Introduction to Data Analysis

Practical Skills & Tools
to Pull Insight from Data
from Day 1

Published by Noah Gomez

Introduction

This e-Book is designed to teach practical data analysis based on a small number of fundamental principles and techniques. It's based on a simple idea: **data is nothing more than organized information**.

An object or idea can be broken down into pieces of information, which, when organized, become data. For example, an apple is just an apple, unless we decide to write down that its weight is 1 kilogram. Now we have organized information.

Once information is *organized*, we can think *critically* about it. For example, if we write that two other apples weigh 0.5 kilograms and 1.5 kilograms respectively, the average weight of the apples is 1 kilogram—a critical thought. **in other words, data *analysis* consists of thinking critically about organized information**.

This example is simple, but the concept is key. This book takes a bottom-up approach to learning data analysis based on the above ideas.

We will **define** data and organize it into **tables**, as well as **manipulate** those tables to answer different **questions**.

We will explore how all data analysis is open to **interpretation**.

We will outline key **terms** you'll need to understand how to think about data, and we will reinforce the importance of *good* **questions** as drivers for the organization and interpretation of **information**.

This book provides a **framework** for virtually all **data exercises** you will ever perform. It shows how to use Excel to **visualize** data in 5 essential charts, outlines 23 essential Excel **functions** that will help you answer a majority of questions, discusses how **averages** and **correlations** are capable of answering most questions, and explains how you can go from amateur to master with speed, practice, and advanced tools and techniques.

To check your understanding, there are 6 section quizzes to ensure you takeaway key information (with answers at the end of the book). More importantly, we'll perform a sample analysis using data you'll download in the next section. You will also perform three business cases based on the data, where you can exercise what you've learned in the book.

My hope is that this e-Book makes you feel like data is simple, yet extremely valuable. I hope it will plant the seed for your understanding, so that you can drive valuable insights for any organization using only the small set of tools, methods, and techniques we'll tackle herein.

Remember, data is nothing more than organized information. Data analysis is nothing more than thinking critically about that organized information.

What You'll Need

Before we get started, you'll need to download some data we'll use throughout this book. In addition, you'll need to ensure you have access to Microsoft Office Excel—don't worry if you don't have a license. You can use the free version online using the guide on AnalystAnswers.com through the link below.

Download the Superstore Dataset

Get Microsoft Excel Online Free

Table of Contents

Purpose: Obtaining Skills for Introductory-Level Data Analyst Jobs

The purpose of this e-Book is to provide you with the knowledge and skills you need for an introductory-level data analyst job. What you'll learn here represents the fundamental knowledge and skills for almost any data analysis role, and once you've learned it, the sky is the limit.

That said, **data analysis is a skill like any other**. To master it, you need to practice, practice, practice.

But you must ensure it's GOOD practice! Anyone can open Excel and start crunching numbers, but good data analysts understand the fundamentals well so they can gather, structure, process, and analyze data that drives real insights.

They know how to use a small set of functions and tools to answer most questions, and they know how to format their spreadsheets so they're easy to understand and review. These are all things we'll cover in this book.

With GOOD practice and time, you can progress quickly. To become a master, you will need **advanced tools**, **advanced techniques**, and perhaps most importantly, you'll need to be **fast**. However, the basics never change. So read attentively and have your Excel table ready!

First things first, we need to understand what data is.

What is data?

Data is organized information. Information is all around us, all the time. But we can't devote energy to analyze and structure at every moment. In his book *Thinking Fast and Slow*, renowned economist Daniel Kahneman discusses this idea, explaining that the human mind consists of two systems – a fast one and a slow one.

The fast one "operates automatically and quickly, with little or no effort and no sense of voluntary control," whereas the second "allocates attention to the effortful mental activities that demand it."[1]

Throughout the day, the fast system is dominant. We unconsciously observe our surroundings, discuss with others, make haphazard choices and form fleeting opinions. On the other hand, when we work, pay bills, or make important decisions, we engage the slow system, which ***thinks critically***.

Data is organized information that we structure using **system two**. And the way we structure it is with **columns and rows**.

Columns & Rows

The basic structure of all data is columns and rows. Each row represents one unit of observation, and each column represents a qualitative or quantitative trait about that unit of observation. Using the example of Apples from earlier, Apple 1, Apple 2, and Apple 3 are all units of observation, and weight is a quantitative trait.

For a more tangible example, let's look at cells N1 to R4 in the super store dataset (if you didn't download it yet, you can do so here).

Product ID	Category	Sub-Category	Product Name	Sales
FUR-BO-10001798	Furniture	Bookcases	Bush Somerset Collection Bookcase	261.96
FUR-CH-10000454	Furniture	Chairs	Hon Deluxe Fabric Upholstered Stacking Chairs, Rounded Back	731.94
OFF-LA-10000240	Office Supplies	Labels	Self-Adhesive Address Labels for Typewriters by Universal	14.62

The first row, also called a **header row**, consists of titles that explain the data in the columns below them—simple enough. The first column holds **unique values** that identify the unit of observation, which in this case is a product.

The next two columns show category and sub-category information related to each product, while the next column shows the product name and the last column shows the value of the product sold.

Globally, what can we say these columns and rows represent? They show product information in the superstore. Before this information was organized in a table, it was nothing more than information about different products (like the Apples from the introduction). Now that it's structured, it's **data**.

When the structure of rows and columns consists of (1) a header row, (2) at least one observation, and (3) at least one trait, the structure is called a **data table**. Tables are the fundamental structure of all data. But be careful, because *not all columns and rows are data tables* – they need unique IDs.

[1] Thinking Fast and Slow, Daniel Kahneman, pg. 20-21

Unique ID (aka Primary Key)

For rows and columns to be valuable, they must be structured as a data table containing **unique IDs**. Unique IDs are non-repeating names for each unit of observation recorded on each row. They tell you what the columns describe. If a table does not have a unique ID, **then it is not a data table**!

For example, in the table from the previous section, the unique ID is the column titled Product ID. We know that every other column in the table describes the product identified in that column. Imagine we remove it, such as in the table below.

Category	Sub-Category	Product Name	Sales
Furniture	Bookcases	Bush Somerset Collection Bookcase	261.96
Furniture	Chairs	Hon Deluxe Fabric Upholstered Stacking Chairs, Rounded Back	731.94
Office Supplies	Labels	Self-Adhesive Address Labels for Typewriters by Universal	14.62

If I ask you to tell me what the category, sub-category, product name, and sales columns describe, could you?

You might be able to intuitively infer based on the names that this tables addresses products, but you cannot tell me more based on the information available because the categories repeat (such as Furniture), and the sub-categories are ostensibly additional information about the categories.

Without a unique ID, also known as a **primary key**, the table is not very useful because we cannot draw conclusions from it—we cannot say what the rows and columns describe.

So remember, **all data tables MUST have a unique ID**. If not, we need to manipulate it to produce one. Manipulating to maintain primary keys is a key skill, so let's look at it in the next section.

Original vs Aggregated Data

Let's bring down our snippet from above.

Product ID	Category	Sub-Category	Product Name	Sales
FUR-BO-10001798	Furniture	Bookcases	Bush Somerset Collection Bookcase	261.96
FUR-CH-10000454	Furniture	Chairs	Hon Deluxe Fabric Upholstered Stacking Chairs, Rounded Back	731.94
OFF-LA-10000240	Office Supplies	Labels	Self-Adhesive Address Labels for Typewriters by Universal	14.62

This data table shows us **raw data** – that is, the most detailed, or ***granular***, form of the information available. It is the **original** data collected.

Imagine now that we want to see product information by Category. We need to **manipulate** the data so that **the Category become the primary key for the table**. A quick look at the Category column shows that there are only two entries available: Furniture and Office Supplies.

However, there are three different products associated with these categories, and the sub-category, product name, and sales columns also have three different data entries. How can we condense the 3-item columns into 2 rows needed to create the Category Unique ID?

The output of the manipulation would look like this:

Category	Count of Product ID	Count of Sub-Category	Count of Product Name	Sum of Sales
Furniture	2	2	2	993.9
Office Supplies	1	1	1	14.62

As you can see, to maintain a non-repeating unique ID based on Category, we must combine the columns with more than 1 entry per category. For **qualitative data** (Product ID, Sub-Category, Product Name), we've **counted** the number of entries. For **quantitative date** (Sales), we've **summed** the amounts.

This process is called **aggregation** – combining information from multiple rows using **aggregation functions**. Note that aggregation functions always convert **qualitative** data into quantitative data by **counting** them, while quantitative data can be manipulated with other functions such as sum and average.

In most cases, data is not presented in its raw form. Instead, it is pre-processed, aggregated, and presented in a "digestible" way.

A common cause of confusion for data analysts working across departments stems form the difference between raw and aggregated data. Non-data professionals often request support on aggregated datasets, which is less useful for data analysts, who prefer to start with granular data to draw conclusions.

So aggregated data is combined information using the aggregation functions. But what are they?

The 6 Aggregation Functions

There are 6 aggregation functions: SUM, COUNT, COUNT DISTINCT, MINIMUM, MAXIMUM, and AVERAGE. The table below shows how they apply to qualitative and quantitative data.

Aggregation Function	Impact on Qualitative Data	Impact on Quantitative Data
Sum	Cannot be applied	Sums the values
Count	Counts the number of entries	Counts the number of entries
Count Distinct (CountD)	Counts the number of entries, ignoring duplicates (i.e. "Tub, Tub, Tub, Shower"= 2 with CountD)	Counts the number of entries, ignoring duplicates (i.e. "1, 1, 2, 3" = 3 with CountD)
Maximum	Cannot be applied	Returns the highest value
Minimum	Cannot be applied	Returns the lowest value
Average	Cannot be applied	Returns the arithmetic mean of values

Remember, **qualitative data is always counted**, while quantitative data can be treated with any of the 6 aggregation functions.

Data Sets

We mentioned before that data tables are columns and rows with Unique IDs, it's important to make the distinction between tables and ***sets***. As early as elementary school, we've all been exposed to data tables. In many cases, we refer to them as "sets."

Why? Because it's easy and intuitive. When you have two tables, it seems more reasonable to call them "set 1" and "set 2." But this is not entirely correct.

"Data set" is a term used to refer to many different types of data objects, of which a data table is only one. Others include schemas, points, arrays, records, and files. In fact, the word "set" is an umbrella term for use-cases outside the scope of data analysis, such as in programming languages.

It's important to make this distinction now because as you advance in data analysis and learn more about structures other than tables, vocabulary becomes critically important (see the key terms section).

Strictly speaking, a **data set is a collection of one or more tables, schemas, points, and/or objects that are grouped together either because they're stored in the same location or because they're related to the same subject**.

Beyond the structure itself, **location** and **topicality** are defining criteria for data sets. In most cases, data sets refer to related information – that is, rows and columns that describe unique IDs on the same topic.

However, this is not always the case. Database managers face many challenges when it comes to storing data. In big organizations, minimizing costs and therefore reducing space is a business requirement.

As a result, database managers often store unrelated data in the same place to optimize storage. This data is considered a "set" because of its location, even though it is not related to the same topic.

"Data analysis boils down to (1) manipulating unique IDs, (2) applying the 6 aggregation functions, (3) taking averages and related statistical functions, and (4) applying correlations. That's it. Once you understand and can apply these skills to answer questions, you're doing the job of a data analyst."

Section Quiz 1

1. What is data?
2. What is the basic structure of all data?
3. What is a data table?
4. What special trait distinguishes a data table from other collections of columns and rows?
5. What is an aggregation?
6. What is raw data?
7. What are the six aggregation functions?
8. What is the difference between a data set and a data table?

Answers here.

What is data *analysis*?

As we discussed above, data is organized information, and its most common form is a data table. **Data analysis**, however, consists of thinking critically about organized information.

But what does it mean to think critically? **In short, critical thinking involves asking questions and answering them by manipulating data, in most cases with the 6 aggregation functions.**

The starting point of any analysis is a targeted, unambiguous, and unbiased question.

In many cases, the question may come naturally as you look at the data. Take a minute to look at the superstore data again—don't read on just yet. Take 60 seconds to look at the data columns and see what questions comes to mind.

Here is a list of questions that you might ask:

1. Which states have the highest sales?
2. Which product sells most often?
3. Which product generates the most sales?
4. Which product generates the most profit?
5. What impact does ship mode have on sales?

These questions may seem simple, but it's very easy to ask **bad** questions. Here are examples of *bad* questions and explanations for why:

1. Why has furniture become such a good category?
2. How great are sales in the state of California?
3. How have 2-quantity orders increased our sales?
4. How poorly did the corporate segment do again in 2017?
5. Shouldn't we focus on the South region?

Explanations:

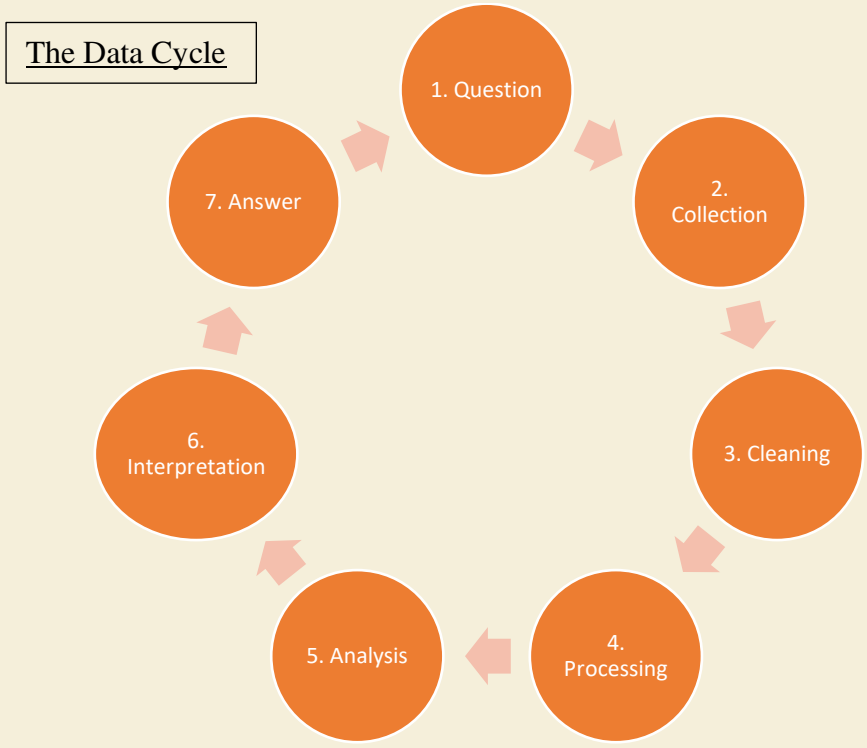
1. This question assumes furniture has become a good category, which would lead the analyst to prove that point even if furniture is a poor performing category. Moreover, the word "good" is ambiguous. Do we mean high sales compared to other categories? Do we mean high profit compared to previous performance? Do we mean compared to its past performance of other categories? It's unclear.
2. This question triggers a response before it is answered. It *is leading* because it insists upon a positive result in California. It is not an objective inquiry.
3. This question assumes 2-quantity orders have increased sales, which leads the analyst to prove this point. However, 2-quantity orders may have decreased sales because consumers spend less per item.
4. This question leads the response by indicating an opinion, "poorly." Moreover, it does not demand an insightful explanation—even if the corporate segment performed poorly, there should be a question as to *why*. A better version would be "how did the corporate segment's performance compare to last year? What happened?"
5. This question insists on a decision before the analysis is complete, rather than asking a neutral question. A better version could be, "which region holds the most potential for future sales?"

