- LOOKUPs are **unidirectional** and must work with indices fixed to the left side (VLOOKUP) or top (HLOOKUP) of the work area.
- Because the VLOOKUP references a col_index, it's **unable to dynamically update** whenever you insert a column or columns in the table_array.

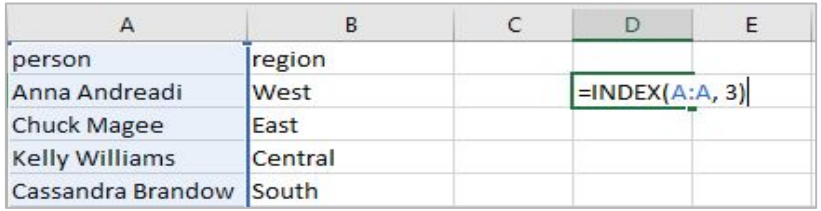


Let's see what happens when we "break" the VLOOKUP by inserting a column.

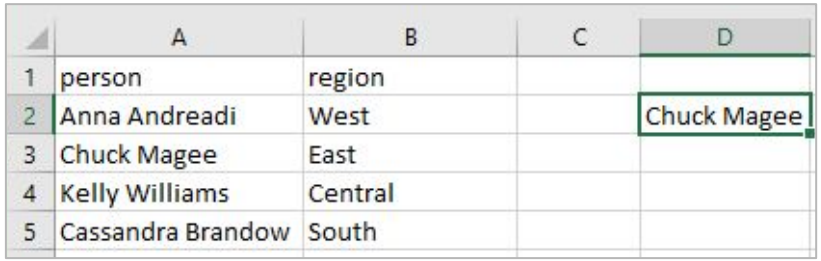
INDEX and MATCH

INDEX

INDEX is a function that returns **the value at the intersection of a row and column** in a given range.



	A	B	C	D	E
1	person	region			
2	Anna Andreadi	West		=INDEX(A:A, 3)	
3	Chuck Magee	East			
4	Kelly Williams	Central			
5	Cassandra Brandow	South			



	A	B	C	D
1	person	region		
2	Anna Andreadi	West		Chuck Magee
3	Chuck Magee	East		
4	Kelly Williams	Central		
5	Cassandra Brandow	South		

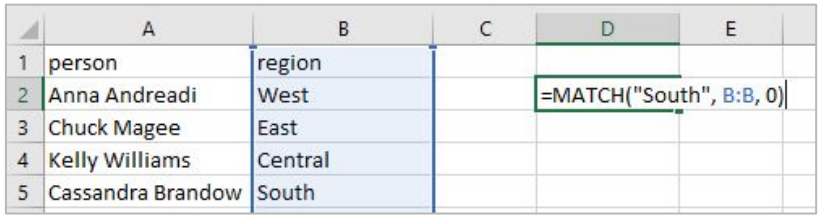
Syntax:

```
=INDEX(Array, Row_num, Column_num)
```

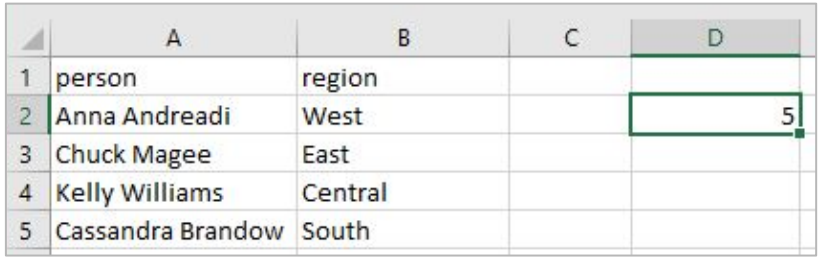
Example: I want to know the name of the person in the third row of the “person” column.

MATCH

Returns **the position of an item in an array** that matches a value.



	A	B	C	D	E
1	person	region			
2	Anna Andreadi	West		=MATCH("South", B:B, 0)	
3	Chuck Magee	East			
4	Kelly Williams	Central			
5	Cassandra Brandow	South			



	A	B	C	D
1	person	region		
2	Anna Andreadi	West		5
3	Chuck Magee	East		
4	Kelly Williams	Central		
5	Cassandra Brandow	South		

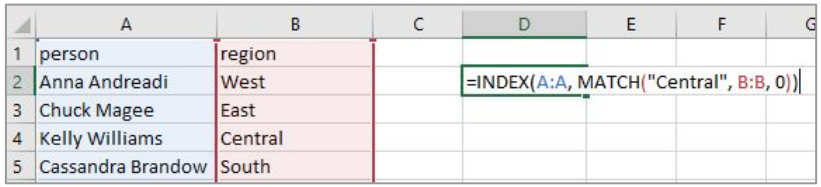
Syntax:

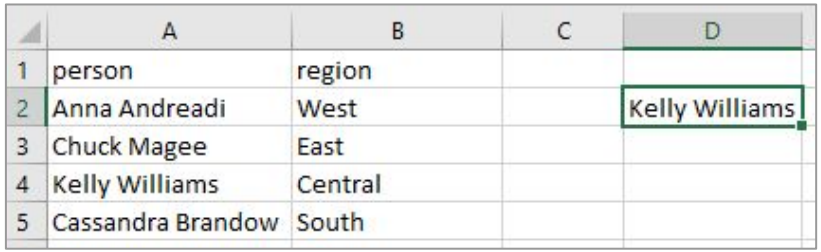
```
=MATCH(Lookup_value, Lookup_array, Match_type)
```

Example: I want to find the first occurrence of “South” in the column containing region.

INDEX/MATCH LOOKUP

Specifies lookup column + **returns the value column independently**.





Syntax:

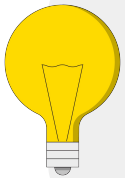
```
=INDEX(Return_value_range,
MATCH(Lookup_value,
Lookup_value_range, Match_type))
```

Example: I want to return the name of the person whose region is "Central".

Guided Walk-Through:
INDEX/MATCH Demo, Part 1

Open the “**index_demo”** spreadsheet, and let’s prepare the worksheet with these steps.

1. In Cell **E2**, type: `=INDEX(A:A, 4)`
2. In Cell **E3**, type: `=MATCH(“Central”, B:B, 0)`



For now, we’ll always use a “0” as the third argument. The “0”parameter requires an **exact** match. In the next section, we’ll see how to use an **inexact** match.

Guided Walk-Through:

INDEX/MATCH Demo, Part 2

We will now combine **INDEX** and **MATCH** into a nested formula.

3. In Cell **I2**, type: `=INDEX(A:A,MATCH(H2,B:B,0))`
4. Copy it down to **I3**.

The inner **MATCH** looks up the “region” of interest in the B:B column, returning the matching row number. The row number from the **MATCH** above is then used in the **INDEX** to look up and return a value in Column A:A.

Guided Walk-Through:

Revising VLOOKUPS With INDEX/MATCH

Let's recreate our VLOOKUPs from earlier using INDEX/MATCH. In a new column, use **INDEX/MATCH** to look up the customer names. Open up the `returns` sheet.

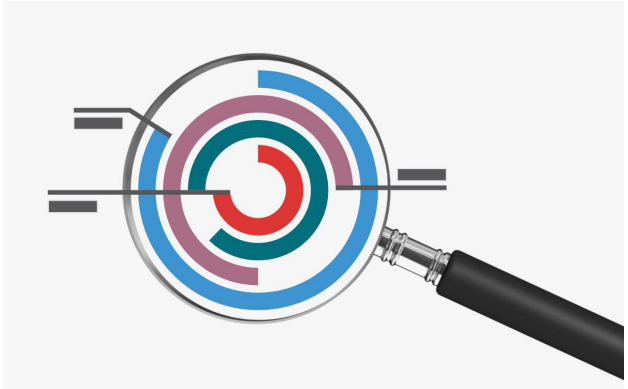
1. In Column N, create a "Customer Name-IM" (IM to indicate Index/Match) column, and type:
   `=INDEX(orders!AH:AH,MATCH(A2,orders!A:A,0))`
2. Copy this down to all rows.

XLOOKUP

What Is XLOOKUP?

XLOOKUP is closely related to the other LOOKUPs, but it...

- Only requires the **lookup_value**, **lookup_array**, and **return_array** arguments.
- *Does not* require the return_array to be to the right of the lookup_array (it can be on either side).
- Does not need a match_mode argument and defaults to exact.



XLOOKUP Syntax

The **value** to look for in the first column of a table.

Array that contains the answer you want to return.

```
=XLOOKUP(lookup_value, lookup_array, return_array)
```

Array in which the lookup_value can be found.

Guided Walk-Through:
“If Not Found” XLOOKUP

“If not found” XLOOKUP allows us to **set a return value** if a match isn’t found, negating the use of “IFERROR” or “IFNA.”

“If not found” XLOOKUP accepts four arguments:

```
=XLOOKUP(lookup_value, lookup_array, return_array, value_if_not_found)
```

Guided Walk-Through:
XLOOKUP Search Mode

XLOOKUP Search Mode allows us to specify the direction of the lookup.

```
=XLOOKUP(lookup_value, lookup_array, return_array, value_if_not_found,
match_mode, search_mode)
```

- **Match_mode** specifies exact or inexact match.
- **Search_mode** specifies the direction of the lookup: 1 is first item to last and -1 is last item to first.

Guided Walk-Through:
Nested XLOOKUP

A **nested XLOOKUP** can perform vertical AND horizontal lookups, just like an INDEX/MATCH.

```
=XLOOKUP(lookup_value, lookup_range, XLOOKUP(lookup_value, lookup_range, return_range))
```.

Don’t forget that a nested XLOOKUP will need **two** end parentheses!

Partner Exercise:

XLOOKUP or INDEX / MATCH Practice

1. Redo the VLOOKUPs we did earlier with XLOOKUP or INDEX/MATCH for the **category**, **sales**, and **profit** columns in the `returns` sheet.
2. Redo the **profit_margin** and **margin_category** columns.

Work with a partner, checking in with each other after answering each question.

Aggregate Functions

Discussion:
When Counting Only Gets You So Far

The regional sales director needs to know the number of sales that have been reviewed yet remain unapproved. Looking at the spreadsheet on the right, the ask seems simple enough — you can even count it by hand.

But what if there were over 800 rows of data?
What could you do to speed up this process while ensuring your calculation is accurate?

Reviewed by Sales	Reviewed by Management	Approved	Shipped
x	x	x	x
x	x	x	x
x	x	x	
x	x	x	x
x	x	x	x
x	x	x	x
x	x	x	x
x	x		x
x	x	x	
x	x	x	x
x			
x	x	x	x
x	x	x	x
x	x		x

Meet Aggregate Functions

- **Aggregate functions summarize data** using formulas in Excel.
- They play an integral role when using **PivotTables**.

Commonly Used Aggregate Functions	
• MIN • MAX • SUM • SUMIF • AVERAGE	• COUNT • COUNTIF • COUNTIFS • COUNTA • COUNTBLANK

Aggregate Functions | Average and COUNT

```
=AVERAGE(number1, [number2], ...)
```

Finds the **average** of a range of numbers.

```
=COUNT(value1, [value2], ...)
```

Counts the **number of numeric values** in a range.

Aggregate Functions | Minimum and Maximum Values

```
=MIN(number1, [number2], ...)
```

Finds the **minimum value** of a range of numbers.

```
=MAX(number1, [number2], ...)
```

Finds the **maximum value** of a range of numbers.

Aggregate Functions | SUM and SUMIF

```
=SUM(number1, [number2], ...)
```

Finds the **sum** of a range of numbers.

```
=SUMIF(range, criteria, [sum_range])
```

Finds the **sum of values in a range** that meet the given criteria.

Aggregate Functions | COUNTIF and COUNTIFS

```
=COUNTIF(range, criteria)
```

Counts the number of **values in a range** that meet the given criteria.

```
=COUNTIFS(criteria_range1, criteria1,
[criteria_range2, criteria2]…)
```

Counts the number of values in a range using specific criteria and can take in **many ranges with many criteria**.

Aggregate Functions | COUNTA and COUNTBLANK

```
=COUNTA(value1,[value2], ...)
```

Counts the number of **non-blank cells**, not just the number of numeric cells.

```
=COUNTBLANK(range)
```

Counts the number of **blank cells** in the range.

Group Exercise:

Using Aggregate Functions | Your Turn

In a small group, practice together the tasks below using the `orders` worksheet. You may use your existing analysis worksheet or create a new one. Ready, set, go!

- `MIN` to find the lowest sale.
- `MAX` to find the highest order quantity.
- `AVERAGE` to find the average profit margin.
- `SUM` to find the total sum of profits.
- `SUMIF` to find the total sales in the “Technology” category.
- `COUNT` to count the number of orders.
- `COUNTIF` to count the number of orders from the “Home Office” customer category.
- `COUNTIFS` to count the number of furniture orders in the “Consumer” customer category.
- `COUNTA` to count the number of returns (use the “Reason_returned” column).