The right question sets you up for success. Bad questions result in inconclusive analyses.

So we know that data analysis starts with good questions, but what does it look like in practice? What steps does an analyst need to take? To answer this, let's look at the **data cycle**.

The Data Cycle

Data analysis answers questions, but the analysis itself is part of a larger cycle, which consists of question → collection → cleaning → processing → **analysis** → interpretation → answer. We won't investigate all of these steps because they are outside the scope of this introduction, but it's important to see where analysis fits in to the big picture.

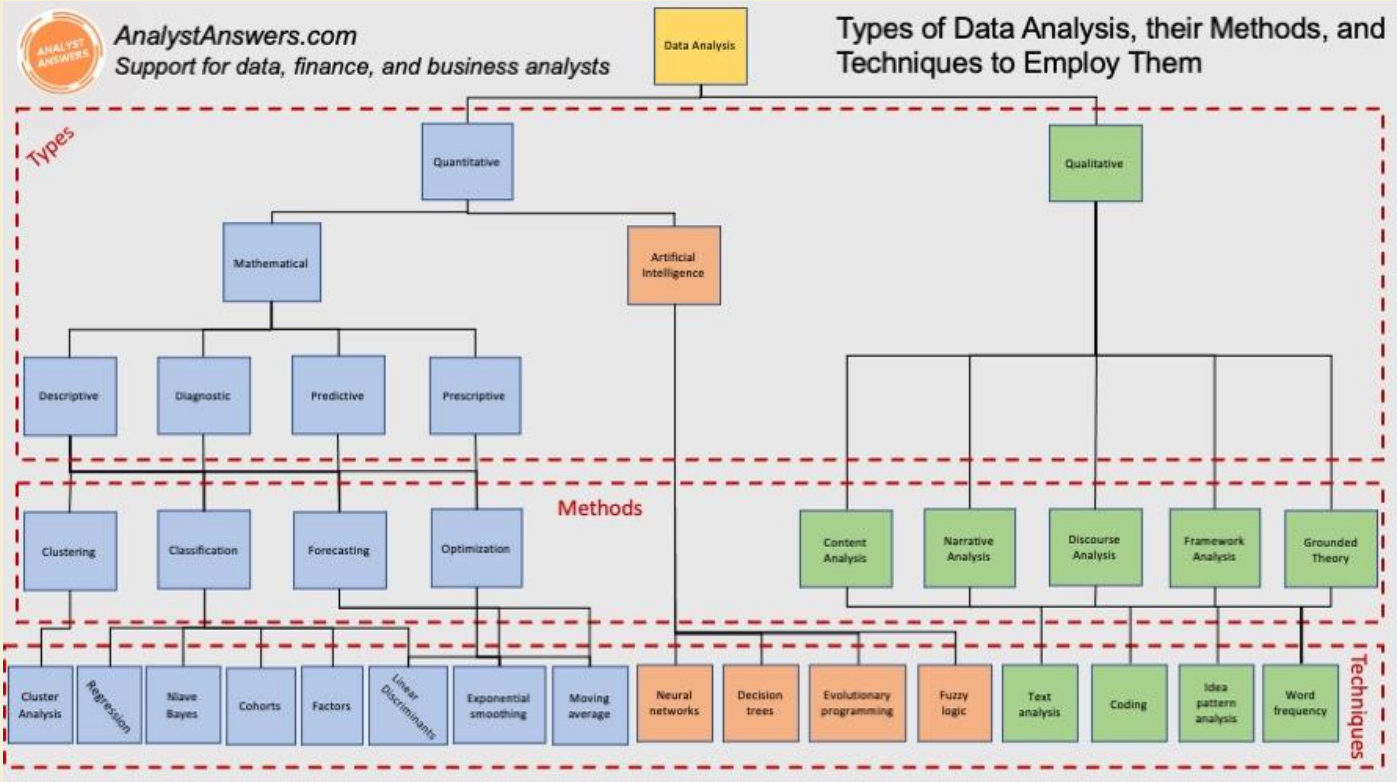


As you can see, analysis is the fifth step of seven. Before we can analyze it, data must be collected, cleaned, and processed. After the analysis, the data must be interpreted to provide the answer to the original question.

In this way, data analysis is concerned with manipulations of the data using a handful of techniques. What are these techniques? To answer this, let's look at *types* of data analysis.

Types of Data Analysis

There are over **30 different types, methods, and techniques** (see the image below), and each of these can be personalized and customized to the liking of each analyst.



As you can see, qualitative and quantitative analysis is the first division, followed by mathematical and AI branches. Mathematical types are further broken down into descriptive, diagnostic, predictive, and prescriptive types, which are buzz words in the data world.

Descriptive analysis aims to describe existing data, whereas predictive analysis aims to foresee what the data will look like in the future. Diagnostic analysis aims to take the results of descriptive analysis and identify the root cause of information under question (good questions are essential!). Prescriptive analysis aims to propose concrete plans to treat what's been diagnosed.

Don't feel overwhelmed by the diagram. Most of the techniques you see above are reserved for experienced analysts, and few if any master all of them. Moreover, these 30 items are extensions of a few basic ideas that we'll cover in the next sections.

Specifically, data analysis boils down to (1) manipulating unique IDs, (2) applying the 6 aggregation functions, (3) taking averages and related statistical functions, and (4) applying correlations. That's it. Once you understand and can apply these skills to answer questions, you're doing the job of a data analyst.

Now that we understand what data analysis is conceptually, it's time to get practical and learn key functions in Excel. Then we'll examine 5 essential charts that you can use in virtually any analysis. From there, we'll move into 5 basic statistics that will help you answer almost any question.

Make sure you have the Superstore data table open. We'll need it for the rest of the book!

Section Quiz 2

1. What is data analysis?
2. What is the starting point of any analysis? And what are its three key components?
3. What is the data cycle? How does analysis fit into it?
4. What are data types, methods, and techniques?
5. What four techniques constitute the job of a data analyst?

Answers here.

Excel is Your Best Friend… So Let's Get to Know Him

I've talked a lot about how data analysis consists of a small number of fundamental skills that you refine as you progress. The same applies to the tools you use. Excel is the basic data analyst software. As you grow, you'll evolve to use other tools, but Excel is the perfect place to start.

That said, you need good Excel fundamentals. In this section we'll cover 23 Excel functions and the importance of pivot tables.

23 Essential Excel Functions

The most important Excel functions you need to know are in the following table.

Function	Description
Aggregation Functions	
=SUM(number1,[number2],...)	Adds the cells or arrays referenced
=COUNT(value1, [value2], ...)	Counts the number of non-empty numerical cells (=COUNTA() can be used for non-numerical cells)
=AVERAGE(number1, [number2], ...)	Takes the arithmetic mean of the values in cells
=MIN(number1, [number2], ...)	Returns the smallest number from a selection of multiple numbers
=MAX(number1, [number2], ...)	Returns the largest number from a selection of multiple numbers
=SUBTOTAL(function_num,ref1,[ref2],...)	Returns the result of a function (i.e. sum, average) from *visible* cells only
IF Statements	
=IF(logical_test, value_if_true, [value_if_false])	Returns a selected entry when the logic test is true, and another entry when the logic test is false
=IFERROR(value, value_if_error)	Returns a specified entry when the first value or expression results in an error
Aggregation Functions + IF Statements	
=SUMIF(range, criteria, [sum_range])	Adds values based on criteria
=COUNTIF(range, criteria)	Counts values based on criteria
=AVERAGEIF(range, criteria, [average_range])	Averages values based on criteria
=MINIFS(min_range, criteria_range1, criteria1, [criteria_range2, criteria2], ...)	Returns the smallest number based on criteria
=MAXIFS(max_range, criteria_range1, criteria1, [criteria_range2, criteria2], ...)	Returns the largest number based on criteria
Concatenations	
=CONCATENATE(text1, [text2], ...)	Combines text from selected cells into one cell
Text Functions	
=LEFT/RIGHT(text,[num_chars])	Return a specified number of characters from the left or right side of a cell
=LEN(text)	Returns the number of characters in a cell
=MID(text, start_num, num_chars)	Returns a specified number of characters from selected a starting point in a cell
=SEARCH(find_text, within_text, [start_num])	Finds subtext within a primary text
Conditional Location	
=INDEX(array, row_num, [column_num])	Returns the cell in an array based on row and column number
=MATCH(lookup_value, lookup_array, [match_type])	Returns the position of a cell within a one-dimensional array
=INDEX+MATCH	COMBO: uses the result of a match function as the row or column number in the index function
=XLOOKUP(lookup_value, lookup_array, return_array, [if not found], [match_mode], [search_mode])	Returns the value in an array based on its position relative to a select value in a first array

NOTE: there is no function for **Count Distinct** in Excel, even though it is a core aggregation function. The function to calculate distinct number of values in a cell is =SUM(1/COUNTIF(array,array)).

The nested (i.e. function inside another function) COUNTIF function using arrays will return an array with the COUNT (repeating) of each cell. By dividing 1 by this array, we calculate the fraction that each instance represents of the whole. By SUMing these numbers, we get the total number.

For example, look at the first two entries in column G, Customer Name. We can see that there are two instances of Claire Gute.

F	G	H
Customer ID	Customer N.	Segment
CG-12520	Claire Gute	Consumer
CG-12520	Claire Gute	Consumer
DV-13045	Darrin Van H	Corporate
SO-20335	Sean O'Doni	Consumer

When we use =COUNTIF() to return the number of times this name appears, it will obviously return 2. However, when we divide 1 by 2, we'll get 0.5. Now we simply place the same formula in each row, and 0.5 will appear twice. When add them together, it will be 1 – giving us a distinct count (non-repeating) of the Claire Gute name.

G	H	I	J	V	W	X	Y
Customer N.	Segment	Country	City				
Claire Gute	Consumer	United State	Henderson		=1/COUNTIF($G$2:$G$3,G2)		
Claire Gute	Consumer	United State	Henderson		0.5		
Darrin Van H	Corporate	United State	Los Angeles				
Sean O'Doni	Consumer	United State	Fort Lauderdale		Sum		
Sean O'Doni	Consumer	United State	Fort Lauderdale		1		
Brosina Hoff	Consumer	United State	Los Angeles				

This formula will make more sense one you read the =SUM() and =COUNTIF() sections below.

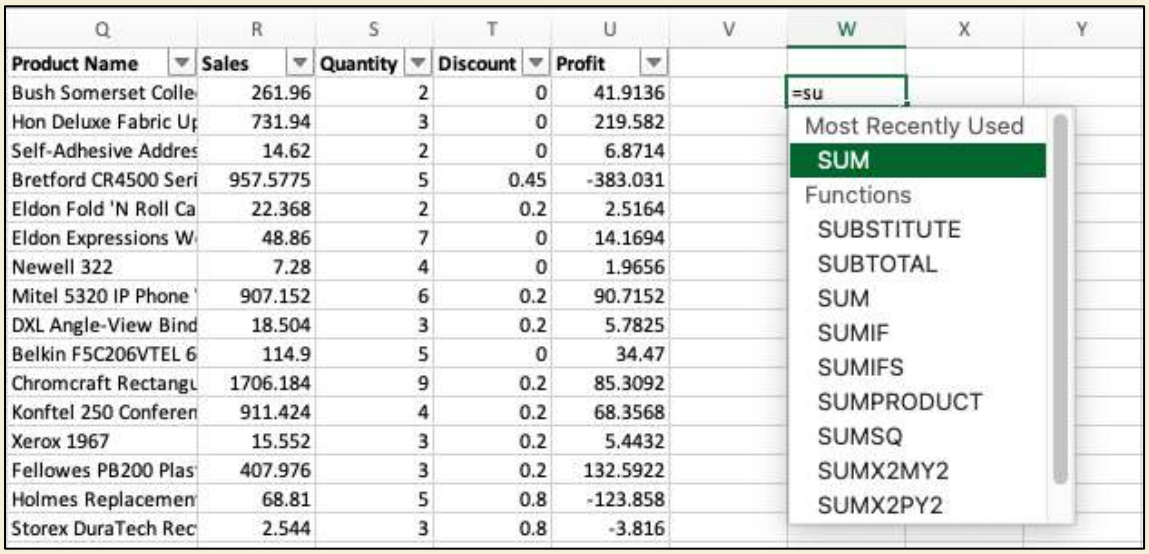
=SUM(number1,[number2],...)

In the superstore spreadsheet, we're using cells Q1:U6.