Group Exercise:

Stretch: Using Aggregate Functions: Answers

	Formula (columns may be different from yours)	Value
Lowest sale	`=MIN(orders!N:N)`	$0.44
Highest order quantity	`=MAX(orders!R:R)`	14
Average profit margin	`=AVERAGE(orders!P:P)`	0.67%
Total sum of profits	`=SUM(orders!O:O)`	$ 19,884
Total sales in the “Technology” category	`=SUMIF(orders!J:J,"Technology",orders!N:N)`	$ 2,230,546
Count number of orders	`==COUNTA(UNIQUE(B:B))`	23,746
Home Office orders	`=COUNTIF(orders!AC:AC,"Home Office")`	768
Number of **Consumer** furniture orders	`=COUNTIFS(orders!AC:AC,"Consumer",orders!J:J,"Furniture")`	2301
Number of returns	`=COUNTA(returns!E:E)-1`	1177

Cleaning and Aggregating Data With Excel

PivotTables

Discussion:

Connecting Hypotheses to Analysis

Let's take a few minutes to revisit our hypothesis-driven questions.

List out hypotheses:

1. Do different types of customer segments make more returns?
2. Do different categories get returned more often?
3. Does shipping speed impact whether an item gets returned?

How would we answer these?

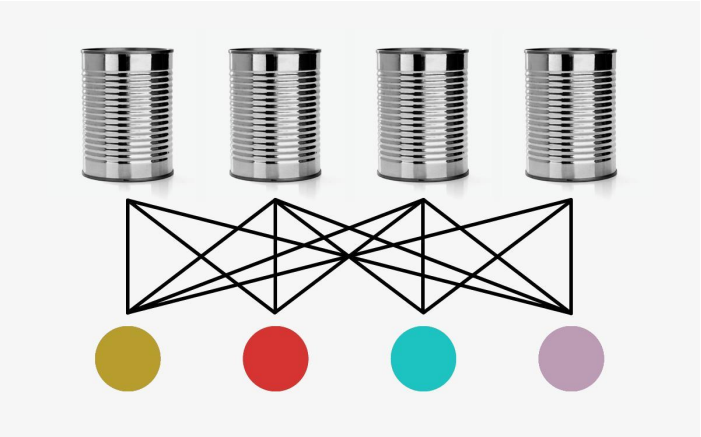
Use the "class_exercises" tab to keep track of the above!

Meet PivotTables

We want to see the **average sale per customer segment.**

With a bit of googling, we could find the **AVERAGEIF()** function. But what if we think we'll add sales in future from new customer segments?

PivotTables allow you to quickly create **dynamic aggregations, slices, and filters**.



Four Components of a PivotTable

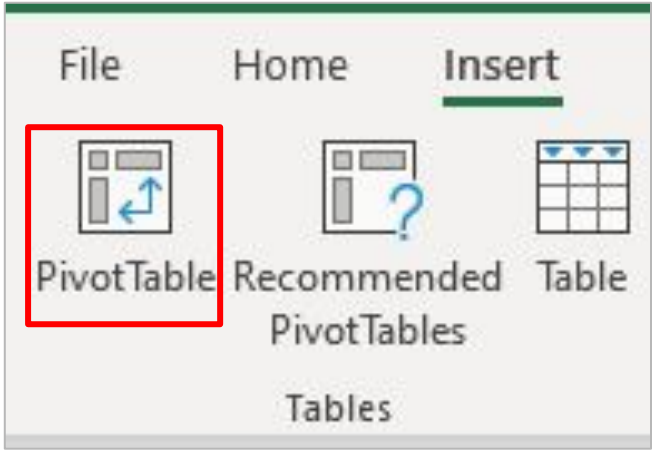
Components	Questions to Ask of Data
Filter	Which data should we include in our PivotTable?
Rows	What unique data values do we want to have as rows in our table?
Columns	What unique data values do we want to have as columns in our table?
Values	What values will be in the cells of our table?

Computers Out:

Creating a PivotTable

Select any cell in the “orders” sheet. Then:

- Create a PivotTable by clicking “**PivotTable**” on the “Insert” ribbon.
- Verify that the table/range auto-selected is the **orders** data. If not, you can change it before moving on.

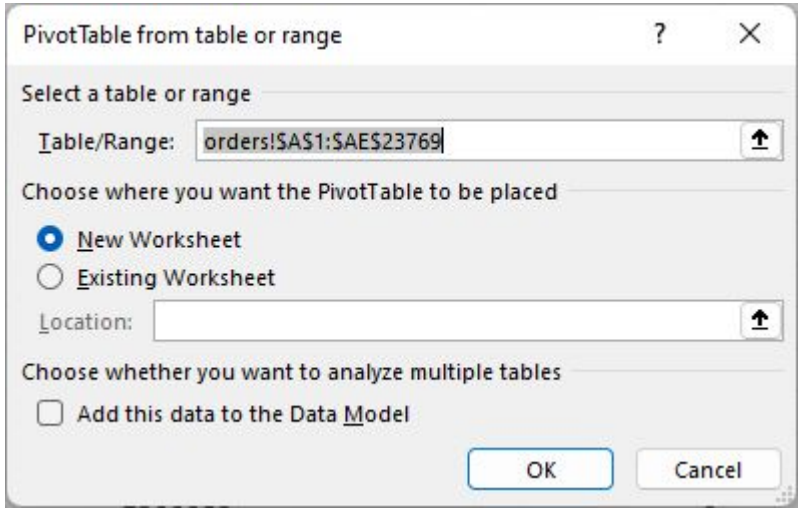


Computers Out:

Creating a PivotTable | Setting It All Up

Decide where you want to put the PivotTable, and...

- Create a **new worksheet** for this example (this is the default option).
- Click **OK**



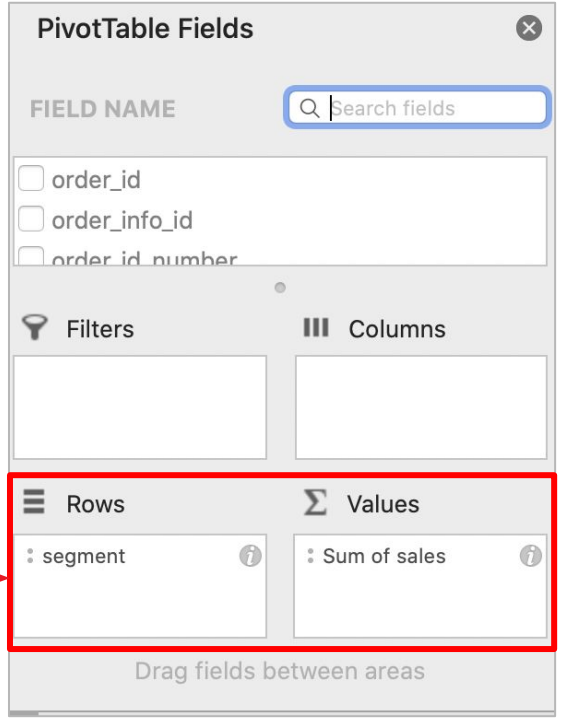
Guided Walk-Through:

Finding the Average Sales | PivotTable Field

Let's find the ***average sale per customer segment*** in our **orders** data:

1. Drag *"segment"* to the **Rows** field.
2. Drag *"sales"* to **Values** field.

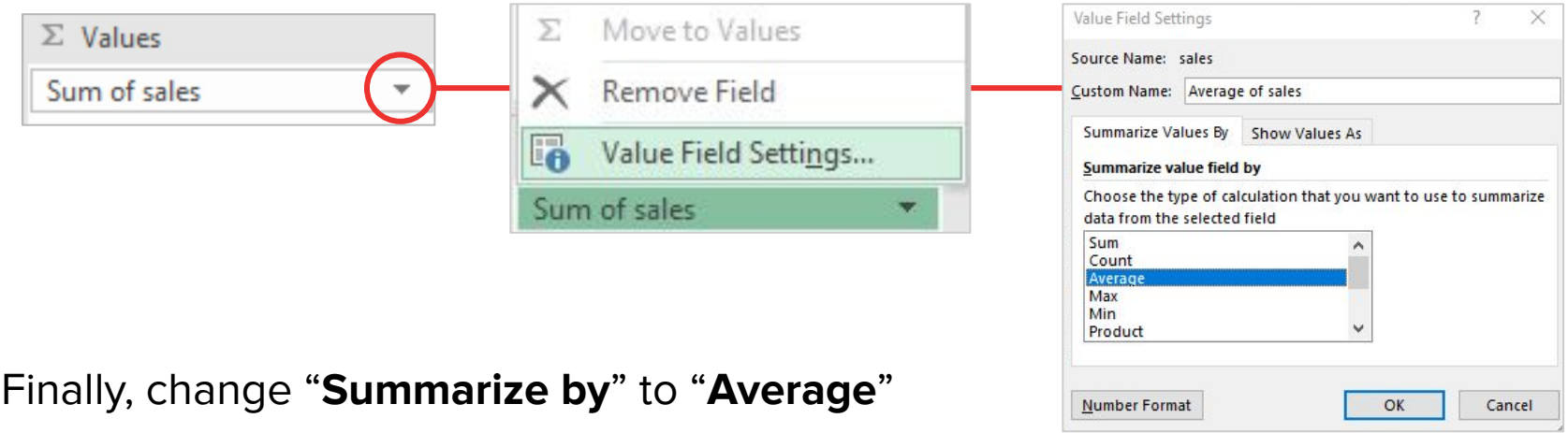
Your PivotTable Fields list should now look like this.



Guided Walk-Through:
Finding the Average Sales | Value Field Settings

3. To change the “Sale” value from “**sum**” to “**average**”:
   - on a Mac, click “i” next to “Sum of sales” in Values
   - on a PC, click the small triangle

Then click “**Value Field Settings…**”



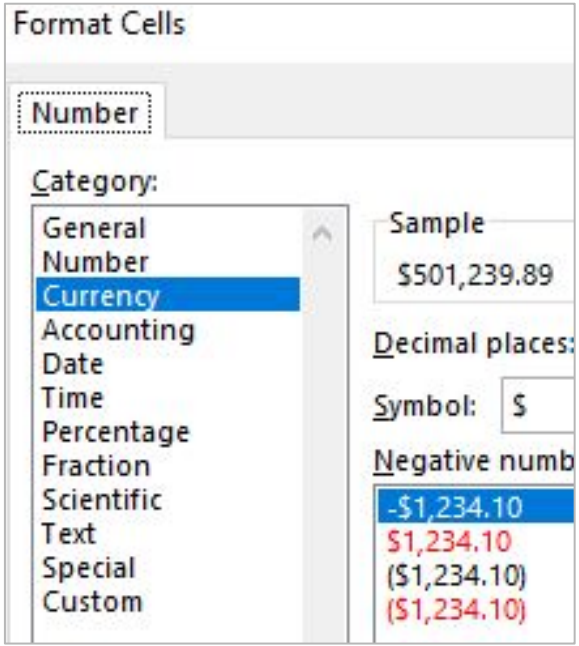
Finally, change “**Summarize by**” to “**Average**”

Guided Walk-Through:
Finding the Average Sales | Formatting

This will properly take in the average sales, but first let's make sure our formatting is correct.

These averages are dollar amounts, so

- Go to "**Number Format** > **Number**."
- Select "**Currency**" (or click the "$" button).



Guided Walk-Through:

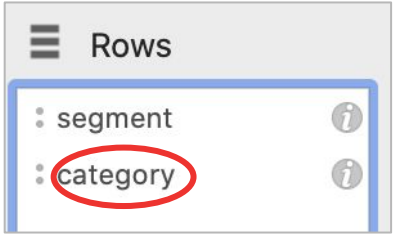
Finding the Average Sales *and* Categories

Next, let's find the **average sale per segment AND per category**. Drag "category" into the Rows field above segment.

Row Labels	Average of sales
Furniture	**$403.37**
Consumer	$410.29
Corporate	$392.77
Home Office	$401.49
Office Supplies	**$123.73**
Consumer	$119.43
Corporate	$132.88
Home Office	$121.12
Technology	**$483.12**
Consumer	$458.56
Corporate	$510.59
Home Office	$506.80
Grand Total	**$247.30**

For every segment, we now see the **average for each segment and each category**. But, what if we want category and *then* segment?

Let's change the order and look at the results.



Guided Walk-Through:

Finding the Average Sales *and* Categories

It'd be much easier to parse and compare this data if the segments were presented down the rows and the categories were presented across the columns.

Let's move "**category**" to Columns.

Segment

Category

Average of sales	Column Labels			
Row Labels	**Furniture**	**Office Supplies**	**Technology**	**Grand Total**
Consumer	$410.29	$119.43	$458.56	$241.55
Corporate	$392.77	$132.88	$510.59	$255.13
Home Office	$401.49	$121.12	$506.80	$250.40
Grand Total	**$403.37**	**$123.73**	**$483.12**	**$247.30**

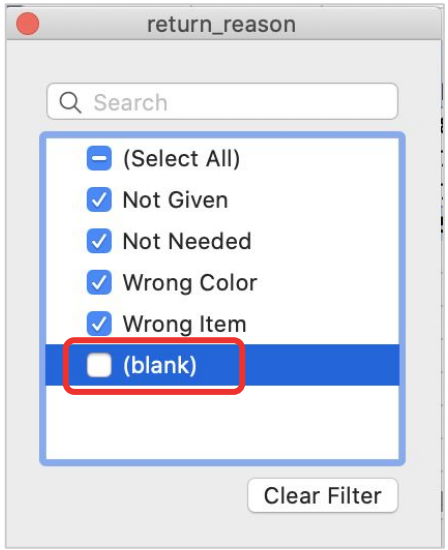
Guided Walk-Through:

Average Sales by Segment, Category, *and* Returned

Ready to get even more info from our data? Let's find the ***average sales per segment and category for only returns.***

The structure of our PivotTable is going to remain the same, but we need to filter some of the data out of the calculations.