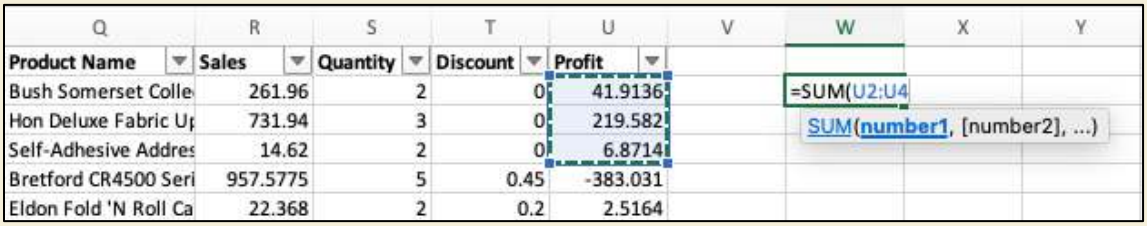
Q	R	S	T	U
Product Name	Sales	Quantity	Discount	Profit
Bush Somerset Colle	261.96	2	0	41.9136
Hon Deluxe Fabric Up	731.94	3	0	219.582
Self-Adhesive Addres	14.62	2	0	6.8714
Bretford CR4500 Seri	957.5775	5	0.45	-383.031
Eldon Fold 'N Roll Ca	22.368	2	0.2	2.5164

Imagine you want to take the total profit from the first three products. The =SUM() formula allows you to do this easily.

a. Place your cursor in an empty cell, such as cell W2.
b. Press the "=" sign to signal that you're entering a formula to Excel
c. Start typing SU… until you see the autofill for the =SUM() formula appears



d. Click ”tab” to engage the formula
e. Use the arrow keys to navigate your cursor to cell U2 (you will remain in the “number1” input section within the function)



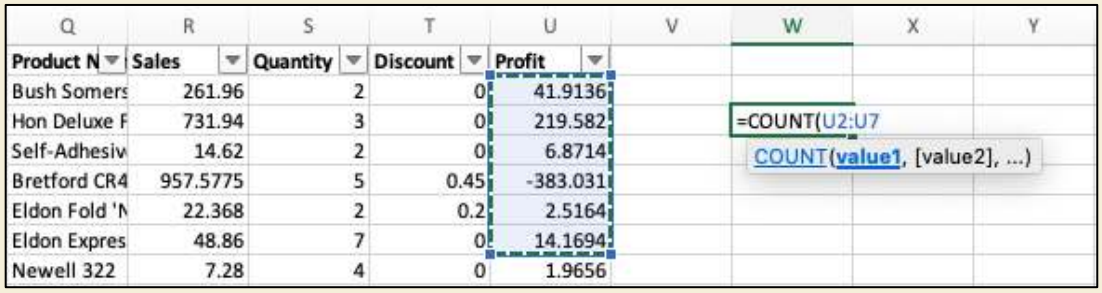
f. While pressing and holding SHIFT, move your cursor down two cells to U4 with the arrow keys
g. Press enter – the result is the sum of those three profits

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

In the superstore spreadsheet, we're using cells Q1:U7.

Imagine you want to count the number of sales in a list. The =COUNT() formula allows you to do this easily.

a. Place your cursor in an empty cell, such as cell W2.
b. Press the "=" sign to signal that you're entering a formula to Excel
c. Start typing CO… until you see the autofill for the =COUNT() formula appears
d. Click tab to engage the formula
e. Use the arrow keys to navigate your cursor to cell U2 (you will remain in the "number1" input mode)
f. While pressing and holding SHIFT, move your cursor down five cells to U7 with the arrow keys



Q	R	S	T	U	V	W	X	Y
Product N	Sales	Quantity	Discount	Profit				
Bush Somers	261.96	2	0	41.9136				
Hon Deluxe F	731.94	3	0	219.582		=COUNT(U2:U7		
Self-Adhesiv	14.62	2	0	6.8714		COUNT(value1, [value2], ...)		
Bretford CR4	957.5775	5	0.45	-383.031				
Eldon Fold 'N	22.368	2	0.2	2.5164				
Eldon Expres	48.86	7	0	14.1694				
Newell 322	7.28	4	0	1.9656				

g. Press enter – the result is a count of non-empty cells, which in this case is 6

Be careful to **only** use =COUNT() with cells that have numbers. If you use it on cells with *text*, the formula will return zero. For text, use another formula, =COUNTA().

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

The average formula works just like =SUM() and =COUNT(), except it returns the arithmetic mean. Try it on cells U2:U6. The result should be -22.42952.

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

Let's use the same cells again from the superstore data table. Imagine now that you want to find the smallest number in an array of numbers. The best way to do this is with =MIN().

1. Place your cursor in an empty cell, such as cell W3.
2. Press the "=" sign to signal that you're entering a formula to Excel
3. Start typing MI… until you see the autofill for the =MIN() formula appear
4. Click tab to engage the formula
5. Use the arrow keys to navigate your cursor to cell U2 (you will remain in the "number1" input section within the function)

Q	R	S	T	U	V	W	X
Product N	Sales	Quantity	Discount	Profit			
Bush Somers	261.96	2	0	41.9136			
Hon Deluxe F	731.94	3	0	219.582		=MIN(U2:U6)	
Self-Adhesiv	14.62	2	0	6.8714			
Bretford CR4	957.5775	5	0.45	-383.031			
Eldon Fold 'N	22.368	2	0.2	2.5164			
Eldon Expres	48.86	7	0	14.1694			

6. While pressing and holding SHIFT, move your cursor down five cells to U6 with the arrow keys
7. Press enter – the result is the smallest number in the array, which is in this case -383.031.

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

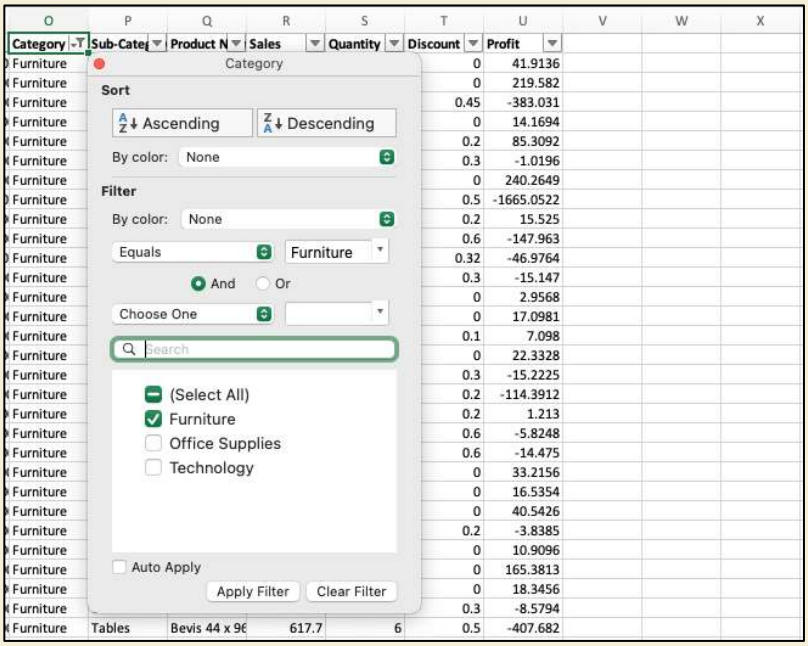
The =MAX() formula works just like =MIN(), except it returns the **largest** number in a range. Try it on cells U2:U6. The result should be 219.582.

=SUBTOTAL(function_num,ref1,[ref2],...)

In the Superstore data table, we're again using columns O to U.

Imagine you're working with a data table that you have already **filtered** (you engages the filter function by navigating to Data > Filter and saw the drop-down arrows appear at the head of each column), and you would like to perform one of the aggregation functions above to the visible cells only.

For example, imagine we filter our data table Category field for Furniture such as in the image below. By filtering, we tell Excel to show us only the rows with the select filter criteria.



If you use =SUM() to calculate the total profit in the column, the function will ignore the filter and take the total column, which you can see if you =SUM(US:U9979) before and after filtering column O as in the picture below – the value will be the same.

However, by using SUBTOTAL, you can choose sum in the first argument in the syntax (number 9), and the result will show the sum of visible cells only. You can see these results in cells W2:W26 below on the filtered table. Go ahead and try it for yourself using the =SUBTOTAL() formula shown in the image below.

	O	P	Q	R	S	T	U	V	W
1	Category	Sub-Category	Product Nam	Sales	Quantity	Discount	Profit		
2	Furniture	Bookcases	Bush Somerset C	261.96	2	0	41.9136		SUM
3	Furniture	Chairs	Hon Deluxe Fabri	731.94	3	0	219.582		285,078.20
5	Furniture	Tables	Bretford CR4500	957.5775	5	0.45	-383.031		=SUM(U2:U9979)
7	Furniture	Furnishings	Eldon Expression	48.86	7	0	14.1694		
12	Furniture	Tables	Chromcraft Recta	1706.184	9	0.2	85.3092		SUBTOTAL
25	Furniture	Chairs	Global Deluxe Sta	71.372	2	0.3	-1.0196		18,226.24
26	Furniture	Tables	Bretford CR4500	1044.63	3	0	240.2649		=SUBTOTAL(9,U2:U9979)
29	Furniture	Bookcases	Riverside Palais R	3083.43	7	0.5	-1665.0522		
31	Furniture	Furnishings	Howard Miller 13	124.2	3	0.2	15.525		

=IF(logical_test, value_if_true, [value_if_false])

Let's imagine you want to determine the products whose profit is more than 40% of its sale value (also known as profit margin). We can create a formula to easily tell us which products have this trait using an =IF() statement.

a. Calculate the profit margin by dividing cells in column U by adjacent cells in column R (see cell W5 in image below).
b. Create a logical argument to test whether the profit margin is greater than 40% (see cell X5 below). This will return a TRUE/FALSE.
c. Us the TRUE/FALSE in the =IF() statement and assign "Yes" as the value_if_true argument, and "No" as the value_if_false argument. Note that because these are text arguments, we have to put them in quotation marks.
d. Alternatively, place all of the arguments directly in the =IF() statement as shown below in cell AA4.

	R	S	T	U	V	W	X	Y	Z	AA
	Sales	Quantity	Discount	Profit		Profit margin (%)	Logical Test	IF Statement		IF Statement
	261.96	2	0	41.9136		16%	FALSE	No		No
	731.94	3	0	219.582		30%	FALSE	No		No
	14.62	2	0	6.8714		47%	TRUE	Yes		=IF((U4/R4)>0.4,"Yes","No")
4	957.5775	5	0.45	-383.031		=U5/R5	=W5>0.4	=IF(X4,"Yes","No")		

=IFERROR(value, value_if_error)

Imagine you want to perform a calculation that you know will return an error message. For example, you want to see how much profit you made compared to the discount provided on the product ordered (also known as the Profit to Discount ratio). However, when the discount is 0, the divisor in the fraction will be a 0 and Excel will return the error message "#DIV/0!". You cannot divide by zero!

To avoid showing the error message, we can use =IFERROR(). In the superstore data table, enter =U2/T2 in cell W2. You will see the error message. In cell X2, enter =IFERROR(U2/T2,"No Discount") as shown in the image below. Press enter, and you will see that we have our text "No Discount" instead of the error message. Remember to put the text in quotation marks!

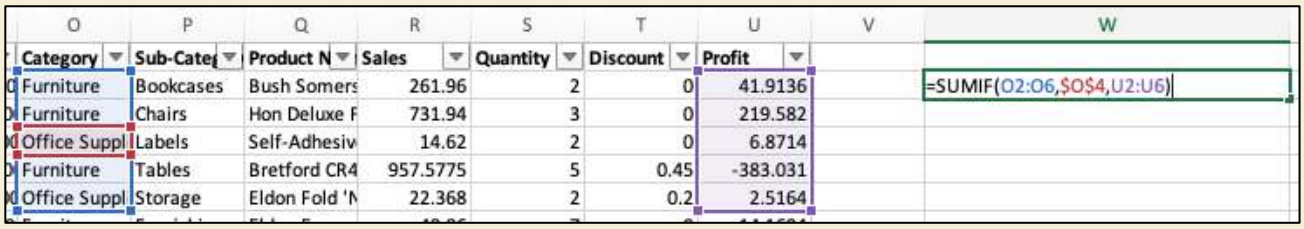
=SUMIF(range, criteria, [sum_range])

Let's continue to use cells O1:U6. Imagine you want to sum the profits from orders whose category is "Office Supplies." In other words, you want a conditional sum.

a. Place your cursor in an empty cell, such as cell W2.
b. Press the "=" sign to signal that you're entering a formula to Excel
c. Start typing SUMI… until you see the autofill for the =SUMIF() formula appear
d. Click tab to engage the formula
e. The "range" argument should include the cells on which we want to apply the sum criteria, in this case O2:O6. Use the arrow keys and press and hold SHIFT to select those cells.

   PRO TIP: use the F4 key to anchor the O4 cell. Anchoring means the formula will continue to reference O4 regardless into which cells we move the formula. You know a reference is anchored because the $ sign appears before the column letter and row number (see image below).

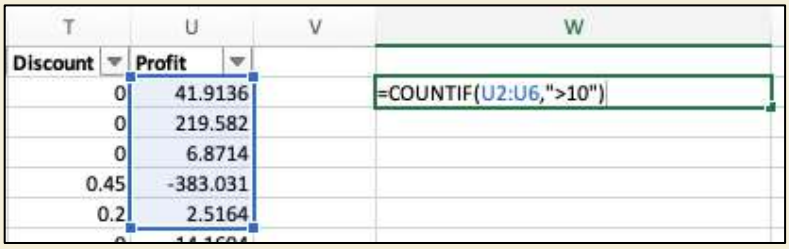
f. The "criteria" argument should include a reference to the cell with criteria, which is "Office Supplies" in this case. Use the arrow key to select either cell O4 or O6.
g. The sum_range argument is the adjacent array of cells that will be summed based on the criteria, in this case cells U2:U6. Use the arrow keys and press and hold SHIFT to select those cells.
h. Press enter – the result is the sum of those three profits, in this case 9.388 (see image below)



=COUNTIF(range, criteria)

The =COUNTIF() function is very similar to =SUMIF(), but you don't need a conditional range because the "counting" logic is applied to numerical and non-numerical values in the same way.

Let's use cells U2:U6. Imagine you want to count the number of ordered products whose profits that are greater than 10. To do so, type =COUNTIF(U2:U6,">10") in cell W2. This will return 2 (see image below) because only two orders generated a profit greater than 10.



Note that the criteria, even when it's not text, must be in quotation marks for =COUNTIF().

=AVERAGEIF(range, criteria, [average_range])

The =AVERAGEIF() formula works the same way as =SUMIF(), only it returns the arithmetic mean of the average_range rather than the sum.

=MINIFS(min_range, criteria_range1, criteria1, [criteria_range2, criteria2], ...)

Let's continue to use cells U2:U6. Imagine you want to find the minimum profit that is greater than 200.

a. Place your cursor in an empty cell, such as cell W2.
b. Press the "=" sign to signal that you're entering a formula to Excel
c. Start typing MINI… until you see the autofill for the =MINIFS() formula appears. You'll note that MINIFS is plural. Excel does not support a single conditional MIN function, but this does not impact the user intent because we can use only one condition even though several are available.
d. Click tab to engage the formula
e. The "min_range" argument should include the cells in which we want to find the minimum, in this case U2:U6. Use the arrow keys and press and hold SHIFT to select those cells.
f. The "criteria_range1" argument should include a reference to the cells or array on which we will apply the >200 criteria. Use the arrow key to select cells U2:U6.
g. The "criteria1" argument is where we will apply our greater than argument. Type ">200" with quotation marks.
h. Press enter – the result is the minimum value greater than 200, or 219.58 (see image below)

=MAXIFS(max_range, criteria_range1, criteria1, [criteria_range2, criteria2], ...)

The =MAXIFS() function works the same way as MINIFS, only it will return the largest value based on conditions within an array.

=CONCATENATE(text1, [text2], ...)

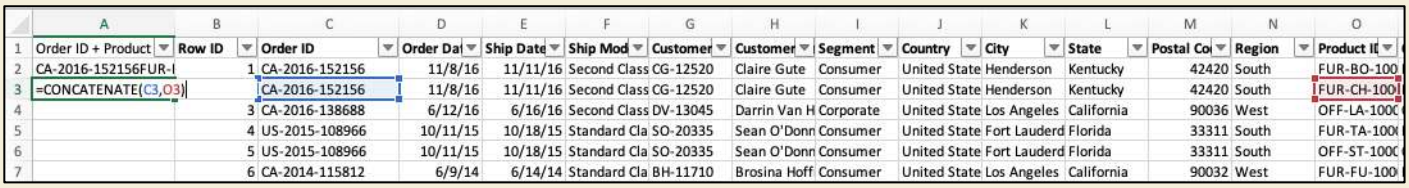
Imagine you want to combine multiple cell values to create a unique ID for a line that otherwise does not have one. This is a classic use case in data analysis, and the best way to do it is with =CONCATENATE().

Let's look at cells B7:U13 in the Super store worksheet. Can you point to a unique identifier for each line (other than the row number)? Currently, there isn't one, and this is not uncommon.

We can see that the Order ID is the same for each line. In fact, the only non-numerical columns without a repeating value are Product ID, Product Name, and Sub-Category. It's just luck that Sub-Category does not repeat in this Order ID, because we can see it repeats for other Order IDs such as US-2015-150630.

That leaves Product ID and Product Name, which are two different ways of writing the same information. We can conclude, then, that the identifier for each row is a combination of Order ID and product information.

To create this unique ID in a row by itself, we use =CONCATENATE (Order ID, Product ID). See the image below.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Order ID + Product	Row ID	Order ID	Order Da	Ship Date	Ship Mod	Customer	Customer	Segment	Country	City	State	Postal Co	Region	Product ID
2	CA-2016-152156FUR-I	1	CA-2016-152156	11/8/16	11/11/16	Second Class	CG-12520	Claire Gute	Consumer	United State	Henderson	Kentucky	42420	South	FUR-BO-100
3	=CONCATENATE(C3,O3)		CA-2016-152156	11/8/16	11/11/16	Second Class	CG-12520	Claire Gute	Consumer	United State	Henderson	Kentucky	42420	South	FUR-CH-100
4		3	CA-2016-138688	6/12/16	6/16/16	Second Class	DV-13045	Darrin Van H	Corporate	United State	Los Angeles	California	90036	West	OFF-LA-100
5		4	US-2015-108966	10/11/15	10/18/15	Standard Cla	SO-20335	Sean O'Donn	Consumer	United State	Fort Lauderd	Florida	33311	South	FUR-TA-100
6		5	US-2015-108966	10/11/15	10/18/15	Standard Cla	SO-20335	Sean O'Donn	Consumer	United State	Fort Lauderd	Florida	33311	South	OFF-ST-100
7		6	CA-2014-115812	6/9/14	6/14/14	Standard Cla	BH-11710	Brosina Hoff	Consumer	United State	Los Angeles	California	90032	West	FUR-FU-100

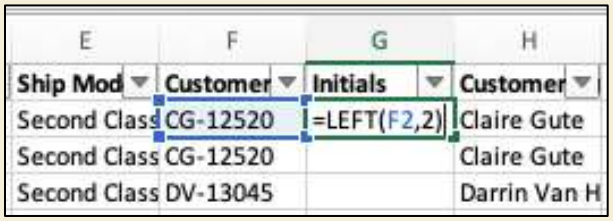
When we copy/paste this formula all the way down the column, we have a new field with no repeating values – a primary key.

=LEFT/RIGHT(text,[num_chars])

Note that in the concatenate section above we inserted a new column, which displaced the values of the other columns to the right. But we ignore that here.

In many cases, data analyst need to isolate parts of the contents in a cell. Imagine you want to anonymize your customer names by using only their initials. Customer initials are the first two characters in the Customer ID column. To extract them, we can use the =LEFT() function.

a. Create a new column by clicking on the column G header (click the letter G).
b. Right click and select "Insert"
c. In the new cell G1, which is empty, write =LEF… until the autofill for the =LEFT() function appears
d. Press tab to engage the function
e. Use the arrow keys to reference cell F2 for the "text" argument
f. Write 2 without quotations for the num_chars argument
g. Hit enter – you should have the first two characters of F2, which in this case is "CG"
h. The same steps apply for the =RIGHT() function, only the starting point is from the right side of the cell. In this example, =RIGHT() would return "20". Give it a try!



E	F	G	H
Ship Mod	Customer	Initials	Customer
Second Class	CG-12520	=LEFT(F2,2)	Claire Gute
Second Class	CG-12520		Claire Gute
Second Class	DV-13045		Darrin Van H

=LEN(text)

Imagine you want to identify the number of characters in a cell. This is the purpose of the =LEN() function. Simply reference a cell for the "text" argument and the function will return the number of characters. Referencing cell F2, for example, would return 8.

=MID(text, start_num, num_chars)

Imagine you want to extract a portion of a cell's contents but not starting from the left or right. For example, you would like to extract the 2-character sub-category flag in the Product ID column. The =MID() functions makes this very easy.

a. Create a new column by clicking on the column O header (click the letter O).
b. Right click and select "Insert"
c. In the new cell O1, which is empty, write =MI… until the autofill for the =MID() function appears
d. Press tab to engage the function

e. Use the arrow keys to reference cell N2 for the "text" argument
f. Write 5 without quotations for the start_num argument because the sub-category id starts at that character
g. Write 2 without quotations for the num_chars argument, because the sub-category id is two characters long
h. Hit enter – you should have the sub-category id from cell N2, which in this case is "BO"

M	N	O	P
Region	**Product ID**	**SubCat ID**	**Category**
South	FUR-BO-10001798	=MID(N2,5,2)	
South	FUR-CH-10000454		Furniture
West	OFF-LA-10000240		Office Suppli
South	FUR-TA-10000577		Furniture

=SEARCH(find_text, within_text, [start_num])

Imagine you would like to find the location of a specific text within a cell. You might want to find this for use as the "text" argument in the =MID() function we just looked at. The =SEARCH() function does this for.

Let's try it on the same example from the =MID() function above. In Cell O3, write =SEARCH("-",N2), where "-" is the text we're looking for and N2 is the cell we're examining (see image below). This will return 4.

N	O	P
Product ID	**SubCat ID**	**Category**
FUR-BO-10001798	BO	Furniture
FUR-CH-10000454	=SEARCH("-",N2)	Furniture
OFF-LA-10000240		Office Suppli
FUR-TA-10000577		Furniture

We can then plug in this function to the start-num argument in the =MID() function, **plus 1**. We must add one because we want to start just after the hyphen, so the function would be =SEARCH("-",N2)+1. The result is the same, "BO".

N	O	P
Product ID	**SubCat ID**	**Category**
FUR-BO-10001798	=MID(N2,SEARCH("-",N2)+1,2)	Furniture
FUR-CH-10000454	5	Furniture
OFF-LA-10000240		Office Suppli
FUR-TA-10000577		Furniture

=INDEX(array, row_num, [column_num])

Sometimes, we what to know the value of a cell at a specific position within a table or array. We can return this value using the =INDEX() function.

For example, image you want to find the sale value of Lena Hernandez's purchase of the TEC-AC-10002167 product. We know this is in row 49 (because we filtered for it), and we know Sales are in column 19. You would write into an empty formula, such as W2, =INDEX(A1:V9979,49,19). The result is **45**.

In practice, you will probably never use =INDEX() alone. Instead, you will combine it with =MATCH(). Let's look at it now.

=MATCH(lookup_value, lookup_array, [match_type])

The match function is similar to =INDEX(), except it returns the location of a value within a one-dimensional array. For example, imagine you want to know what row Lena Hernandez's order of product TEC-AC-10002167 is. We know what Order ID and Product ID are the unique IDs for each row, so we can create a new concatenation that includes Lena Hernandez's name for the =MATCH() function, which will be the value of the lookup_value argument.

a. Create a new column B by clicking on the "B" in the Excel row letter, then right-clicking > Insert
b. Title this column Primary Key
c. In the new cell B2, write the formula =CONCATENATE(C2,H2,O2), which will return CA-2016-152156Claire GuteFUR-BO-10001798 (Order ID + Customer Name + Product ID).

B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
Primary Key	Order ID	Order Da	Ship Date	Ship Mod	Customer	Customer	Segment	Country	City	State	Postal Co	Region	Product II	Category
=CONCATENATE(C2,H2,O2)	CA-2016-152	11/8/16	11/11/16	Second Class	CG-12520	Claire Gute	Consumer	United State	Henderson	Kentucky	42420	South	FUR-BO-100	Furniture
CA-2016-152156Claire GuteFUR-CH-10(	CA-2016-152	11/8/16	11/11/16	Second Class	CG-12520	Claire Gute	Consumer	United State	Henderson	Kentucky	42420	South	FUR-CH-100(	Furniture
CA-2016-138688Darrin Van HuffOFF-LA	CA-2016-138	6/12/16	6/16/16	Second Class	DV-13045	Darrin Van H	Corporate	United State	Los Angeles	California	90036	West	OFF-LA-100(	Office Suppli
US-2015-108966Sean O'DonnellFUR-TA	US-2015-108	10/11/15	10/18/15	Standard Cla	SO-20335	Sean O'Donn	Consumer	United State	Fort Lauderd	Florida	33311	South	FUR-TA-100(	Furniture
US-2015-108966Sean O'DonnellOFF-ST	US-2015-108	10/11/15	10/18/15	Standard Cla	SO-20335	Sean O'Donn	Consumer	United State	Fort Lauderd	Florida	33311	South	OFF-ST-100(	Office Suppli
CA-2014-115812Brosina HoffmanFUR-F	CA-2014-115	6/9/14	6/14/14	Standard Cla	BH-11710	Brosina Hoff	Consumer	United State	Los Angeles	California	90032	West	FUR-FU-100(	Furniture

From there, we can use this concatenation as the lookup_value argument in the =MATCH() function.

a. In cell W2, type =MATCH(CA-2016-169194Lena HernandezTEC-AC-10002167,B1:B9979,0) and press enter. The result should be **49** — the row in which the concatenation appears.

As with =INDEX(), it's rare to use =MATCH() alone, instead, we combine the two: INDEX + MATCH.

INDEX+MATCH

Imagine you are told to retrieve the profit for Lena Hernandez's purchase of the TEC-AC-10002167 product, and you don't want to use any hard-coded numbers.

To do this very easily, we can use INDEX+MATCH.

a. In cell X2, type =INDEX(A1:V9979,MATCH(B49,B1:B9979,0),MATCH(S1,A1:V1,0)).
   - The first argument defines the range of the search, which is the whole table A1:V9979.
   - The second argument defines the row number based on a MATCH of the concatenation across all of column B.
   - The third argument defines the column number based on a MATCH of the Sales column name across all of the column headers.
b. Press enter—the return will be **45**. As you can see, this is a much more efficient approach than using =INDEX() or =MATCH() alone!

=XLOOKUP(lookup_value, lookup_array, return_array, [if_not_found], [match_mode], [search_mode])

The last function we will look at in XLOOKUP. Before this function was developed, many people used VLOOKUP and HLOOKUP, which allow you to lookup a value in an array based on an input at the row (VLOOKUP) or column (HLOOKUP).

You may still see them used, but XLOOKUP is much more flexible, and it's also the direction Microsoft is taking for the future, so I encourage this function.

So, what does =XLOOKUP() do? Much like INDEX+MATCH, it allows you to return the value from an array based on its adjacent position with a value in a first array.

For example, imagine you want to find the profit again for Lena Hernandez's order of product TEC-AC-10002167. In this case you would do the following steps.

a. In cell X2, type =XLOOKUP(B49,B1:B9979,S1:S9979)
   - B49 represents the concatenation we used above for lookup_value in the =MATCH() function (CA-2016-169194Lena HernandezTEC-AC-10002167).
   - The second argument specifies that the looking array is column B.
   - The third argument specifies that the return value should come from column S, the Profit column.
b. Hit enter—the return is **45**.

=XLOOKUP() is, in other words, a quick way than INDEX+MATCH when there is only one direction (row or column) on which you want to provide criteria.