- Let's move "**return_reason**" to the Filter field.



Notice that a dropdown for "return_reason" has been added above the PivotTable.

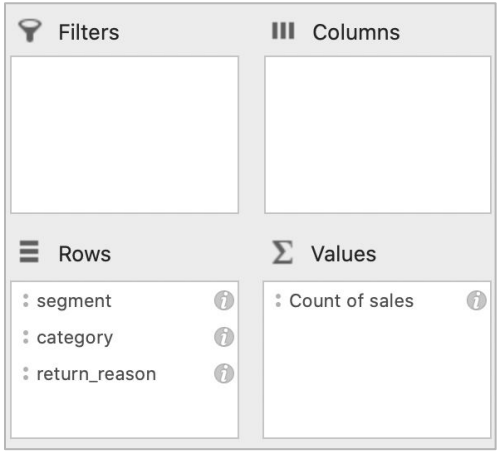
- De-select "**(blank)**" and notice that the numbers now change to only display the orders that have been returned.

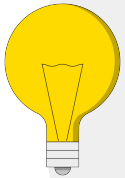
Guided Walk-Through:

Number of Records by Segment, Category, *and* Returned

Let's identify the ***number of orders by segment, category, and return_reason***:

1. Move "**segment**", "**category**", and "**return_reason**" to Rows.
2. Change "**Average of sales**" to "**Count of sales**".
3. Reformat this into a number.



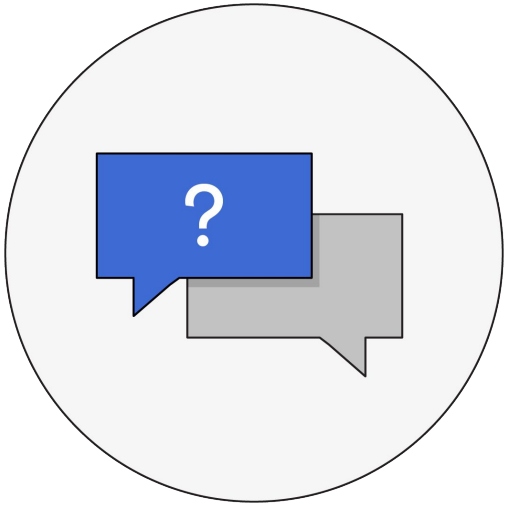


Note: When doing a count, you can use any of the variables for Values, as long as every member of that column has values of some kind.

Discussion:
Guided Walk-Through Recap

How did it go? Let's recap by reflecting on the following. Share your response to the questions below with your class.

1. **How well (or otherwise) does the PivotTable work for this data set? Why?**
2. **What does each row define?**
3. **What does each row contain?**



COUNTing With PivotTables

Discussion:
True or False?

Take a look at this PivotTable from our "Orders" data set.

Row Labels	Count of sales
Consumer	12067.00
Corporate	6968.00
Home Office	4733.00
Grand Total	**23768.00**

True or False:

The consumer segment of the business sold 12,067 dollars worth of products.

Be ready to defend your answer.

Discussion:

The Answer Is...

The consumer segment of the business sold 12,067 dollars worth of products? **False**.

- **12,067** is the *number of rows in the order sheet* from the consumer segment.
- There is one row per order, per product. For example, if a single order includes three distinct products (staplers, pencils and paper), there are three rows. But if a single order contains only one product - even if there are multiples of that product (e.g. 10 of the same pencils) - there will be only one row.
- So, the count is not related to the dollar amount at the register, and it is not simply the number of orders. It's hard to describe simply, but you can think of it as the **number of times different products were ordered.**

Discussion:

What Are We Counting?

This PivotTable is counting the **number** of **cells** in the sales column that contain a value.

Given this information, are the following statements True or False?

1. **There is a total of 23768 sales across segments.**
2. **"Home Office" generated the lowest revenue among the three segments.**

Row Labels	Count of sales
Consumer	12067.00
Corporate	6968.00
Home Office	4733.00
Grand Total	**23768.00**

Discussion:

The Answers Are...

1. There are a total of 23,768 sales across segments — **true, but it depends on how you define a “sale”!**
    a. There is a total of 23,768 rows (across all three segments) that have a value in the “Sales” column.
    b. The number of orders is *less* than this. If each sale is of a distinct product within a distinct order, then this is the number of sales.
2. “Home Office” generated the lowest revenue among the three segments — **we can’t tell!**
    a. Because the number of rows does not represent dollars, the count of sales for “Home Office” does not reflect the amount of revenue (dollars) generated.

To ensure an accurate, streamlined PivotTable, keep in mind that...

Any values that we want in the rows or columns should be discrete, categorical variables.

Solo Exercise:

PivotTables Practice | Prompts

In your existing analysis sheet or in a new worksheet, answer these questions:

1. **Do rates of returns differ significantly by profit margin category?** *(Hint: Remember, we created a margin category definition in the last lesson.)*

2. **What are the minimum, average, and maximum profits for orders by margin category?** *(Extra practice: Convert all to currency.)*

3. **What percentage of all orders occurred in each customer category, broken down by profit margin category?**

4. **What is the highest performing subcategory by sales in the "Consumer" customer category?** *(Hint: Use a PivotTable filter!)*

Solo Exercise:

PivotTables Practice | Methods and Solutions

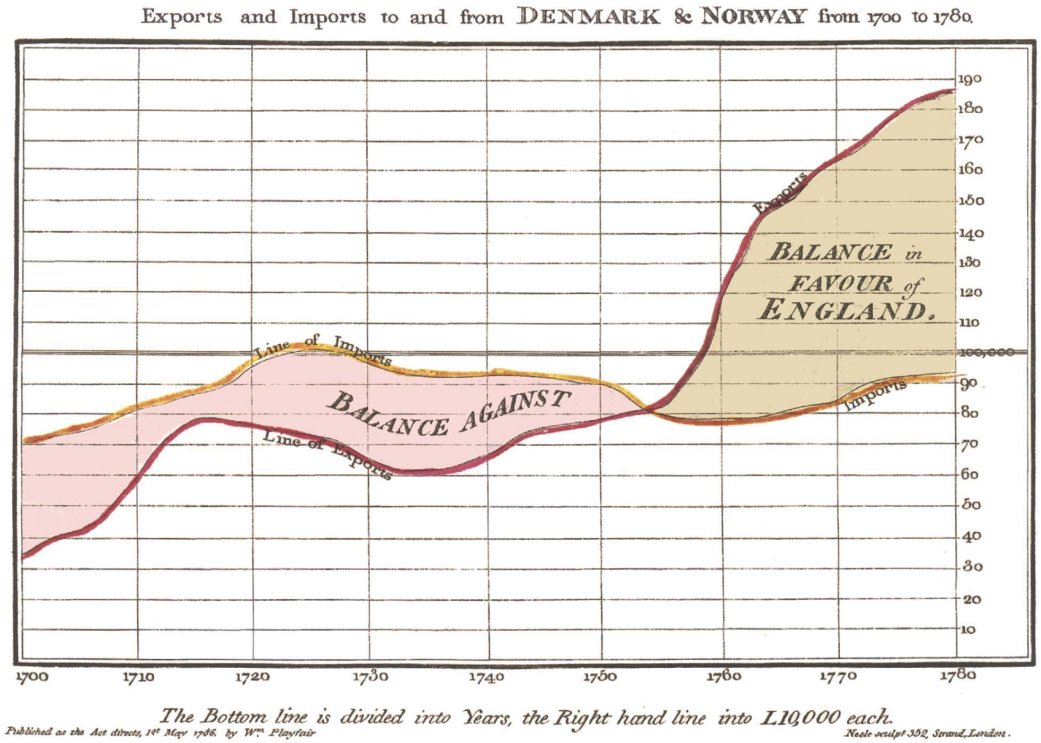
How did you reach your answers? Check out the following methods:

1. Rows: margin_category Values: Count of Sales	**2.** Rows: profit_margin_category Values: Min of Profit, Max of Profit, and Average of Profit
3. Rows: profit_margin_category Columns: segment Values: count of order ID (set to show values as % of grand total)	**4.** Rows: sub_category Values: sum of sales Filter: Consumer segment

Cleaning and Aggregating Data With Excel

Data Visualization

The History of Data Visualization



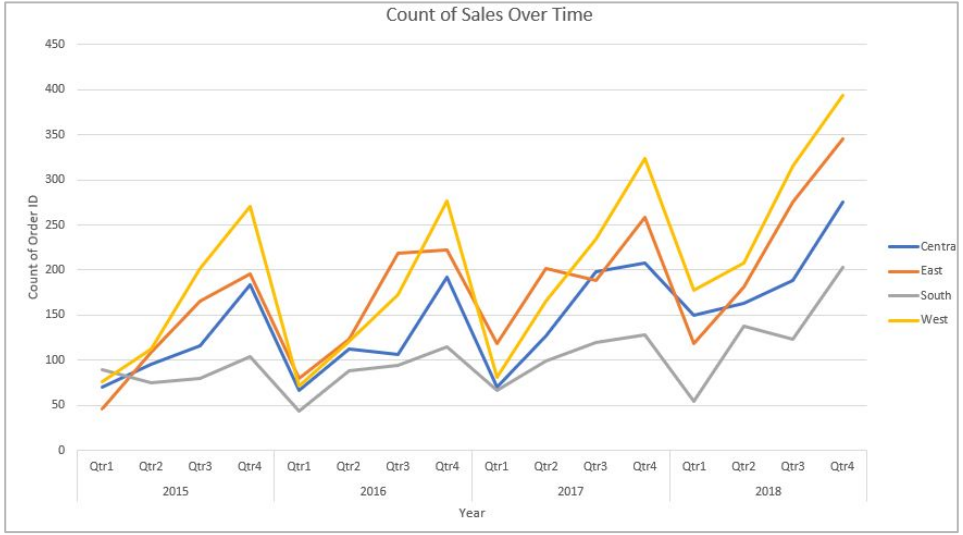
William Playfair's 1786 **time series chart** showing the trade balance between Norway and Denmark.

Source: William Playfair

You Have Exactly 60 Seconds...

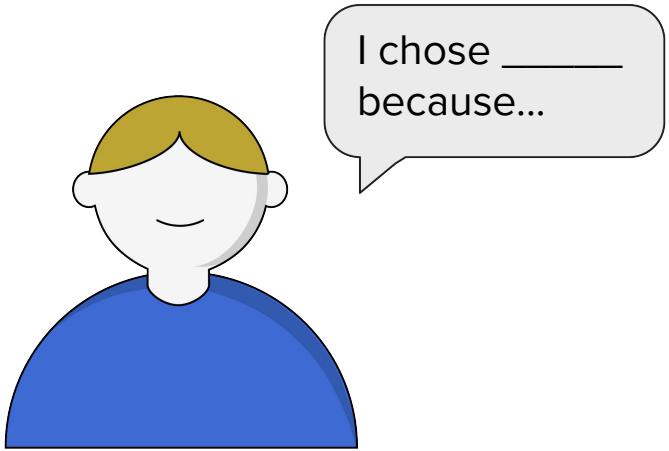
Count of order_id	Column Labels				
Row Labels	Central	East	South	West	Grand Total
2015					
Qtr1	70	46	90	76	282
Qtr2	96	109	75	112	392
Qtr3	116	166	80	202	564
Qtr4	184	196	104	271	755
2016					
Qtr1	66	80	43	71	260
Qtr2	112	123	88	121	444
Qtr3	106	219	94	173	592
Qtr4	192	222	115	277	806
2017					
Qtr1	70	118	66	81	335
Qtr2	127	202	99	166	594
Qtr3	198	188	120	234	740
Qtr4	208	258	128	324	918
2018					
Qtr1	150	118	54	178	500
Qtr2	163	181	138	208	690
Qtr3	189	276	123	315	903
Qtr4	276	346	203	394	1219
Grand Total	2323	2848	1620	3203	9994

...to write one sentence about the sales trends from 2015 to 2018.



Discussion:
Which Did You Choose?

By a show of hands, how many of you used the table on the right and how many used the image on the left? Tell us why!

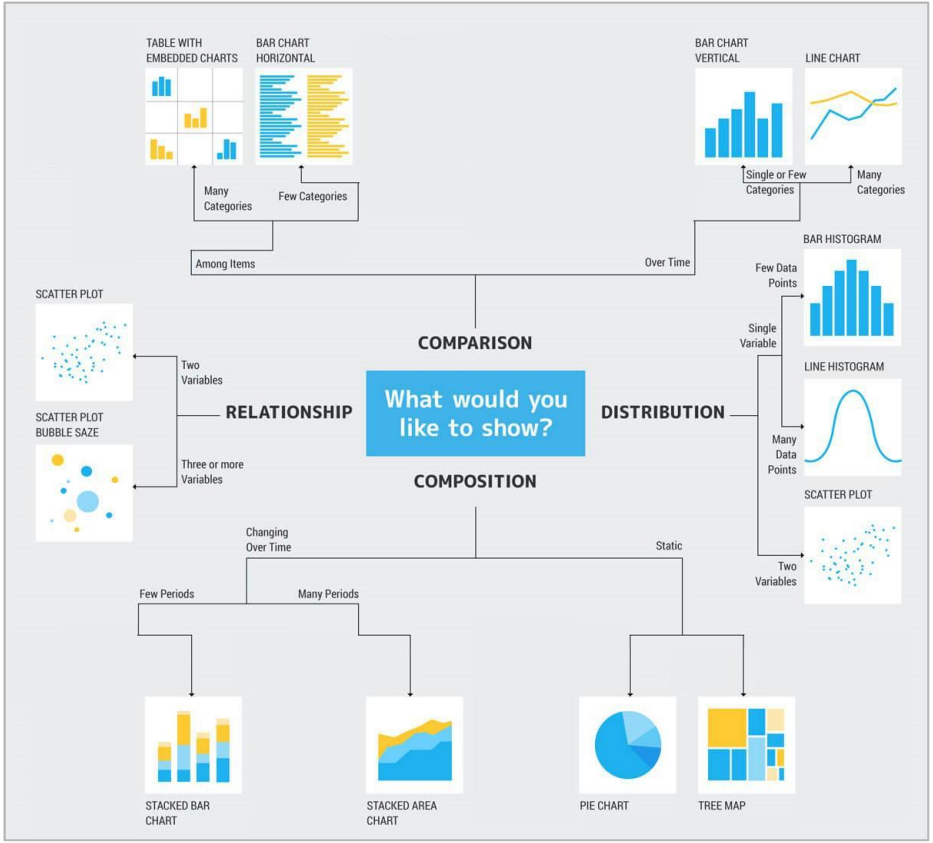


Selecting the Right Visualization for Your Data

Visualization Types

Data visualization methods can be organized by **four distinct categories (use cases)**:

1. Comparison
2. Distribution
3. Composition
4. Relationship



Source: Dataversity

Discussion:

The World as 100 People

Use the four categories — comparison, distribution, composition, relationship — and associated charts as a reference when analyzing the examples provided on this and the following slides.

1. **What chart type did they use?**
2. **What information were they trying to convey?**
3. **Was this effective? Why or why not**

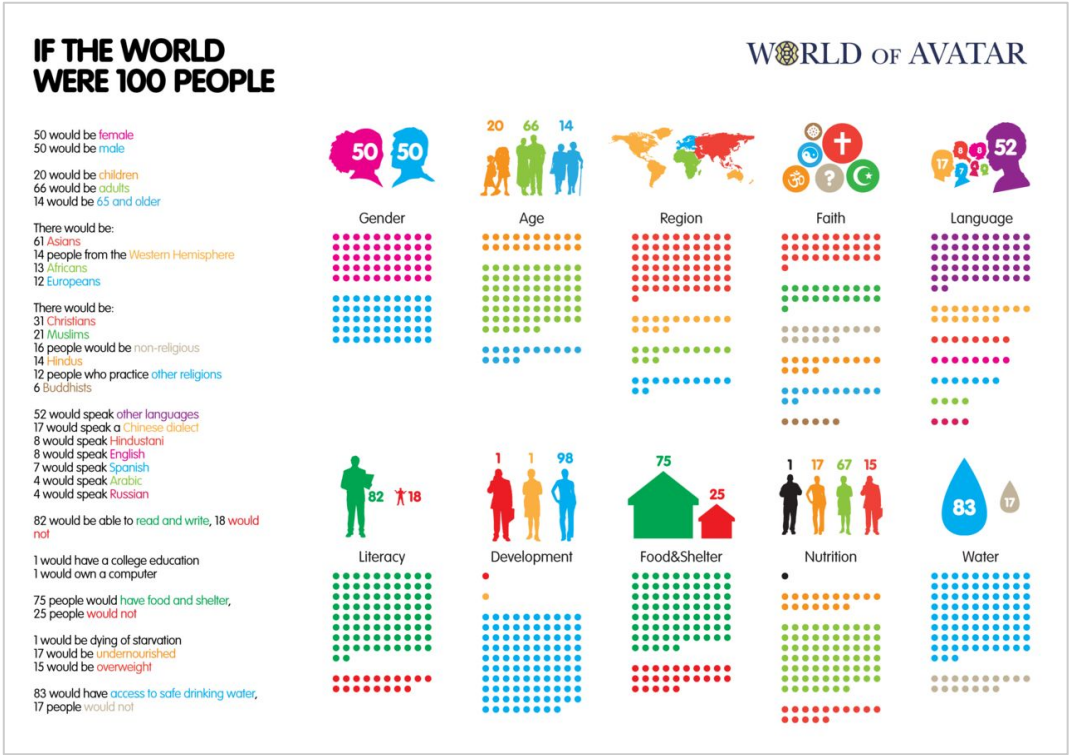
Source: jackhagley.com

Real Cases:

If the World Were 100 People

Here's another way to visualize this information.

What are some of the pros and cons you see?



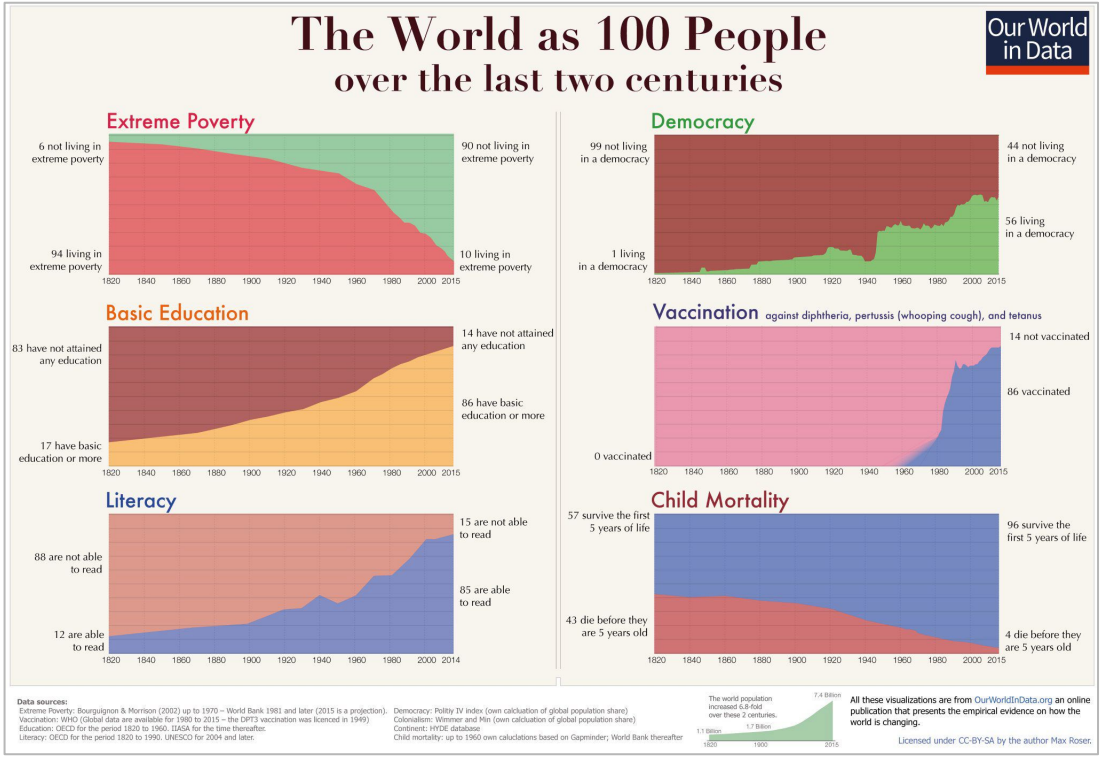
Source: lifehack.org

Real Cases:

The World as 100 People Over the Last Two Centuries

Yet another way to visualize similar information.

What works? What doesn't?



Comparing Data

Visualization Decision Table

You can also choose visualizations based on the number of **quantitative** and **categorical variables** you want to compare at a time.

		Quantitative Data Points			
		1	**2**	**3**	**4+**
Categorical Data Points	**0**	Histogram Box plot Heat map Strip plot Bullet chart Triangle plot Gauge charts	Scatterplot Small multiples: Line graph	Standard Bubble Small multiples: Line graphs	Table Small multiples: Line graphs
	1	Bar chart Pie chart Line chart Dot plot Treemap Small multiples: - Histograms - Bar charts	Scatterplot Dot plot Small multiples: Histograms	Standard Bubble Dot plot Scatterplot matrix	Table Scatterplot matrix

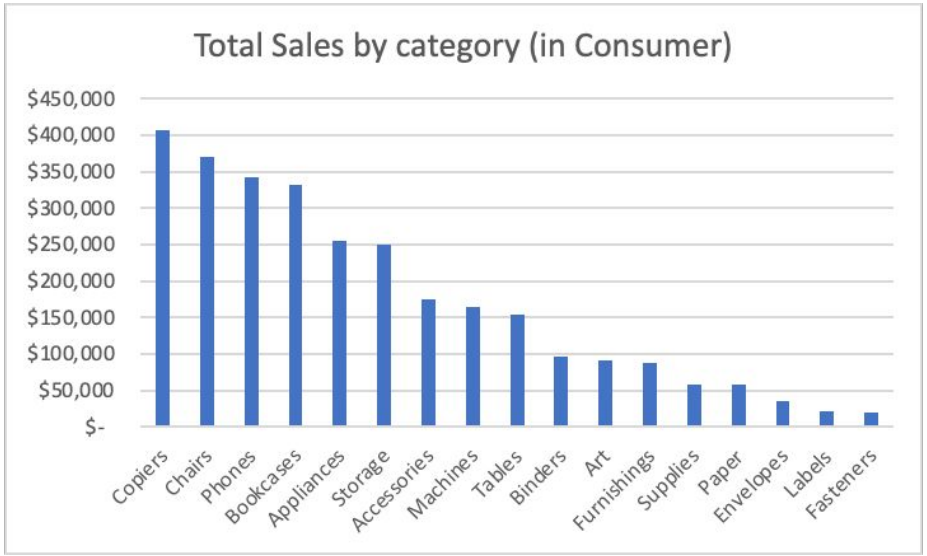
Adapted from Boxplot Analytics

Comparing Data: Bar Charts

What Are Bar Charts?

A type of graph used to **show and compare** the number, frequency, or other measure (e.g. mean) for different numerical and categorical data.

They're the **most commonly used** types of graph because they are simple to create and easy to interpret.

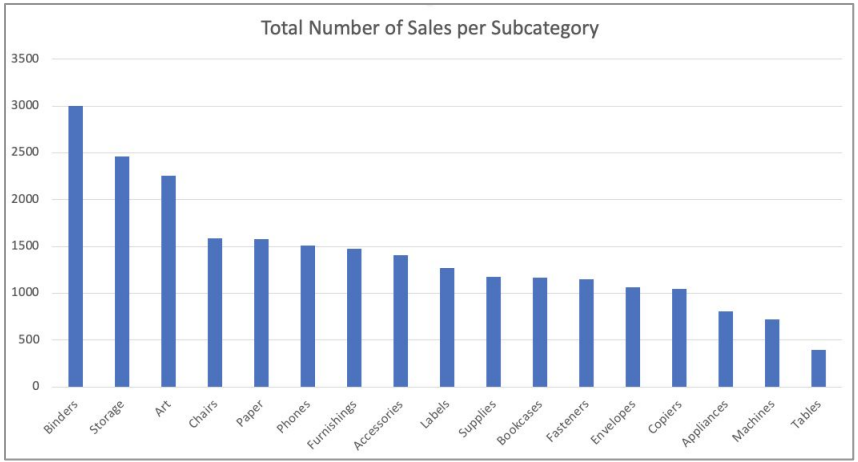


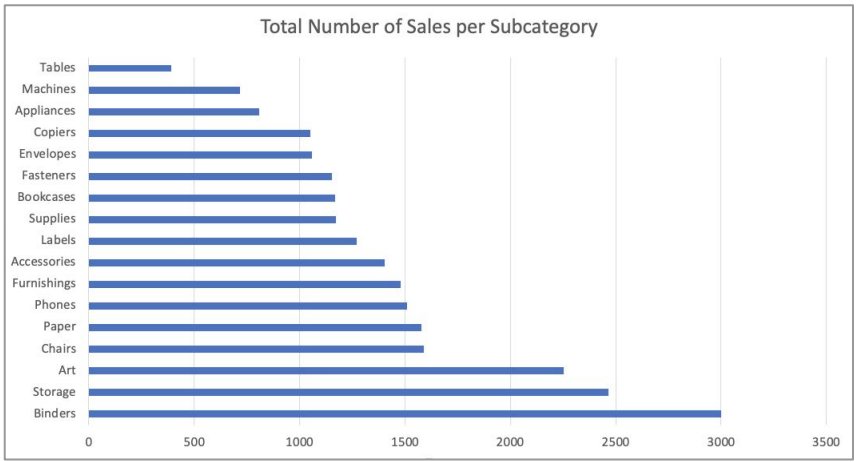
Discussion:

Bar Chart | Traditional vs. Horizontal

These charts feature the same data, but differ in how the data is presented.