Pivot Tables

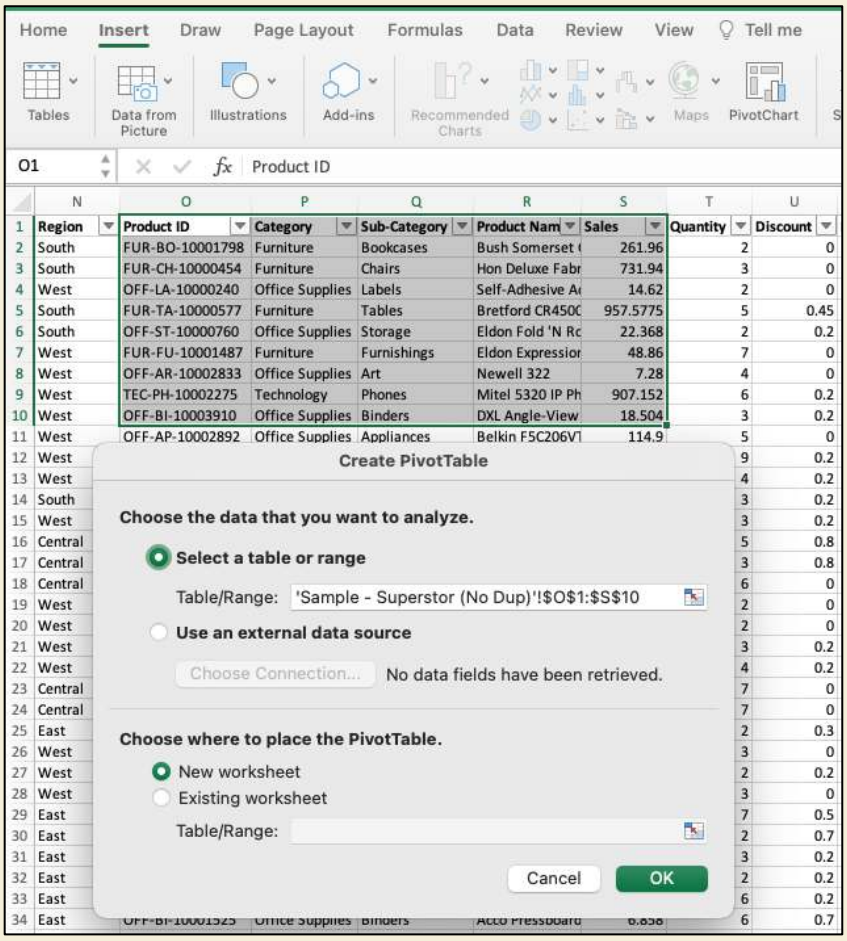
Now that we've looked at functions, it's time to get to know PivotTables. A pivot table is a table of **grouped values** that **aggregates** the individual items of a more extensive table within one or more **discrete** categories.

That may sound overwhelming, but pivot tables are easy to understand in practice, and they're very useful. They take the input from a raw data table and allow you to select dimensions (non-numerical columns) on which you want to see either other dimensions or measures (numerical values).

The best way to understand them is by example, so let's look at one now. Let's use the range O1:S10 in the super store data table.

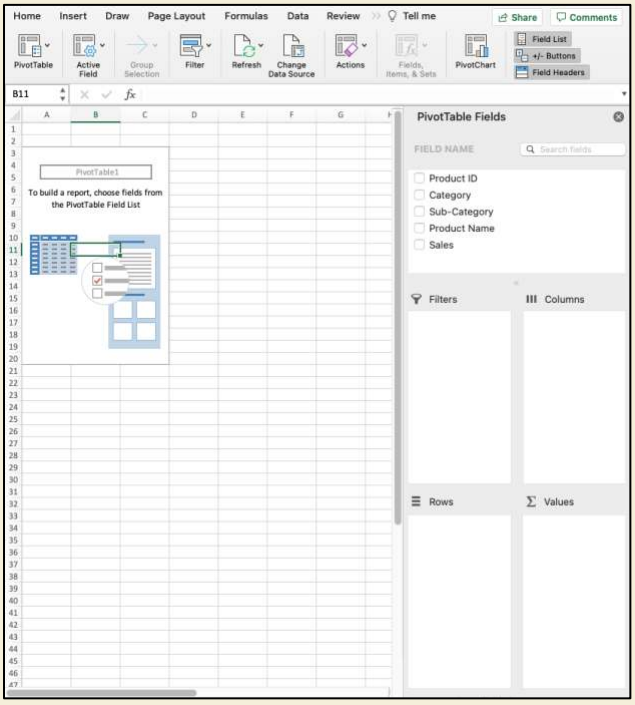
O	P	Q	R	S
Product ID	**Category**	**Sub-Category**	**Product Nam**	**Sales**
FUR-BO-10001798	Furniture	Bookcases	Bush Somerset (	261.96
FUR-CH-10000454	Furniture	Chairs	Hon Deluxe Fabr	731.94
OFF-LA-10000240	Office Supplies	Labels	Self-Adhesive A	14.62
FUR-TA-10000577	Furniture	Tables	Bretford CR4500	957.5775
OFF-ST-10000760	Office Supplies	Storage	Eldon Fold 'N Rc	22.368
FUR-FU-10001487	Furniture	Furnishings	Eldon Expressior	48.86
OFF-AR-10002833	Office Supplies	Art	Newell 322	7.28
TEC-PH-10002275	Technology	Phones	Mitel 5320 IP Ph	907.152
OFF-BI-10003910	Office Supplies	Binders	DXL Angle-View	18.504

a. Highlight this selection with your cursor, then navigate to Insert > PivotTable.
b. You will see a popup that looks like the below.

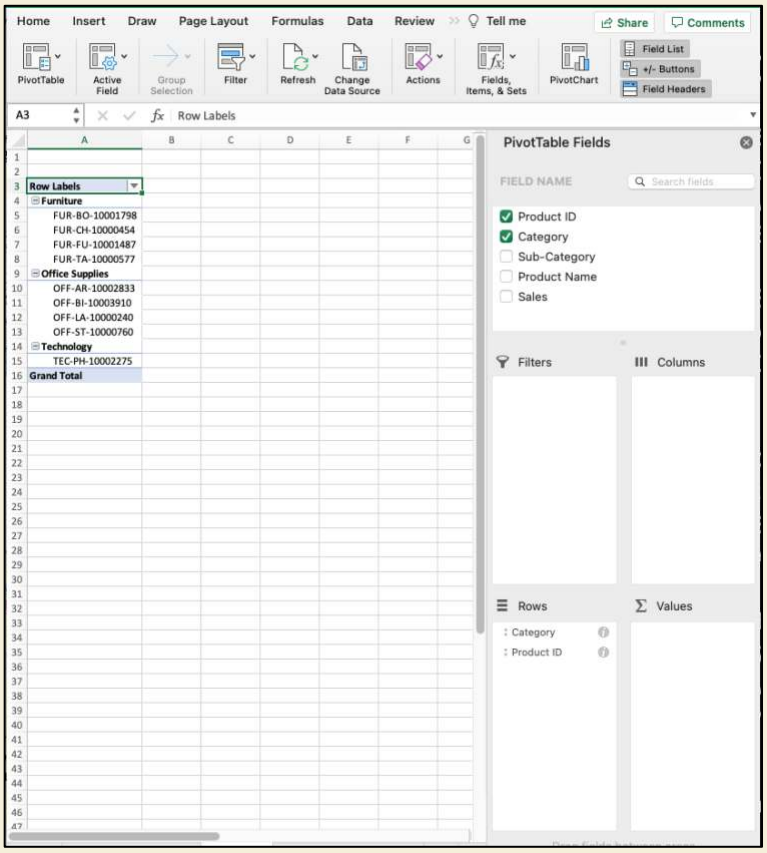


c. You can see that the range you highlighted is already the target range and the table will be created in a new worksheet. Click OK.

d. You will now see the screen below.

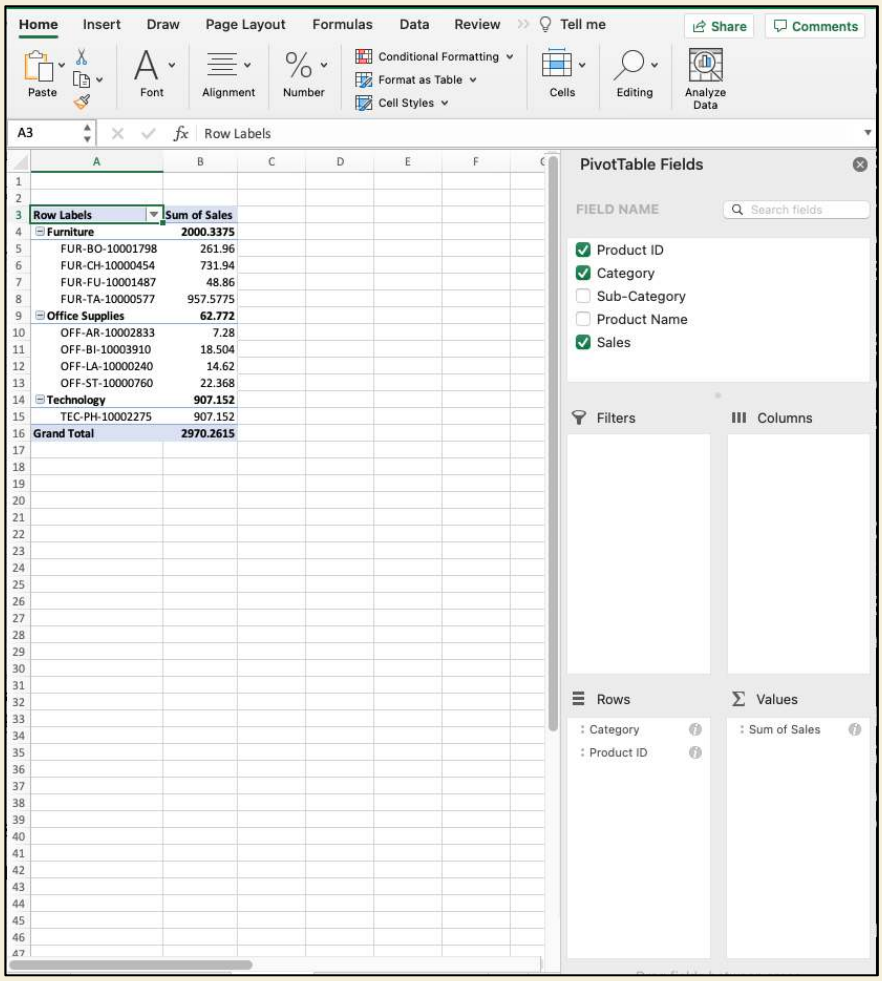


e. You can see the area that will become the PivotTable with instructions to choose fields from the field list on the right-hand side of the spreadsheet. To see how this works, drag the "Category" field from the field list to the Rows section, then drag "Product ID" just underneath it. You should see the following.

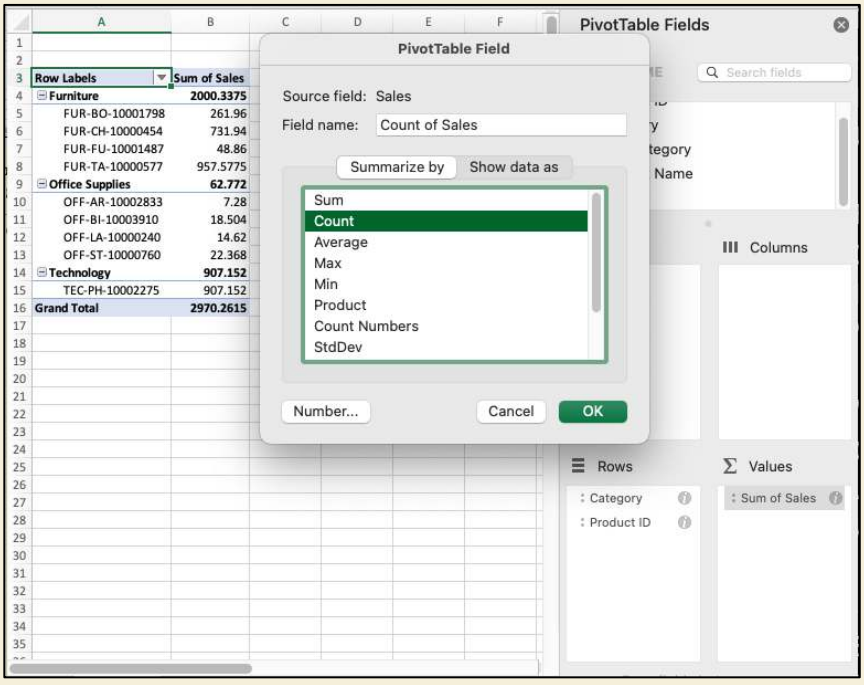


f. As you can see, the products are allocated to each category available in the original data table. **This is how dimensions can be layered on top of each other** to get a different view of the data.

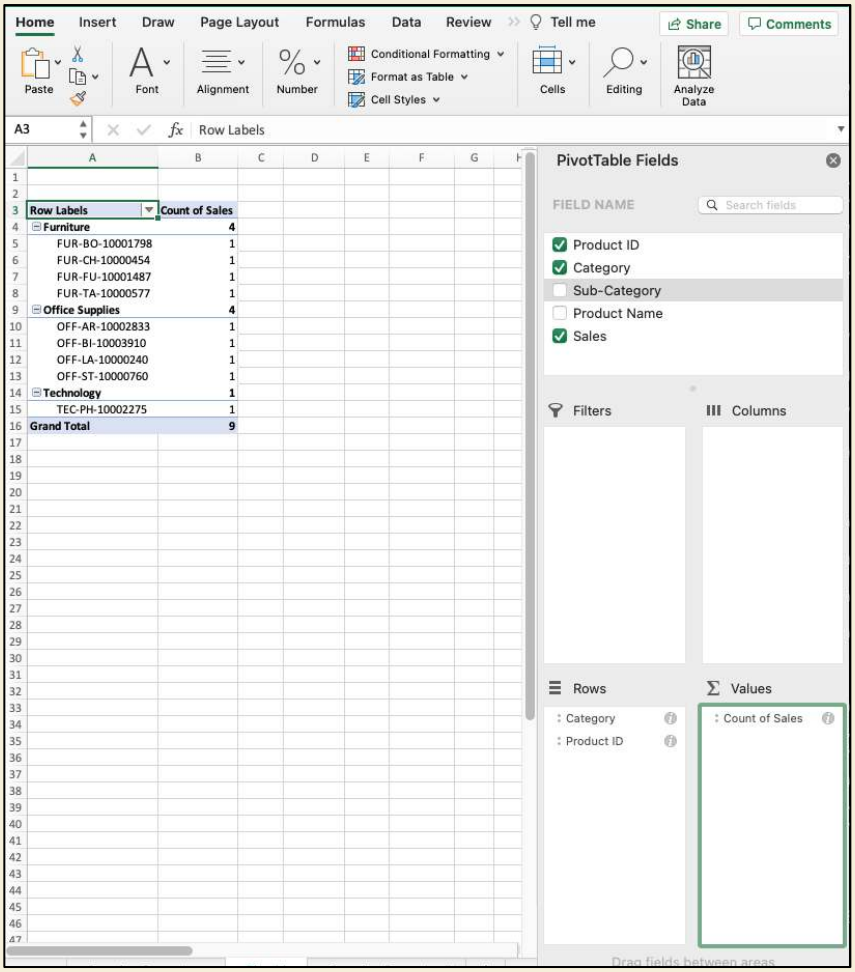
g. But what about numbers? In a PivotTable, numbers should go under the Values section in the bottom right corner. Try now by dragging Sales there. See the image below.