Which one is actually a bar chart?





When to Use Bar Charts

When you have:

- At least one **categorical variable (x axis)**.
- At least one **quantitative variable (y axis)**.

Best for:

- Quickly displaying simple comparisons across categorical variables.

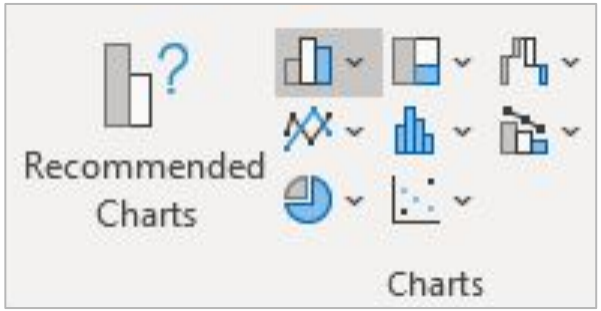
		Quantitative Data Points			
		1	2	3	4+
Categorical Data Points	0	Histogram Box plot Heat map Strip plot Bullet chart Triangle plot Gauge charts	Scatterplot Small multiples: Line graph	Standard Bubble Small multiples: Line graphs	Table Small multiples: Line graphs
	1	**Bar chart** Pie chart Line chart Dot plot Treemap Small multiples: - Histograms - Bar charts	Scatterplot Dot plot Small multiples: Histograms	Standard Bubble Dot plot Scatterplot matrix	Table Scatterplot matrix

Adapted from Boxplot Analytics

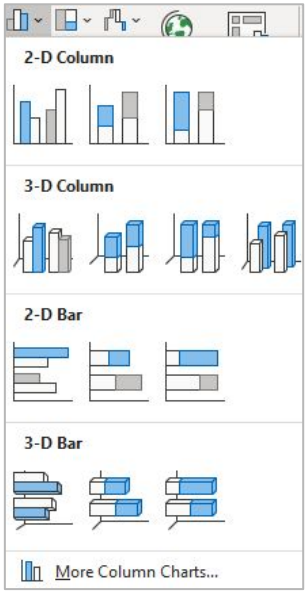
Computers Out:

Navigating Bar Charts in Excel

1. We'll use this part of the ribbon to create our data visualizations.



2. Clicking one of the chart types will display a dropdown menu of the different types of charts.



Guided Walk-Through:

Creating a Bar Chart in Excel

We have our PivotTables from the last lesson. Let's use one of them to create a visualization that will answer this question:

When sales go up, does the profit margin also go up?

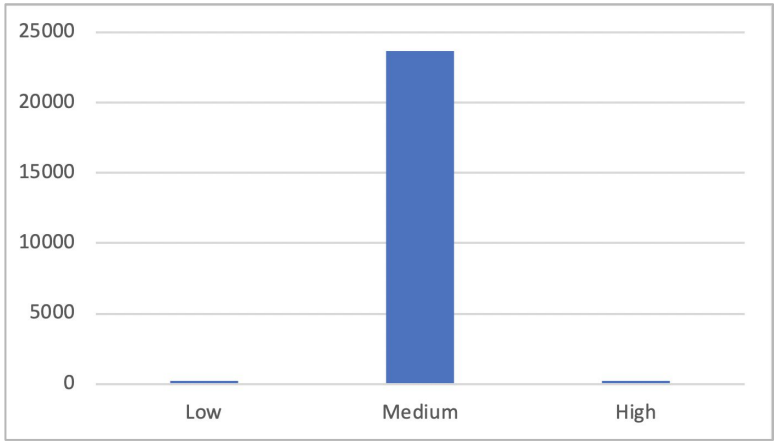
Open a new sheet and call it `charts` — or, if you prefer, you can keep all PivotTables on one sheet and all charts on another.

1. Select any cell in the "**Sales by Profit Margin**" PivotTable.
2. Go to the "**Insert**" tab > "**Chart**" and select a type of bar chart.

What Is Chart Formatting?

Chart formatting allows us to easily understand what this chart represents on first glance.

- **Chart title:** header on a chart
- **Axis title:** labels on each axis
- **Data labels:** labels on each data point



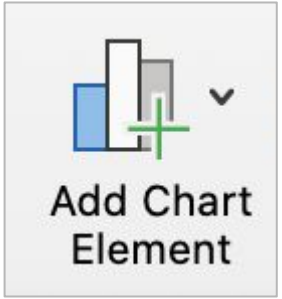
Guided Walk-Through:

Adding Chart Formatting in Excel

Let’s add formatting to this chart.

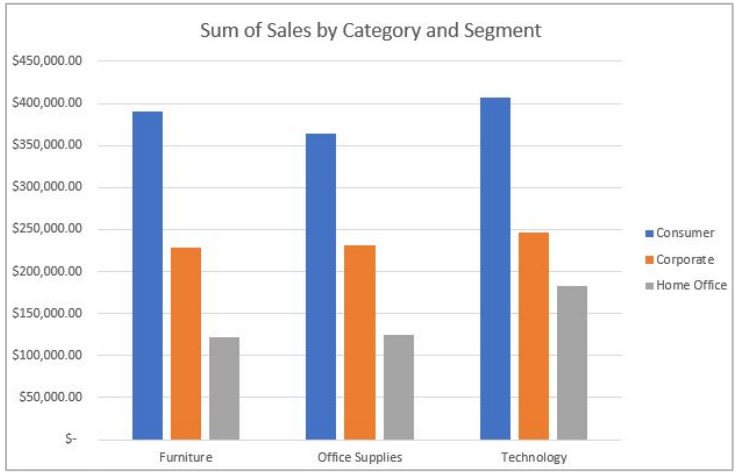
Select the chart, then click on the “**Design**” tab at the top of your workbook.

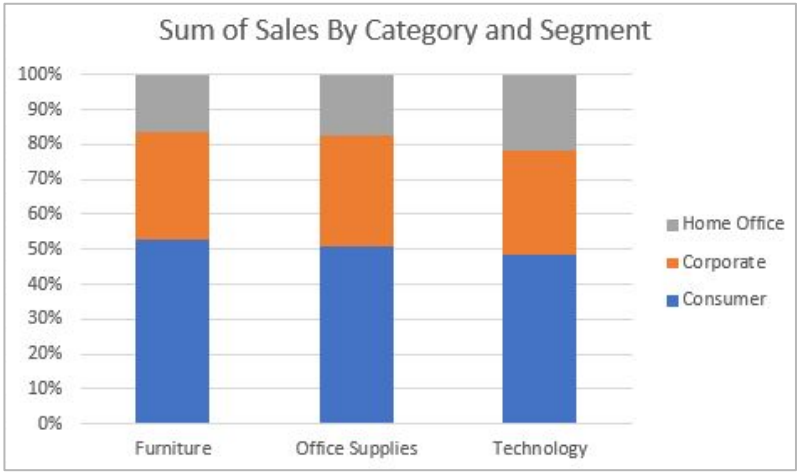
1. **Chart title:** Click the “Add Chart Element” button on the left side of the ribbon. Choose “Chart Title” > “Above Chart.”
2. **Axis titles:** Click the “Add Chart Element” button on the left side of the ribbon. Choose “Axis Titles” > “Primary Horizontal” to name the x axis, then repeat for the y axis by choosing “Primary Vertical.”
3. **Data labels:** Click “Add Chart Element” button on the left side of the ribbon. Choose “Data Labels” > “Outside End.”



Bar Chart | Multi-Column vs. Stacked

A **multi-column** bar chart shows the *whole number* of each numerical variable, whereas a **stacked** bar chart shows *parts* of a whole.





Discussion:

Which Chart Should We Use?

Take a look at the pivot table below.

In this situation, we're looking at the **average profit margin (a percentage) by the customer segments and product categories**.

Which type of chart should we use?

Average of profit_margin	Column Labels		
Row Labels	Furniture	Office Supplies	Technology
Consumer	0.4%	0.8%	0.5%
Corporate	0.3%	0.6%	0.5%
Home Office	0.4%	1.1%	0.7%

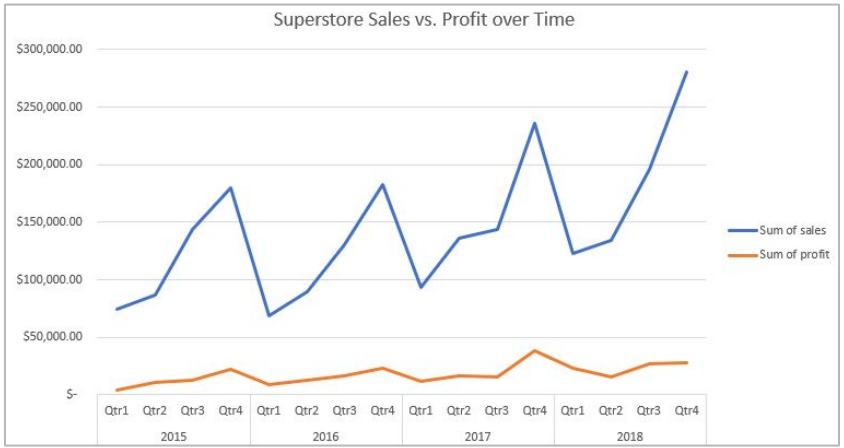
Comparing Data: Line Charts

What Are Line Charts?

Also known as a line graph, it's used to visualize **the value of something over time**.

Consists of at **least two axes**, x and y:

- The x axis represents independent variables.
- Data plotted to the y axis are dependent on x.



When to Use Line Charts

When you have:

- Time series data.
- A quantitative variable.

Best for:

- Displaying changes and patterns over time.

		Quantitative Data Points			
		1	2	3	4+
Categorical Data Points	0	Histogram Box plot Heat map Strip plot Bullet chart Triangle plot Gauge charts	Scatterplot Small multiples: Line graph	Standard Bubble Small multiples: Line graphs	Table Small multiples: Line graphs
	1	Bar chart Pie chart Line chart Dot plot Treemap Small multiples: - Histograms - Bar charts	Scatterplot Dot plot Small multiples: Histograms	Standard Bubble Dot plot Scatterplot matrix	Table Scatterplot matrix

Adapted from Boxplot Analytics

Discussion:

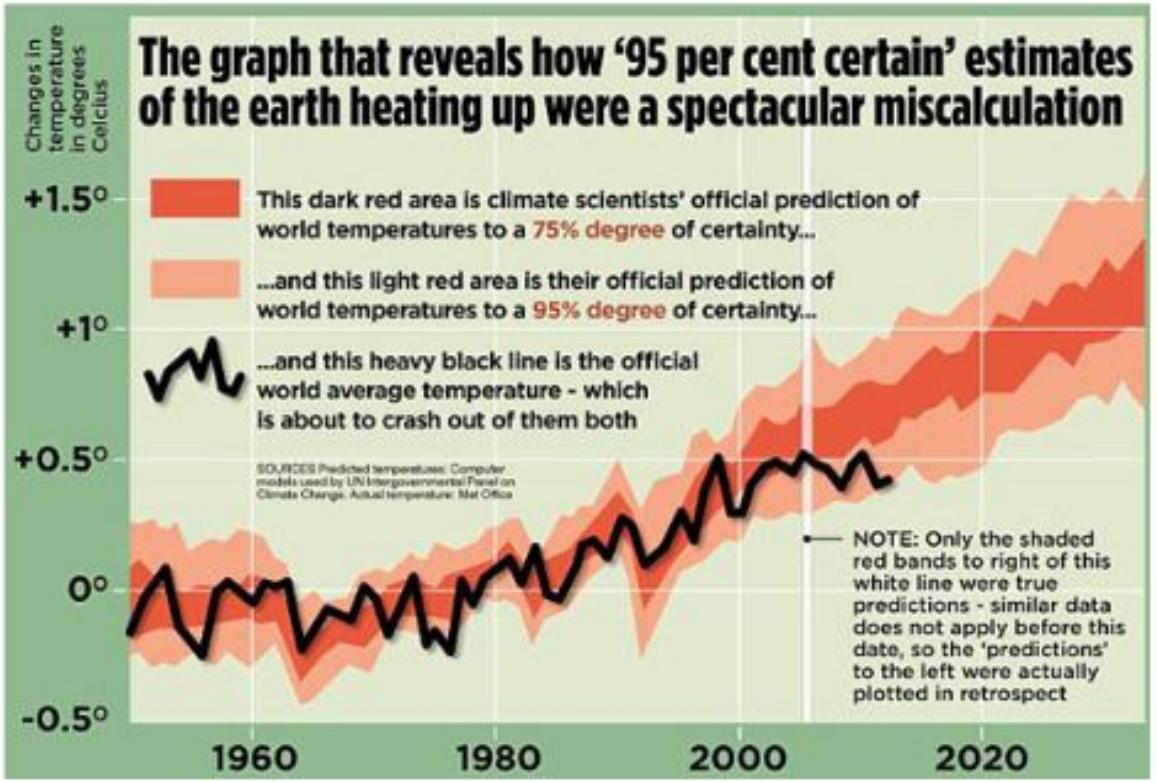
What Can a Line Chart Tell Us?

3 minutes

Let's take a look at this line chart.

What's going on here?

What information can you glean from it?



Source: Statistics How-To

Guided Walk-Through:
Let's Try It!

You're working on a business report that includes trends in sales and profits over time. **How do you create a line chart that shows and compares both trends?**

Try making a new PivotTable with time data. Go to your `orders` table and insert a pivot chart in a new worksheet with these fields:

- **Rows**: "**Order Date**" grouped by month
- **Values**: "**Sum of Sales**," "**Sum of Profit**"

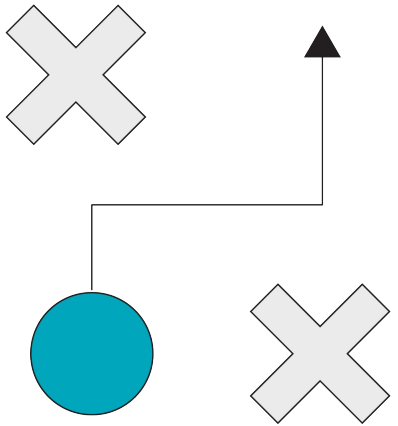
1. Select any cell in your PivotTable.
2. Insert > Chart > Line

Partner Exercise:

Compare and Contrast

So what does your line chart look like? Turn to your partner and compare.

- If your line charts look the same, hooray! Talk about other types of values you can compare using a line chart.
- If your line charts look different, no need to panic! Backtrack your steps to figure out where your approach differed.
- **Why does profit look so small?**



Displaying Relationships Between Two Variables: Dual Axis (Secondary Axis)

What Is a Dual Axis Chart?

Displays the relationship between **two variables using two different chart types**.

As a best practice, both variables should **use the same scale** so as not to confuse the audience.

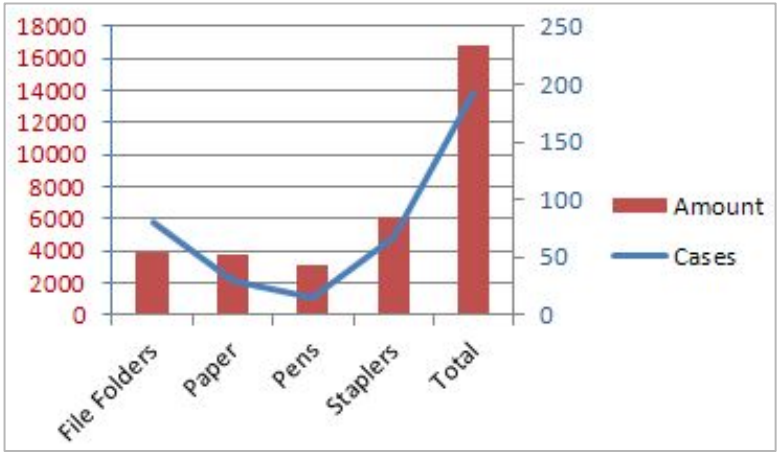
When to Use Dual Axis

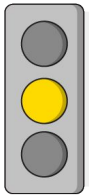
When you have:

- Two sets of categorical variables with matching quantitative data.

Best for:

- Displaying two sets of information together to show a relationship.





Caution: Many data visualization books actually discourage the use of dual axis charts because they can be confusing.

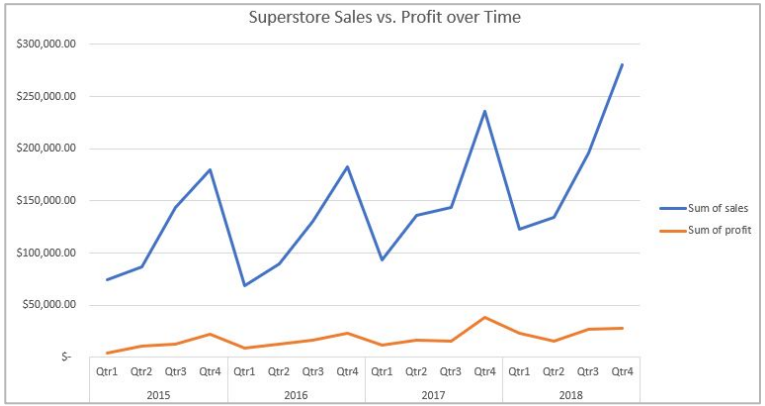
Source: Contexture.com

What Are Secondary Axes?

Secondary axes allow us to compare two variables that are on different scales.

They consist of **three axes**, x axis and two y axes:

- The primary y axis is on the left side of the chart.
- The secondary y axis is on the right side of the chart.



Guided Walk-Through:
Let’s Try It!

Let’s change the axis for “Sum of Profit” to secondary to better show how the trends between “Sales” and “Profit” do or do not relate to each other.

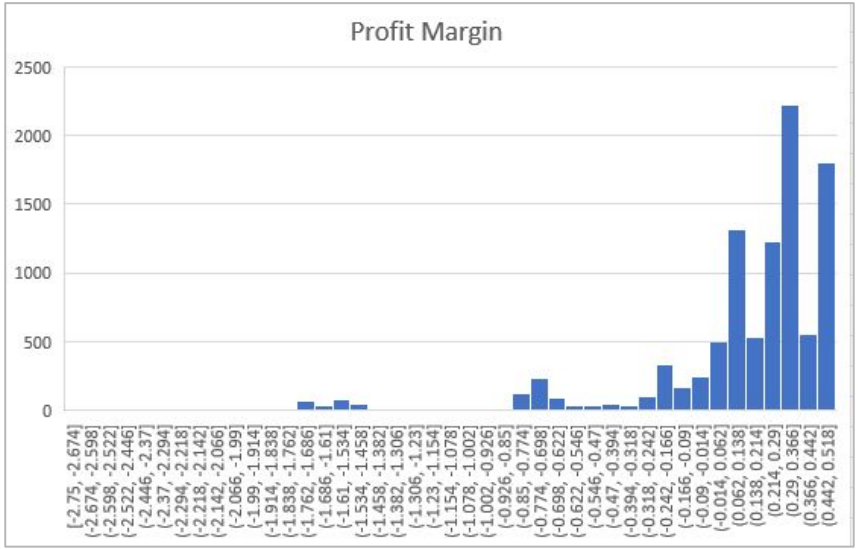
- Right click on the profit line and choose “Format Data Series.”
- In the formatting pane, choose “Secondary Axis.”

Analyzing Distributions: Histograms and Scatterplots

What Are Histograms?

Histograms display the **shape and spread of continuous data** using bars of different heights:

- Each bar groups numbers into ranges.
- Taller bars show that more data falls in that range.



When to Use Histograms

When you have:

- One continuous **quantitative variable** that can be broken down into “bins.”

Best for:

- Displaying the distribution of frequencies across a data set.

		Quantitative Data Points			
		1	2	3	4+
Categorical Data Points	0	**Histogram** Box plot Heat map Strip plot Bullet chart Triangle plot Gauge charts	Scatterplot Small multiples: Line graph	Standard Bubble Small multiples: Line graphs	Table Small multiples: Line graphs
	1	Bar chart Pie chart Line chart Dot plot Treemap Small multiples: - Histograms - Bar charts	Scatterplot Dot plot Small multiples: Histograms	Standard Bubble Dot plot Scatterplot matrix	Table Scatterplot matrix

Adapted from Boxplot Analytics

What Are Scatterplots?

A chart generated by mapping **numeric** values to a pair of perpendicular axes.

Uniform Dots

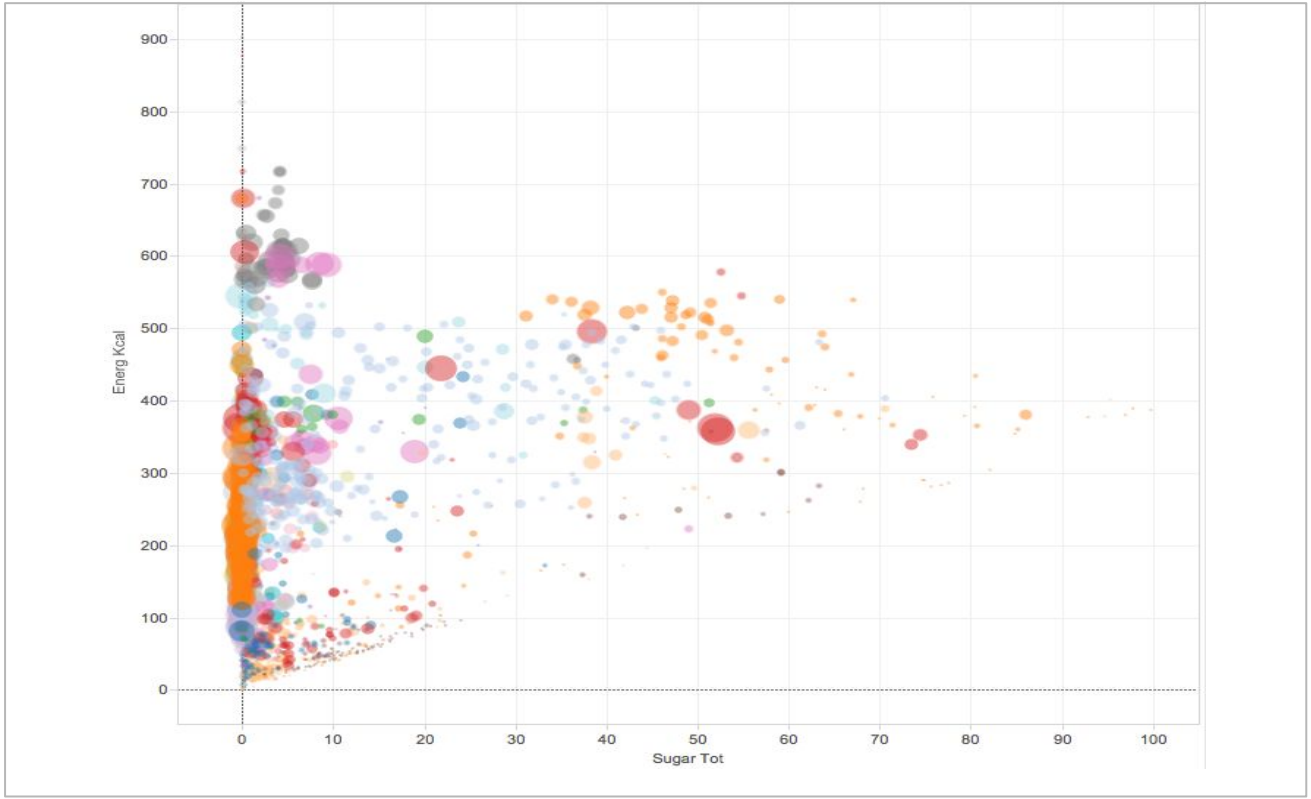
Each dot represents a data point at the intersection of two attributes.

With a Trend Line

Scatterplots With Colored Dots

Scatterplots | Bubble Charts

Scatterplot Variant | Combining Color and Size



When to Use Scatterplots

When you have:

- Two or more **quantitative variables** that can be broken down into “bins.”
  - Do not need categorical variables but can be added for multiple series.

Best for:

- Displaying the relationship between two **quantitative variables**.

		Quantitative Data Points			
		1	2	3	4+
Categorical Data Points	0	**Histogram** Box plot Heat map Strip plot Bullet chart Triangle plot Gauge charts	**Scatterplot** Small multiples: Line graph	Standard Bubble Small multiples: Line graphs	Table Small multiples: Line graphs
	1	Bar chart Pie chart Line chart Dot plot Treemap Small multiples: - Histograms - Bar charts	**Scatterplot** Dot plot Small multiples: Histograms	Standard Bubble Dot plot **Scatterplot matrix**	Table **Scatterplot matrix**

Adapted from Boxplot Analytics

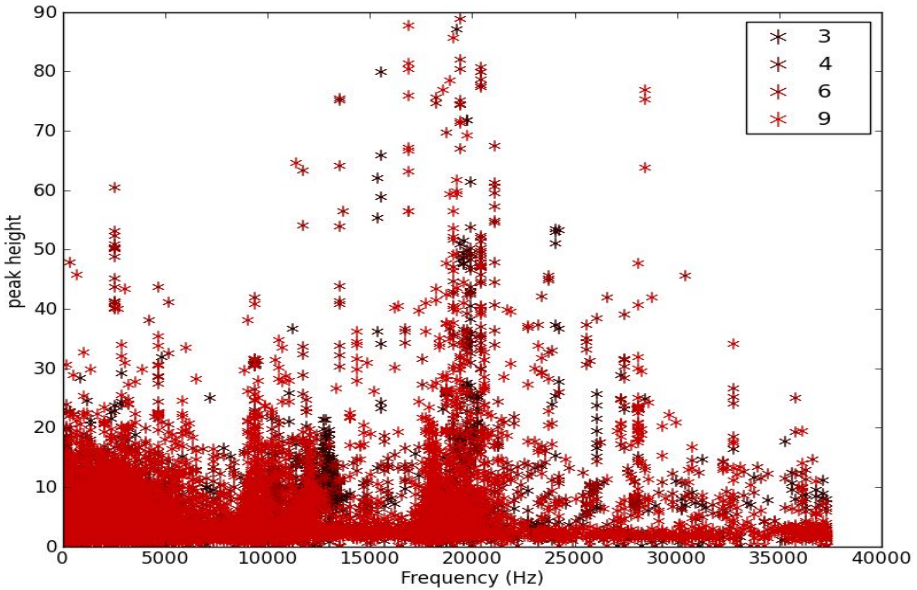
Discussion:

Scatterplot | Use Cases and Drawbacks

Given the use case below for scatter plots:

- Show the association between two, three, or four variables.