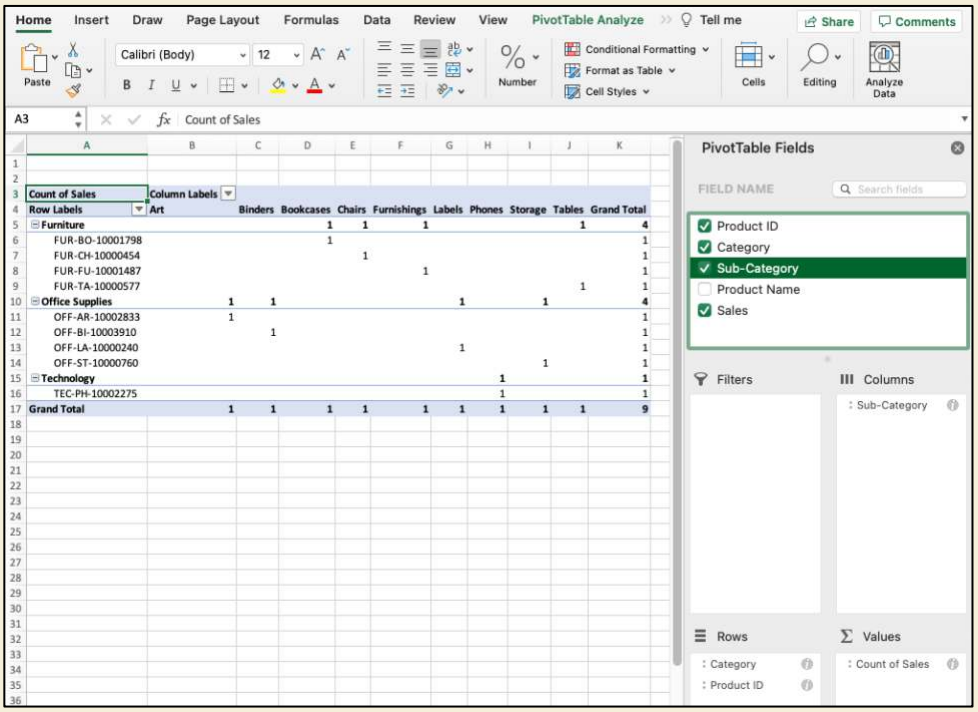
h. You can see in the Values section "Sum of Sales," which means Excel automatically adds the values together at the given **level of detail** (Category and Product ID). We can ask Excel to perform any of the 6 aggregation functions (except for count distinct) by right-clicking the pill in the Values section and changing the "Summarize By" field. Here's what COUNT selection looks like:



i. When you press OK, the values in the PivotTable are counted:



j. We can also add columns to the view, which will further segregate the data. Try this by adding Sub-Category to the columns section. The result should look like this:



k. Finally, we can filter fields by adding them to the upper-left filter section. You can see what this looks like by dragging Product Name to the Filter section and deselecting one or more.

Section Quiz 3

1. What function returns the largest value in an array?
2. What function returns a specified value when an error message occurs on the base calculation?
3. What function computes the sum of an array based on criteria in the same or another array?
4. What function combines the values of multiple cells into one?
5. What function returns the number of characters in a cell?
6. What function returns the position of a cell within an array?
7. What combination of functions returns the value of a cell at a specified row and column location in an array, where the location coordinates are the result of a nested function?
8. What do PivotTables do?

Answers here.

It's All About Averages

Now you know the essential Excel formulas and PivotTables, which means you know how to source data and manipulate it. However, we haven't addressed how to analyze data series with statistics. Recall the definition of data analysis:

Data *analysis* consists of thinking critically about organized information

Specifically, data analysis boils down to (1) manipulating unique IDs, (2) applying the 6 aggregation functions, (3) taking averages and related statistical functions, and (4) applying correlations. That's it. Once you understand and can apply these skills to answer questions, you're doing the job of a data analyst.

In this section, we'll look at point 3 above—applying averages and related statistical measures to analyze data series.

The simplest way to do this is through an average. You already know how to do this by applying the =AVERAGE() function. But what does it really means to take an average?

If you recall middle school math, an average is the sum of all numbers in a series divided by the count of numbers in the series. In function language, it would look like this:

$$\frac{= SUM(array)}{= COUNT(array)}$$

In other words it's:

$\frac{a+b+c+\cdots}{n}$, where n is the count of values in the series, also shown as $\bar{x}$

The result of the average is a value that *represents the central point in the set*. I don't need to explain averages any more than that—you are already familiar with it. However, it's important to understand the idea of a "central point" because it's the key driver for the other 4 statistical calculations we'll look at: variation, standard deviation, covariance, and correlation.

As an example, use the =AVERAGE() function on the first three values in Profit, V2:V4. The result should be 89.46.

Variance

Variance is a very simple way of using the average in a series to determine "how much" the values in a series diverge from each other. The formula is

$\frac{\sum(x_i-\bar{x})^2}{n-1}$, where n is the count of values in the series, also shown as S^2

So what does this formula mean? It means we need to take the average in a series, then subtract the average from each point in the series, one by one, and raise the result of each subtraction to the power of 2 to eliminate any negative values. This will show the "distance" that each point has from the *central point* (average) in the series.

Then, we add those value together and divide by the count of values in the series. In this sense, you can think of variance as "the average of the distance of each point from the center of the series."

As you can imagine, the variance value will be very high because it *squares the difference* of each point to the mean. This means it's hard to interpret the results of variance against the values in the series alone. For example, let's use the function =VAR.S() on cells V2:V4—profit cells. You can do this in cell X2. The result will be 13,006.64.

	T	U	V	W	X
	Discount	Profit		Variance	
	0	41.9136		=VAR.S(U2:U4)	
	0	219.582			
	0	6.8714			
	0.45	-383.031			
	0.2	2.5164			

Alone, that number is not very helpful because we cannot compare it with a benchmark. It's much too high to compare with the values in the set itself.

Instead, **we should use variance always when comparing two independent sets**. So let's use the same =VAR.S() function on cells V5:V7, then V17:V19. The results should be 51,091.79 and 82.54.

	T	U	V	W	X
	Discount	Profit		Variance	
	0	41.9136		13,006.64	
	0	219.582			
	0	6.8714			
	0.45	-383.031		51,091.79	
	0.2	2.5164			
	0	14.1694			
	0	1.9656			
	0.2	90.7152			
	0.2	5.7825			
	0	34.47			
	0.2	85.3092			
	0.2	68.3568			
	0.2	5.4432			
	0.2	132.5922			
	0.8	-123.858			
	0.8	-3.816		=VAR.S(U17:U19)	
	0	13.3176			
	0	9.99			
	0	2.4824			

In other words, we know that set 1 is 0.25 as varied as set 2 (13,006.64/51,091.79) and 157 times more varied than set 3.

What does this mean from an analytical perspective? It means that the profits from the orders and products in the first set are ¼ as spread as those in set two—the points in set 1 are more related.

Standard Deviation

Standard deviation is probably more familiar to you as a concept. Like variance, it represents the data set's collective distance from the mean.

In fact, standard deviation is simply the square root of variance. Here's the formula:

$$\sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}$$, where n is the count of values in the series, also shown as σ

The difference is that standard deviation values are in the same order of magnitude as variance, so it's easier to compare. You might be asking, "why use variance at all?"

Well, standard deviation is slightly less reliable than variance because we're adding an additional operator: squared root. In practice, this means you should always **prefer standard deviation when analyzing one set**, but **variance when comparing multiple sets**.

Let's apply standard deviation to the same cells we applied variance:

V	W	X	Y
Profit		Variance	Standard Deviation
41.9136		13006.63592	=STDEV.S(V2:V4)
219.582		=VAR.S(V2:V4)	=STDEV.S(V2:V4)
6.8714			
-383.031		51091.7913	226.0349338
2.5164		=VAR.S(V5:V7)	=STDEV.S(V5:V7)
14.1694			
1.9656			
90.7152			
5.7825			
34.47			
85.3092			
68.3568			
5.4432			
132.5922			
-123.858			
-3.816		82.53980112	9.085141778
13.3176		=VAR.S(V17:V19)	=STDEV.S(V17:V19)
9.99			

As you can see, the standard deviation is much smaller than variance – by exactly the square root in fact. The number is intuitively easier to understand compared with the values in the series. For example, look at Cell Y5, which is the standard deviation of -383.031, 2.5164, and 14.1694.

At 226.035, the standard deviation shows that the absolute distance from the mean in this data series is 226, which makes sense given the outlier negative -383.

Practically speaking, this high standard deviation means the data points are highly spread – they're not related.

Section Quiz 4

1. In simple terms, what does an average tell us about a data set?
2. What is variance, and what is its key limitation?
3. What is standard deviation?
4. When should you use variance, and when should you use standard deviation?

Answers here.

Except When It's About Correlations

Now that we understand averages and the two functions related to them—variance and standard deviation—you're ready to perform a vast amount of analysis. Indeed, most questions we want to answer about data come from these simple statistical functions.

However, what we've seen so far has been focused on *a single data series*. In many cases, we want to understand the relationship between two data series. For example, we may want to understand the relationship between profits and quantity of products sold. Intuition tells us the two should be highly connected.

To understand the relationship, we need to look at covariance and correlations.

Covariance

Covariance is similar to variance with two key changes: we add a second series (y) to the numerator and remove the squared operator:

$\frac{\sum(x-\bar{x})*(y-\bar{y})}{n-1}$, where n is the count of value in the series, also known as cov(X,Y)

What does this mean? The inclusion of a second variable means the output is an assessment of the relationship between two random inputs, x and y. This allows us to make a comparison directly in the formula. The removal of the square means we're now allowing negative values.

Why? With variance, the goal is to determine the absolute distance of one variable from the series mean, but with covariance, we want to see the relationship, whether negative or positive, between the variables. Let's apply covariance to the same profit arrays as variance and standard deviation, and include now quantity.

	T	U	V	W	X	Y	Z
1	Quantity	Discount	Profit		Variance	Standard Deviation	Covariance
2	2	0	41.9136		13006.63592	114.0466392	=COVARIANCE.S(T2:T4,V2:V4)
3	3	0	219.582		=VAR.S(V2:V4)	=STDEV.S(V2:V4)	=COVARIANCE.S(T2:T4,V2:V4)
4	2	0	6.8714				
5	5	0.45	-383.031		51091.7913	226.0349338	-50.66273333
6	2	0.2	2.5164		=VAR.S(V5:V7)	=STDEV.S(V5:V7)	=COVARIANCE.S(T5:T7,V5:V7)
7	7	0	14.1694				
8	4	0	1.9656				
9	6	0.2	90.7152				
10	3	0.2	5.7825				
11	5	0	34.47				
12	9	0.2	85.3092				
13	4	0.2	68.3568				
14	3	0.2	5.4432				
15	3	0.2	132.5922				
16	5	0.8	-123.858				
17	3	0.8	-3.816		82.53980112	9.085141778	8.4842
18	6	0	13.3176		=VAR.S(V17:V19)	=STDEV.S(V17:V19)	=COVARIANCE.S(T17:T19,V17:V19)
19	2	0	9.99				
20	2	0	2.4824				

We can see that for cells T5:T7 (quantity) and V5:V7 (profit), the multiplied distances to the mean divided by 2 (number of points minus 1) is negative, which is driven both by the negative average of the profit entries, as well as the negative distance of point 2 to the average of cells T5:T7.

Since covariance is negative, we know that the relationship between the variables in inverse—when one goes down, the other goes up. And when one goes up, the other goes down. However, the size of the covariance is also important. Covariance is cell Z6 is -50, while in cell Z17, it's only 8.

Which one displays more tightly related variables? The answer is the first, because 50 is greater than 8. However, you should not depend on covariance for the tightness of the relationship. **Magnitude is only useful in the sense that we see how far away from zero the relationship is**.

When it's too close to zero, we question whether there is a relationship at all. At 8, the covariance is quite small and should not be considered as a statistically significant relationship.

We'll talk about correlation in the next section, which is much more useful for determining "how close" two variables are to each other.

Note that the example above is not ideal because we have a huge outlier in cell V5. Ideally, we would exclude that number. Moreover, another factor we should account for is the limited number of inputs we have. The **law of large numbers** states that more data results in a more accurate analysis. For educational purposes, the sets are small. In practice, 3 values are far too few.

Correlation

So it's all about averages… except when it's about correlations. **We saw that covariance is a useful measure to see the direction of the relationship and to ensure that it's non-zero**.

Correlation, however, provides a value between -1 and 1, called the correlation coefficient, which **standardizes** the closeness. Anything higher than 0.7 is considered strong, while anything less than 0.5 is weak, and anything between 0.5 and 0.7 is considered moderate.