What are a few potential drawbacks?



Raise your hand to share your answer!

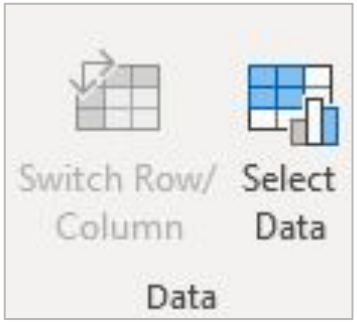
Source: Seaborn Scatterplot

Guided Walk-Through:

Creating Scatterplots in Excel

First, let's create a PivotTable with the following fields:

- **Rows**: "sub_category"
- **Values**: "Sum of profit," "Average of discount"



Note: Excel won't let you insert a scatterplot from a PivotTable. You will need to copy/paste the data into another chart of the sheet as "Paste Values."

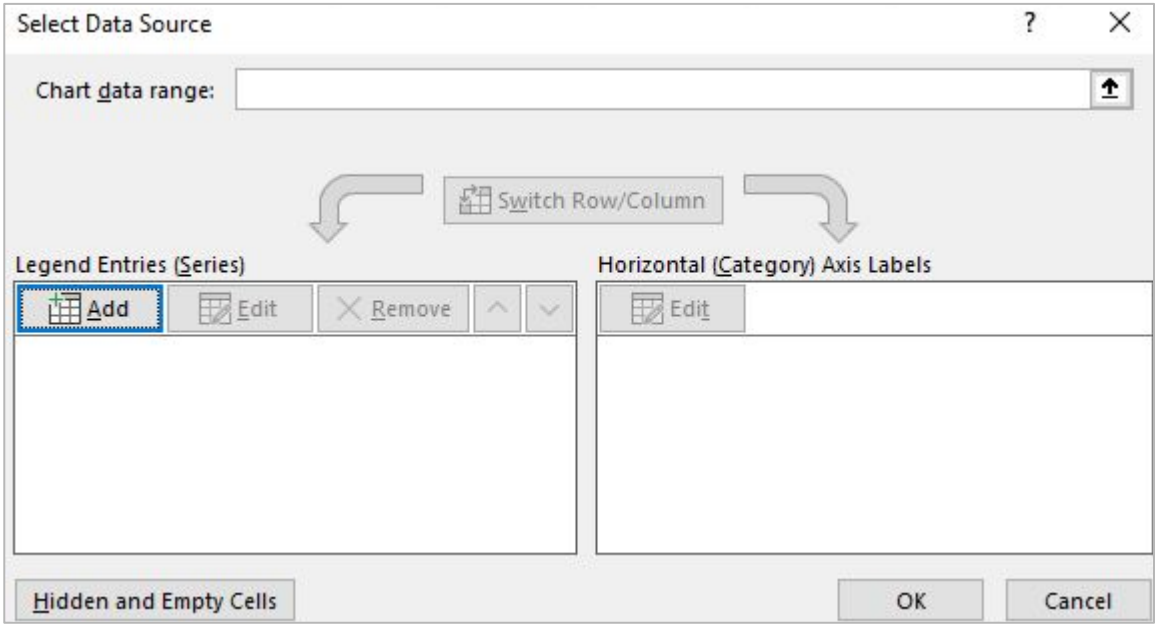
1. Click Insert > Chart > Scatter and select any blank cell to insert a blank chart.
2. On the "Chart Design" ribbon, click "Select Data."

Guided Walk-Through:

Creating Scatterplots in Excel

In order to create each subcategory as its own color, we will need to add each one as a series.

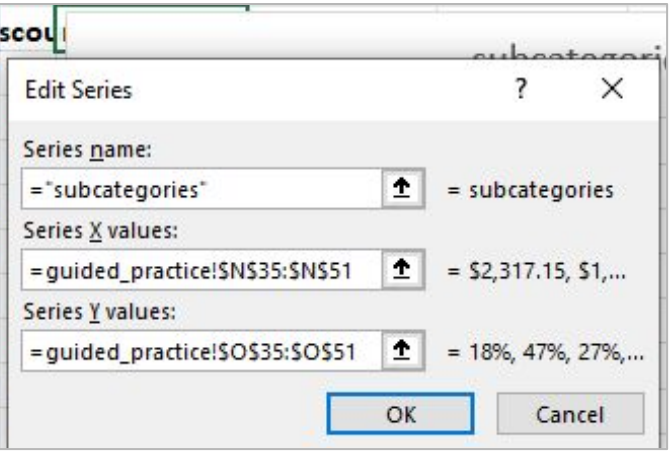
3. Click “Add” to begin selecting your data (“**+**” button on a Mac).



Guided Walk-Through:
Creating Scatterplots in Excel

1. For series name, enter “subcategories”.
2. For the X values, select the range of sums of profit.
3. For the Y values, select the range of averages of discount.
4. Click “OK.”

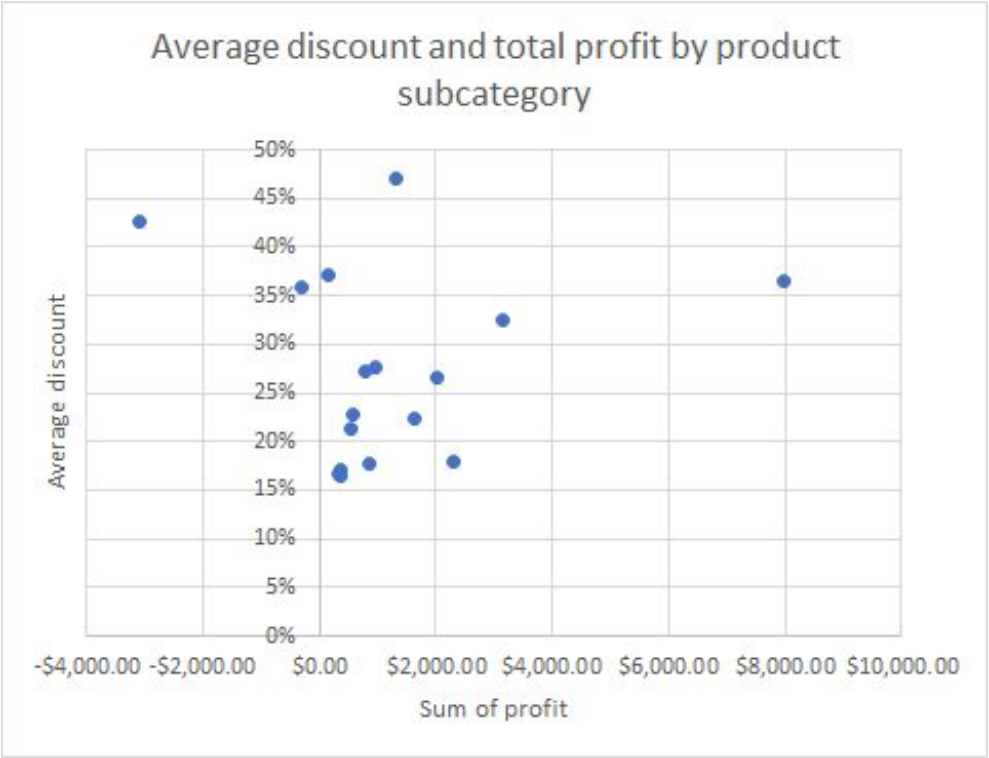
Row Labels	Sum of profit	Average of discou
Accessories	$2,317.15	18%
Appliances	$1,328.05	47%
Art	$784.40	27%
Binders	$950.45	28%
Bookcases	$148.91	37%
Chairs	$3,142.14	32%
Copiers	$7,958.61	36%
Envelopes	$362.69	17%
Fasteners	$362.25	17%



Computers Out:

Creating Scatterplots in Excel

Let's see if we can recreate this scatterplot from our Superstore data!



Computers Out:

Adding Category Data Labels

Let’s start by adding data labels. **Note**: These steps are only available to PC and later versions of Mac. For older versions of Mac, users will need to create a macro.

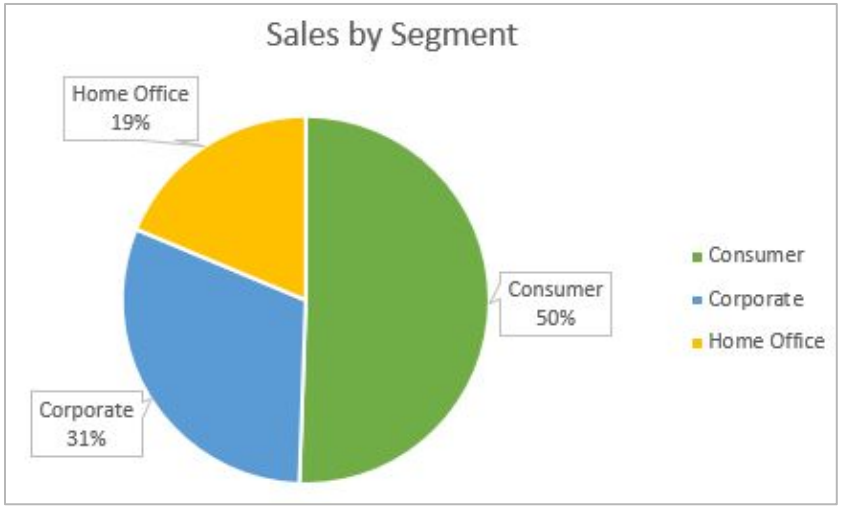
1. Right click on any point on the scatterplot.
2. Choose “Add Data Labels.”
3. Right click directly on any one of the data labels.
4. Choose “Format Data Labels.”
5. Check “Value from Cells” and highlight the names of the subcategories.
6. Click OK.
7. Uncheck “Y Value.”

Exploring Data Compositions: Pie Charts

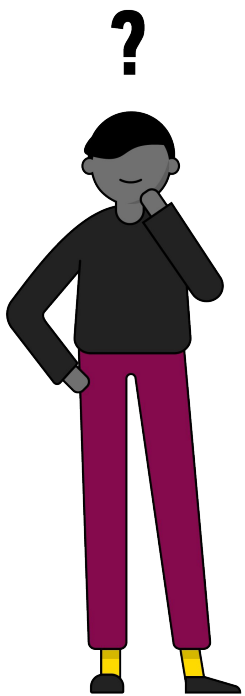
What Are Pie Charts?

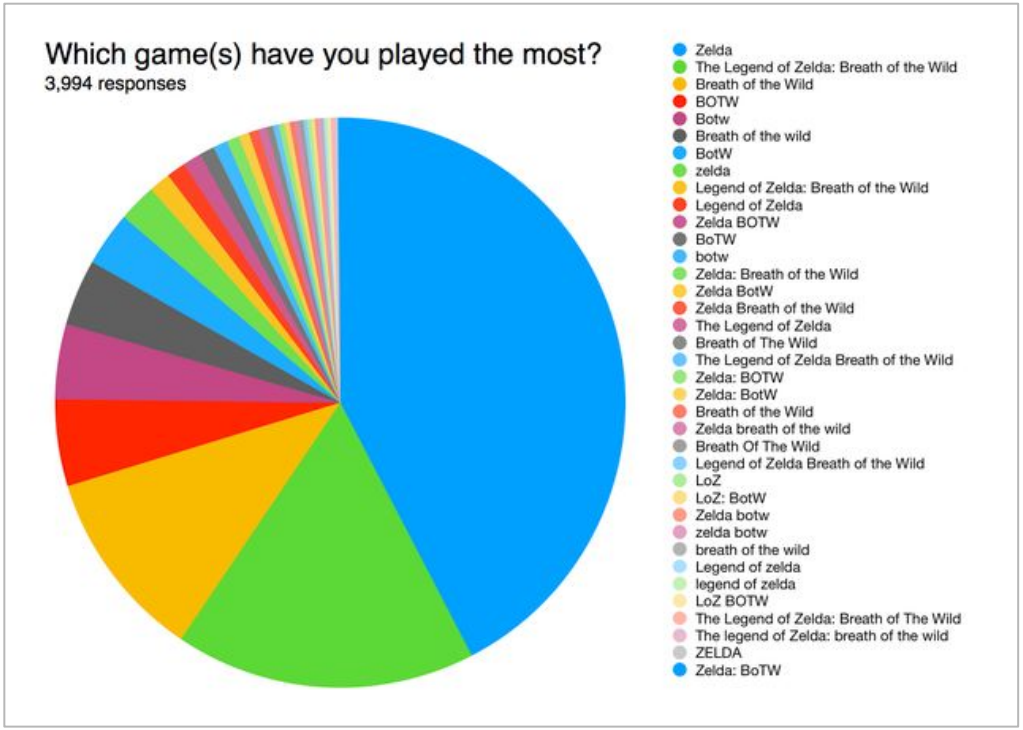
Display information as **a percentage of a whole**.