As with the other 3 statistical computations, correlation depends on the average of two series:

$$\frac{\sum(x-\bar{x})*(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2*\sum(y-\bar{y})^2}}, \text{ also shown as } r$$

You can think of correlation as "the sum of multiplied distances from the mean of two variables, divided by the square rooted product of the sum of distances squared for both series." It's difficult to fully picture the impact of this calculation on a data set because there are many operators, but the key points are:

- The numerator takes "the sum of many products (multiplications)" whereas the denominator takes "the product of many sums," which means the denominator will always be greater
- The numerator can be negative, but the denominator will always be positive

For the first time, we're not using the count of items in the series as the denominator. Instead we're using values from the series themselves. This allows us the standardize the variance against itself, creating a benchmark—i.e. a value between -1 and 1. Let's apply this now to the same cells we've been using:

	T	U	V	W	X	Y	Z	AA
1	Quantity	Discount	Profit		Variance	Standard Deviation	Covariance	Correlation
2	2	0	41.9136		13006.63592	114.0466392	65.06316667	0.988128288
3	3	0	219.582		=VAR.S(V2:V4)	=STDEV.S(V2:V4)	=COVARIANCE.S(T2:T4,V2:V4)	=CORREL(T2:T4,V2:V4)
4	2	0	6.8714					
5	5	0.45	-383.031		51091.7913	226.0349338	-50.66273333	=CORREL(T5:T7,V5:V7)
6	2	0.2	2.5164		=VAR.S(V5:V7)	=STDEV.S(V5:V7)	=COVARIANCE.S(T5:T7,V5:V7)	=CORREL(T5:T7,V5:V7)
7	7	0	14.1694					
8	4	0	1.9656					
9	6	0.2	90.7152					
10	3	0.2	5.7825					
11	5	0	34.47					
12	9	0.2	85.3092					
13	4	0.2	68.3568					
14	3	0.2	5.4432					
15	3	0.2	132.5922					
16	5	0.8	-123.858					
17	3	0.8	-3.816		82.53980112	9.085141778	8.4842	0.448609163
18	6	0	13.3176		=VAR.S(V17:V19)	=STDEV.S(V17:V19)	=COVARIANCE.S(T17:T19,V17:V19)	=CORREL(T17:T19,V17:V19)

The results are all between -1 and 1, and we can see that the correlation between quantity and profit is almost perfect for the first group, but for the last group it is only moderately strong.

You can see how correlations are useful. They allow analysts to identify variables that are closely related. When one of the factors is under the control of the analyst's organization, it may signal that action should be taken.

Section Quiz 5

1. What is covariance?
2. What makes covariance different from variance?
3. What is a correlation?
4. Can correlations be negative?
5. What is different about the correlation formula compared to variance, standard deviation, and covariance?
6. What are correlations useful for?

Answers here.

5 Essential Charts

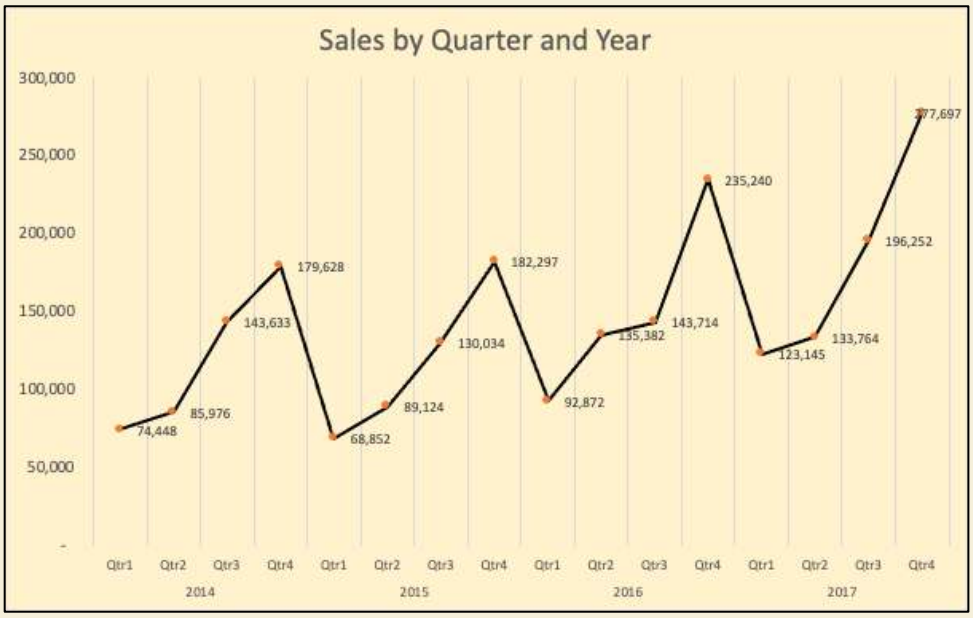
Data analysts dig into data to discover insights. However, they don't bring much value to an organization if they cannot communicate those insights. The best way to communicate is through data visualization.

Data visualization is a fancy way of saying charts or graphs. Good analysis use various forms of graphs to communicate ideas and messages. The good news is they only need five types: line graphs, bar charts, column charts, area charts, scatter plots, and waterfall charts.

PRO TIP: you should never use pie charts! While popular, they're no more valuable to the viewer than the percentage labels used to identify them. You might as well just show a list of percentages.

Line Graphs

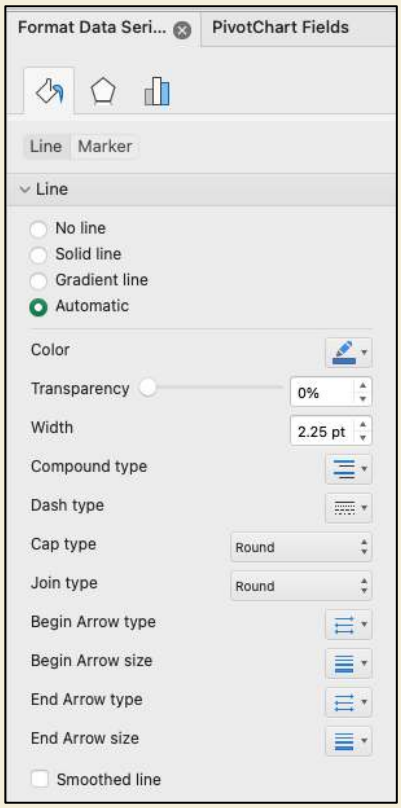
Line graphs are useful for showing events over time. Use them to depict changes over time on the x-axis.



Using the data from our superstore data table, this view shows sales by year and quarter. It's insightful because it shows a trend—each year, sales decline in Q1, and they slowly grow throughout the year to Q4. This is a classic commercial trend explained by the holiday season in Q4.

To build this view:

a. Create a PivotTable in a new tab with the full Superstore data table
b. Drag the Order Date Field to Rows
c. Drag the Sales field to Values and ensure the aggregation type is SUM
d. Click anywhere in the PivotTable, then navigate to Insert>Line>Line in the ribbon
e. Optional: click on one of the horizontal grid lines and press delete (these are not helpful)
f. Double-click the title and substitute the text "Sales by Quarter and Year"
g. Click on the square in the bottom right corner of the chart and drag it to enlarge the view (until the quarters in the axis label are horizontal and not vertical)
h. Optional: click on the legend and press delete – this is not useful because we only have one line
i. Right click the background > Format Chart Area… > Fill, and select a light orange
j. Right click on the axis label and select Add Major Gridlines
k. Click on the line until it's highlighted, then right click > Format Data Series. You will see a popup on the right side of the workbook that looks like this, where you can change the color of the line:



l. Click on the series again, then right click > Add Data Labels (you can format these in the same way as other items by right clicking > Format Data Labels…)
m. Click the dropdown menu (like the one in the image above) and select the color black
n. Then click on the box underneath the Fill sign at the top called "Marker" (like in the image above)
o. Navigate to Marker Options > Built-in > Type > •
p. Navigate to Fill > and select the color orange

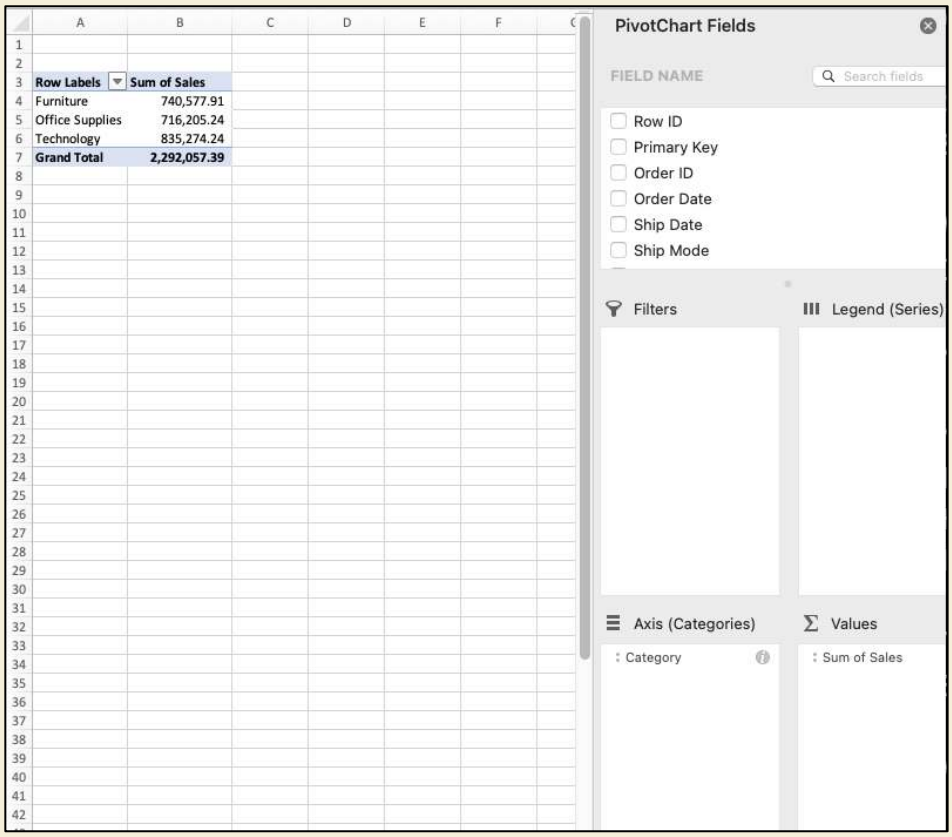
By now you probably understand how formatting work in Excel. You can either access the formatting options from Format > Format Pane in the ribbon, or simply by right-clicking the functionality you want to modify.

We won't cover formatting in the other 4 charts since formatting is a huge topic that's outside the scope of this book. But rest assured, if you want to see it, Excel most likely does it. We'll cover formatting in other contexts on AnalystAnswers.com.

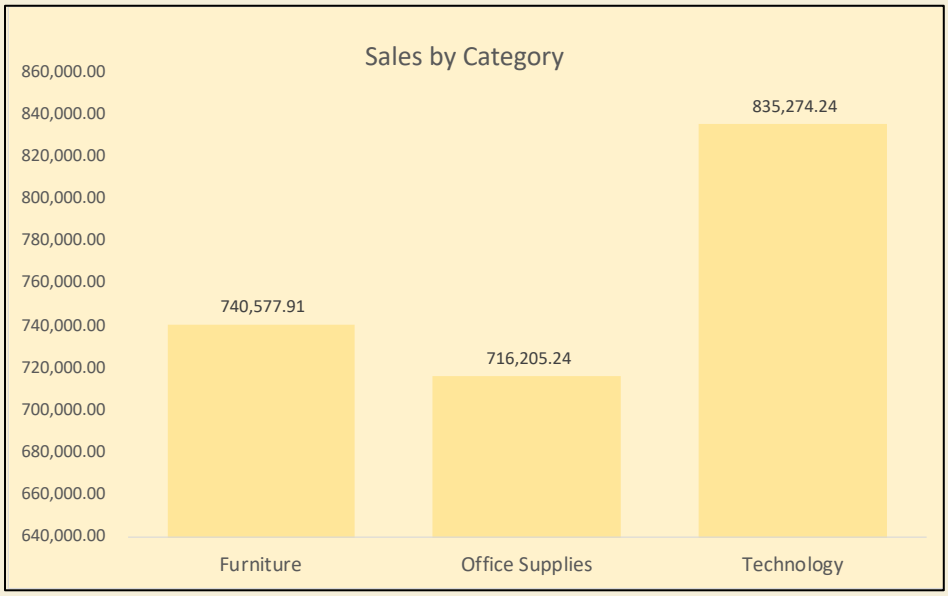
Column Charts

Column charts are useful when comparing the measures of different entries in a dimension. For example, using our super store data table, imagine we want to know the total sales produced **by category type**.

To do this, we should pivot the table again, only this time we drag the category field to rows and sales to the value section.

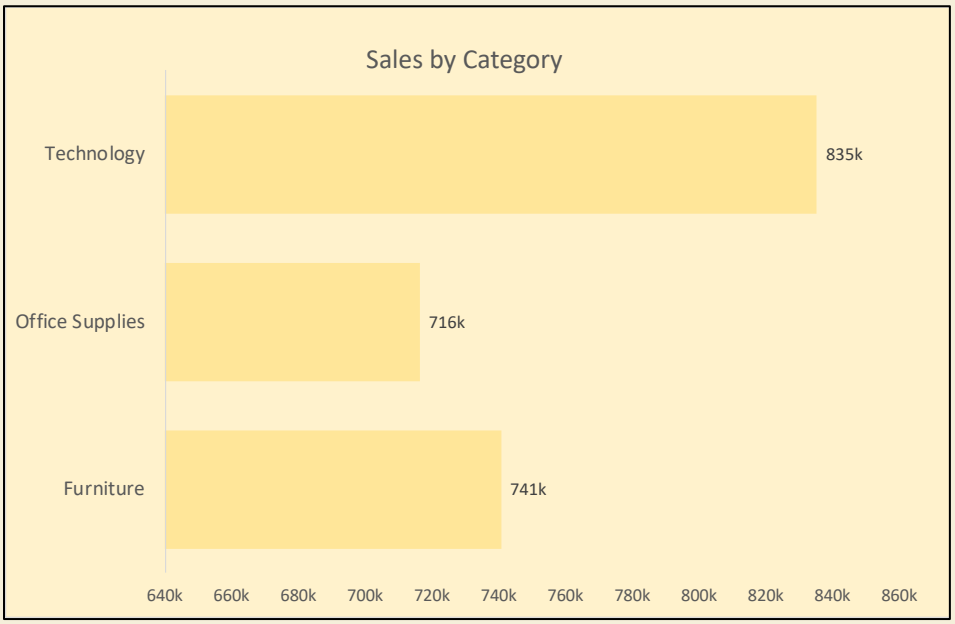


As with the line graph, navigate to Insert > Column > Column to create the basic view and format to your liking.