Pie charts **should never be 3D**, as this can skew the data and confuse the audience!



Please, No





Source: Reddit

When to Use Pie Charts

When you have:

- **Quantitative variables** that add up to 100%.

Best for:

- Displaying how portions of data relate to each other as parts of a whole.

		Quantitative Data Points			
		1	2	3	4+
Categorical Data Points	0	**Histogram** Box plot Heat map Strip plot Bullet chart Triangle plot Gauge charts	Scatterplot Small multiples: Line graph	Standard Bubble Small multiples: Line graphs	Table Small multiples: Line graphs
	1	Bar chart **Pie chart** Line chart Dot plot Treemap Small multiples: - Histograms - Bar charts	Scatterplot Dot plot Small multiples: Histograms	Standard Bubble Dot plot Scatterplot matrix	Table Scatterplot matrix

Adapted from Boxplot Analytics

Guided Walk-Through:

Let’s Make a Pie (Chart)!

Let’s revisit our PivotTable — “**Percentage of orders per customer segment**” — and visualize the result:

- **Row**: Customer segment
- **Values**: Count of “order ID” (show value as % of column total)

1. Select any cell in your PivotTable.
2. Insert > Chart > Pie

Conditional Formatting

This Time You Only Have 15 Seconds...

...to identify the salespersons with the highest and lowest sales revenue in February.

2	*Salesperson*	*Region Covered*	*February 2017 Sales*
3	Jeffrey Burke	Oklahoma	$ 28,000
4	Amy Fernandez	North Carolina	$ 23,138
5	Mark Hayes	Massachusetts	$ 25,092
6	Judith Ray	California	$ 21,839
7	Randy Graham	South Carolina	$ 23,342

2	*Salesperson*	*Region Covered*	*February 2017 Sales*
3	Jeffrey Burke	Oklahoma	$ 28,000
4	Amy Fernandez	North Carolina	$ 23,138
5	Mark Hayes	Massachusetts	$ 25,092
6	Judith Ray	California	$ 21,839
7	Randy Graham	South Carolina	$ 23,342

Meet Conditional Formatting

A tool that allows you to **apply formatting** to a cell or a range of cells and draws attention to specific parts of the data when **a predefined condition** is met.

Before

2	*Salesperson*	*Region Covered*	*February 2017 Sales*
3	Jeffrey Burke	Oklahoma	$ 28,000
4	Amy Fernandez	North Carolina	$ 23,138
5	Mark Hayes	Massachusetts	$ 25,092
6	Judith Ray	California	$ 21,839
7	Randy Graham	South Carolina	$ 23,342

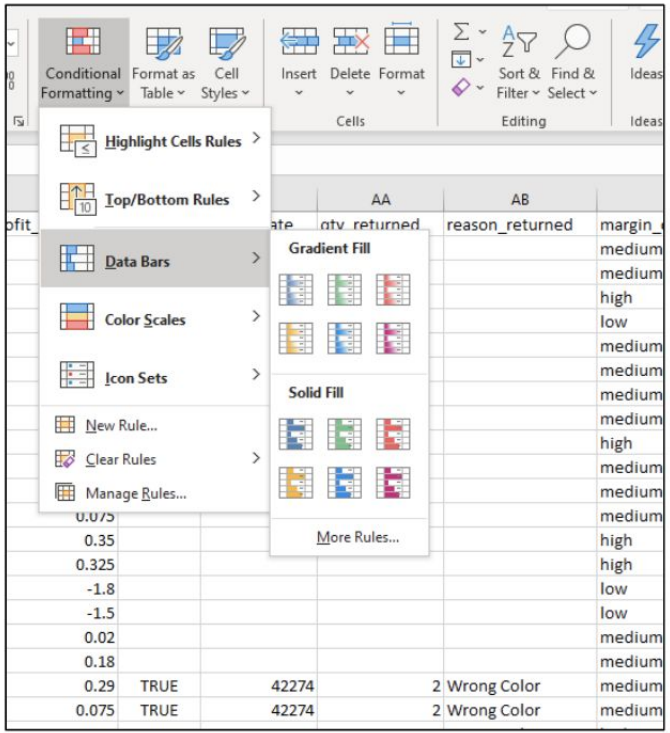
After

2	*Salesperson*	*Region Covered*	*February 2017 Sales*
3	Jeffrey Burke	Oklahoma	$ 28,000
4	Amy Fernandez	North Carolina	$ 23,138
5	Mark Hayes	Massachusetts	$ 25,092
6	Judith Ray	California	$ 21,839
7	Randy Graham	South Carolina	$ 23,342

Applying Conditional Formatting

To set conditional formatting over a data range:

1. Select the cells to which you want to apply conditional formatting.
2. On the “**Home**” tab of the ribbon, select “**Conditional Formatting**”.
3. Choose the type of formatting you want.



Types of Conditional Formatting

Cells Rules

Highlight the data and change the text color according to the value of that cell.

Top/Bottom Rules

Highlight the data and change the text color depending on how the data relates to the rest of the data in the selected range.

Data Bars

Add a color bar to the cell to represent the value in a cell. The higher the value, the longer the bar.

Color Scales

Apply a color gradient to a range of cells. The color indicates where each cell value falls within that range.

Icon Sets

Choose a set of icons to represent the values in the selected cells.

Conditional Formatting in Practice: Heatmap

What do the colors here tell you?

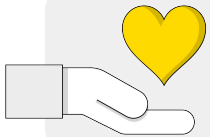
Sum of sales	Column Labels				
Row Labels	2015	2016	2017	2018	Grand Total
Qtr1	$ 74,447.80	$ 68,851.74	$ 93,237.18	$123,144.86	$ 359,681.58
Qtr2	$ 86,538.76	$ 89,124.19	$136,082.30	$133,764.37	$ 445,509.62
Qtr3	$ 143,633.21	$130,259.58	$143,787.36	$196,251.96	$ 613,932.11
Qtr4	$ 179,627.73	$182,297.01	$236,098.75	$280,054.07	$ 878,077.56
Grand Total	**$ 484,247.50**	**$470,532.51**	**$609,205.60**	**$733,215.26**	**$2,297,200.86**

Solo Exercise:

Conditional Formatting

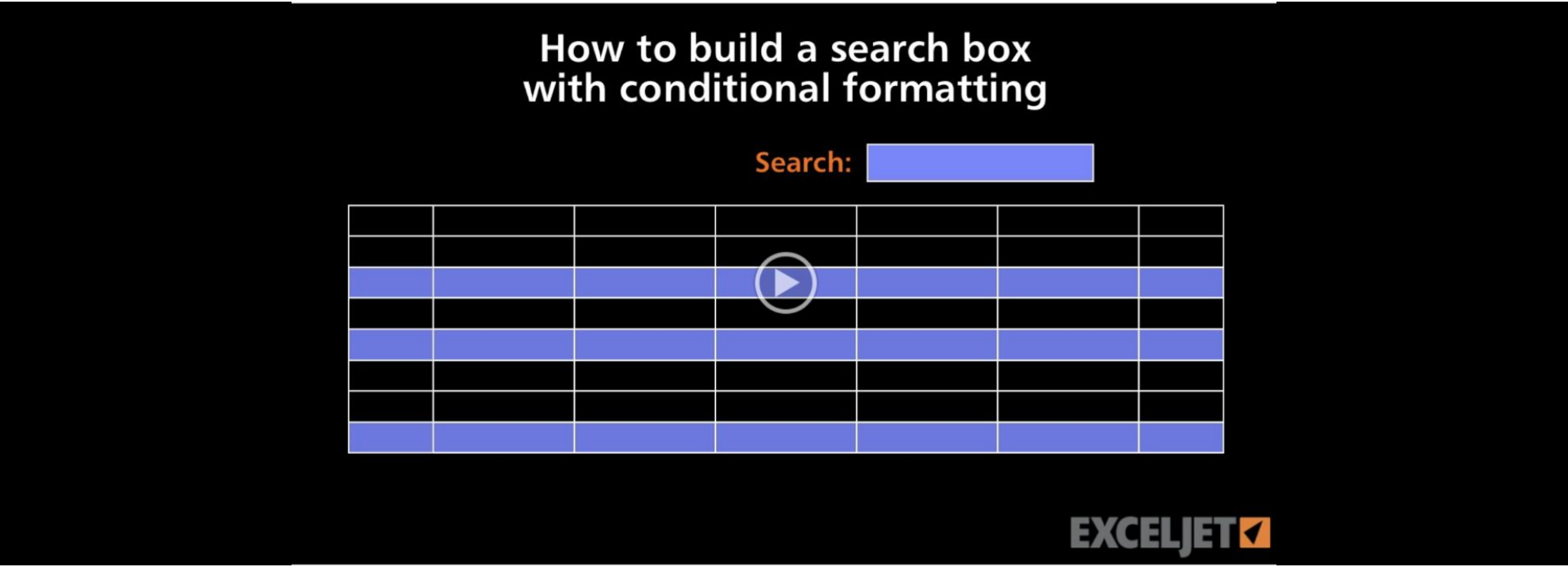
Select any PivotTable we have made up to this point and apply the following conditional formatting types to see what each one looks like!

- ❑ Highlight cell rules
- ❑ Top/bottom rules
- ❑ Data bars
- ❑ Color gradients
- ❑ Icons



Is there one type that works better than the others? If so, be sure to make a note of it for future reference!

Building a Search Box With Conditional Formatting



Solo Exercise:

Optional Homework

Use visualizations to answer the following questions:

1. Which customer segment is making the most returns?
2. Which product subcategory seems to be driving items that are returned in the **consumer** customer segment?
3. We had planned to pursue heavy marketing of technology products to the consumer customer segment next year. Is technology the right product category based on profit margin and return share?

Cleaning and Aggregating Data With Excel

Wrapping Up

What have we learned?

Through our superstore challenge, we have learned how to:

- Summarise our data using **functions**
- Explore our data using aggregate functions
- Join datasets using `LOOKUP` and `XLOOKUP` functions
- Analyse data and generate insights without formulas using **PivotTables**, by quickly create dynamic aggregations, slices, and filters
- Visualize data

Solo Exercise:

Optional Homework

Finish the **`homework (optional)`** 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_customer`** column that identifies customers in the given list.

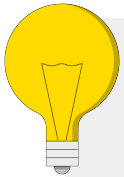
You may need to use formulas we didn't cover in the lesson!

Solo Exercise:

Moving Data With VLOOKUP - Optional Homework

Now let's say we want to bring in more information from the **orders** data set. We don't need everything, just a few columns.

On your own, use **VLOOKUP** to bring the **category, sales,** and **profit** columns from the **orders** worksheet to the **returns** worksheet.



Remember to start from your "**key**" column!

Solo Exercise:

Moving Data With VLOOKUP - Optional Homework - Solutions

Category:

```
=VLOOKUP(A2,orders!A:AJ,7,FALSE)
```

Sales:

```
=VLOOKUP(A2,orders!A:AJ,11,FALSE)
```

Profit:

```
=VLOOKUP(A2,orders!A:AJ,12,FALSE)
```

Solo Exercise:

Optional Homework

Use visualizations to answer the following questions:

1. Which customer segment is making the most returns?
2. Which product subcategory seems to be driving items that are returned in the **consumer** customer segment?
3. We had planned to pursue heavy marketing of technology products to the consumer customer segment next year. Is technology the right product category based on profit margin and return share?

Recap
Today, we...

- Applied data cleaning best practices, including working with **NULLs**.
- Conducted exploratory analyses.
- Manipulated data sets using **LOOKUPs**.
- Summarized data with **aggregate functions** and **PivotTables**.
- Identified the appropriate visualization types for a given data set.
- Created analytics visuals such as bar charts, pie charts, line graphs, histograms, and scatterplots.

Looking Ahead

Optional Homework

- Optional homework

Up Next: Organizing Data with SQL

Additional Resources

- Excel Keyboard Shortcuts
- Relative, Absolute, and Mixed Cell References
- Explaining XLOOKUP
- Using INDEX MATCH
- VLOOKUP vs. INDEX MATCH
- F4 No Longer Changes Absolute Cell References
- Ten PivotTable Problems and Easy Fixes
- Blog Post: Tables and Linking Data Structures in Excel
- Blog Post: Grouping in PivotTables


GA