Bar Charts

Bar charts perform essentially the same role as column charts, meaning they're useful for comparing the values of multiple items in a column. Columns, however, are presented **horizontally**:

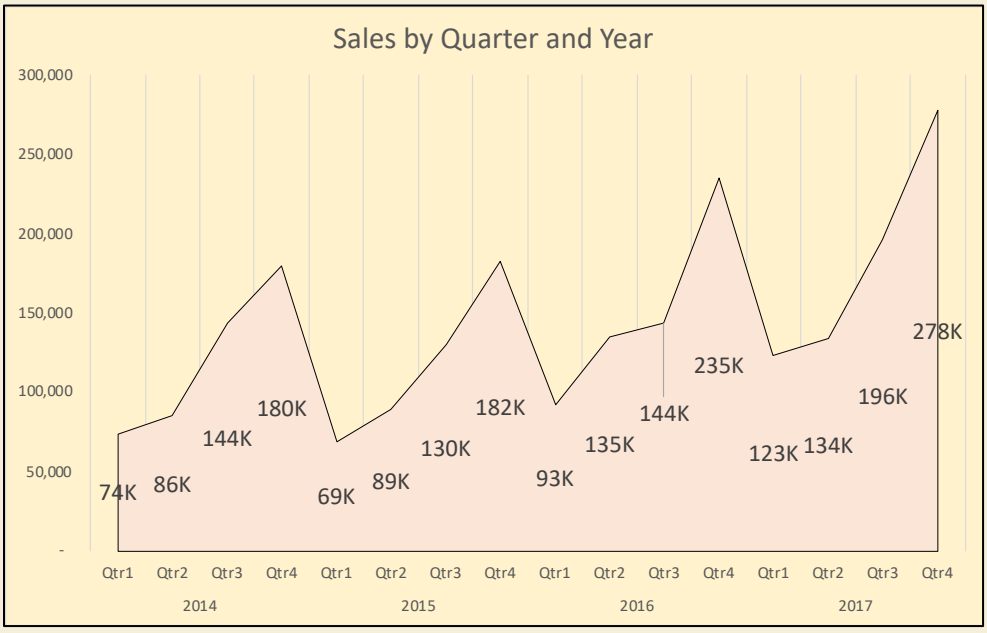


They tend to *accentuate the comparative size of values*, so the reader will unconsciously believe large or small numbers shown as a bar are greater than when shown as a column.

To create this view, copy the bar chart and paste it in a cell adjacent. Navigate to PivotTable Design > Change Chart Type > Bar. The change is automatic, and you can format to your liking.

Area Charts

Area charts are similar to line graphs, but they fill in the area under the graph, which tends to feel like the quantity is larger to the reader. Let's use the same graph we used for line, making it an area chart:



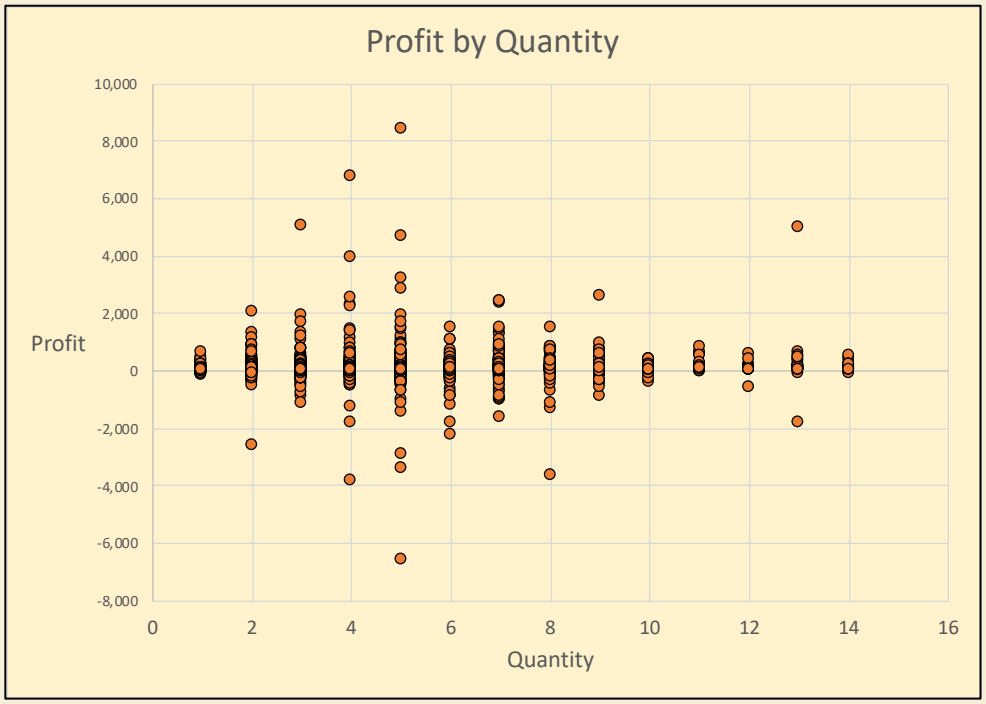
You can create this view by selecting the data and navigating to Insert > Line > Area. Alternatively, you can select the line chart, copy paste a second version, and navigate to PivotTable Design > Change Chart Type > Line > Area.

Scatter Plots

Scatter plots are useful for identifying relationships between two random variables, such as quantity and sales. The below graph is a visualization of the correlation example we looked at earlier. The difference here is that we're taking all the data points, rather than segments of three points.

What does this graph show us? Essentially, there is zero correlation between quantity and profit. Instead, we have many ordered products at quantity 5 – but the profit associated with these amounts is spread across a large negative to positive range.

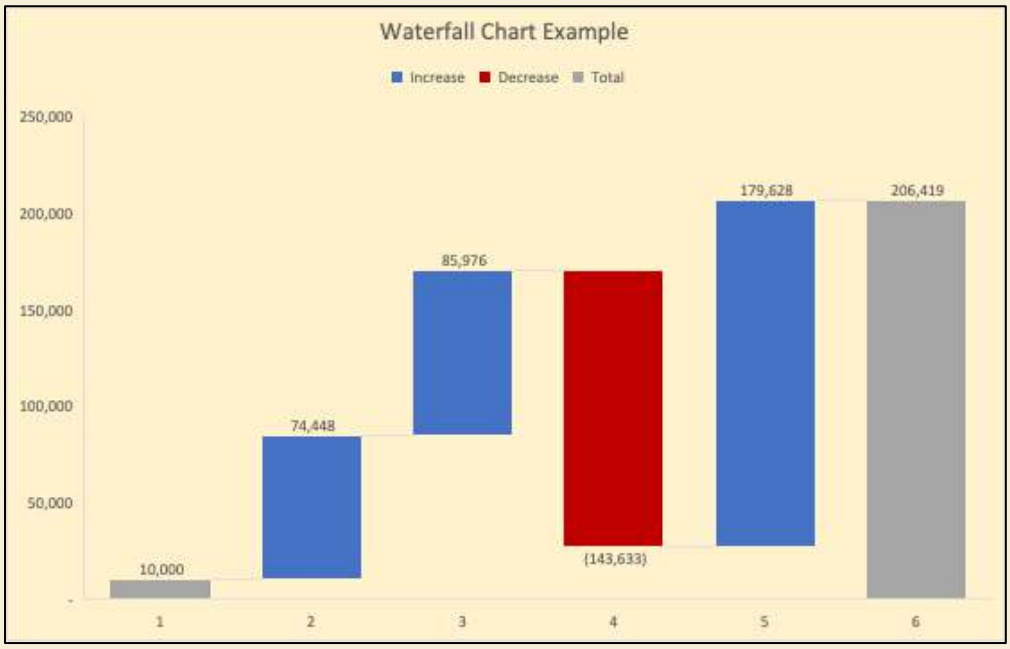
Using the correlation function, we can also calculate that the correlation coefficient is 0.07, which is less than 10%, an extremely low correlation (not to be confused with a 0.7 coefficient, which is strong!).



To create this view, select the data for quantity and profit, then navigate to Insert > Scatter > Scatter. Format to your liking.

Waterfall Charts

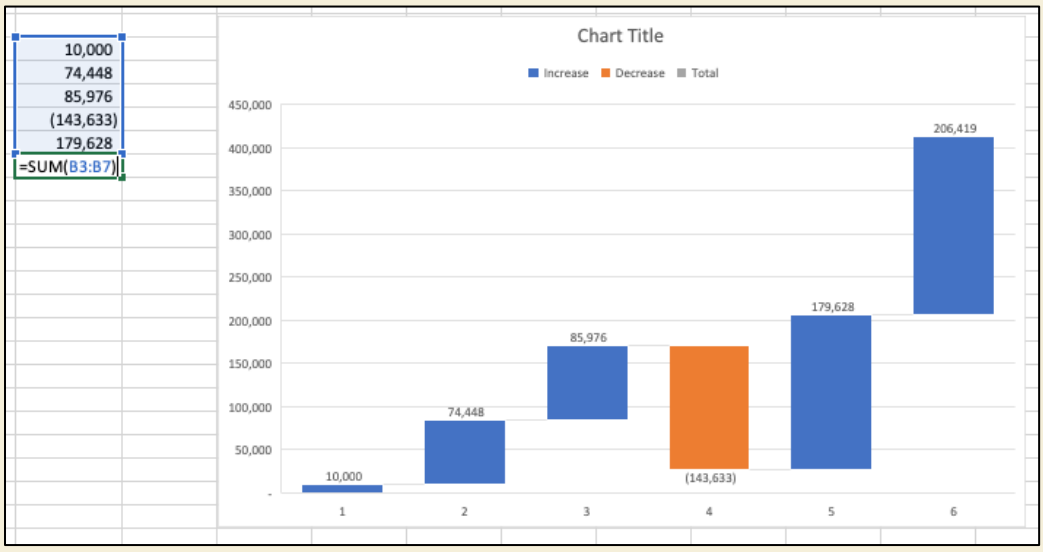
Waterfalls are the last chart we'll look at. These charts are most useful for showing the positive and negative contributions to change in value. For example, imagine we want to show the change in a cash balance that started at 10,000 and ended at 206,418 over a period on 6 days.



The waterfall chart above shows that on days 2, 3, and 5 we have net additions to the cash balance. However, on day 4 we had a significant reduction in the balance. Overall, the contributions ensure the value never becomes negative.

To create this chart, you have to set up the data in a specific way (see image below).

a. Ensure the first value in the series is the starting balance
b. Put the value of additions or reductions to the balance in cells below this first cell
c. Use the =SUM() formula to add the values from the starting cell to the last contribution/reduction
d. Select those values and navigate to Insert > Waterfall Chart
e. You should see the following view



f. From here, you need to right-click the first and last bars (starting and ending balances) > **Set As Total**
g. You can then change the color of the "decrease" legend entry directly on the fill color function under the Home tab. The negative bars will reflect this.

Section Quiz 6

1. Which chart would you use to show change in values for one dimension over time?
2. Which chart would you use to show the relationship between two variables?
3. Which chart would you use to compare the values of several instances of a dimension?
4. Which chart would you use to visually emphasize the magnitude of change over time?
5. Which chart would you use to show the additions and reductions in a balance over time?

Answers here.

23 Key Terms

Like any discipline, data analysis comes with a set of terms that you'll need to know. Here's the list with definitions and brief examples. Some of these definitions will only make sense when you see examples in "real-life," but take the opportunity to get started with them now so you'll understand them the second time later.

We've already covered some of these terms in the book, so you can think of this section as a glossary of terms as well. If we've addressed the word, I'll reference the page or link in the right column.

#	Term	Definition	Brief Example	Reference
1	Aggregation	the combination of dimensions or measures using the level of detail or calculation on other dimensions or measures	Adding the weights of multiple cars in a dimension to see the aggregate weight of all cars	Page Link
2	Average	The arithmetic mean of a data series	AVERAGE(3, 6, 9) = 6	No reference
3	Column	A vertical element in a data table that records one dimension of information for each entry	Profit in the super store data table	Excel workbook
4	Correlation	A measure of the strength and direction of the relationship between two series	The correlation between quantity and profit	Page Link
5	Data	Organized information, usually in the form of a data table	The super store data table	Page Link
6	Data analysis	Answering an unambiguous and unbiased question, data analysis is thinking critically about organized information, mainly through (1) manipulating unique IDs, (2) applying the 6 aggregation functions, (3) taking averages and related statistical functions, and (4) applying correlations	In the next section, we'll look at an example analysis	Page Link
7	Data attribute	Another term for data column, but connected to the entry itself, formally "a single-value descriptor for a data point or data object."	42 in profit or Furniture category for the Unique ID CA-2016-152156Claire GuteFUR-BO-10001798	Excel Workbook
8	Data field	Another term for data column, formally "a location for a predetermined type of data that — collectively with other data fields — describes the place it is stored."	Profit in the super store data table	Excel workbook
9	Data interpretation	The step after data analysis in the data cycle that helps the analyst create meaning from the analysis performed	In the next section, we'll look at an example analysis where we'll interpret the findings	Page Link

10	Data manipulation	The process of altering the structure of a data table either through aggregations or the creation of unique ID	The creation of a concatenation for the MATCH() function	Page Link
11	Data object	A collection of one or more data points that create meaning as a whole. In other words, "data object" is an alternate way of saying "this group of data should be thought of as standalone."	The super store data table is a data object	Page Link
12	Data point	a piece of information that describes one unit of observation, at one point in time, at the data collection level	Cell S2 in the super store data table	Page Link
13	Data set	a collection of one or more tables, schemas, points, and/or objects that are grouped together either because they're stored in the same location or because they're related to the same subject	The superstore data table	Page Link
14	Data table	a collection of columns and rows with unique IDs that organize information about an object or concept	The superstore data set	Page Link
15	Data visualization	The use of visualizations, usually charts and graphs, to display data in a way that makes it easier to communicate	The five essential charts	Page Link
16	Excel function	A built-in functionality in Excel that performs a mathematical, locating, positioning, data type, logical, or character location job	The 23 Essential Excel Functions	Page Link
17	Granularity	The level of detail defined by selected dimensions and measures	In the superstore data table, two subcategory lines under one category with 10 in profit alone is more granular than the category level of detail with 20 in profit combining the two	Page Link
18	Operator	Symbols that indicate mathematical treatment of numbers	+, - , ÷, x	No reference
19	Primary Key	Another word for Unique ID	CA-2016-152156Claire GuteFUR-BO-10001798 in the super store data table	Excel workbook cell B2
20	Qualitative	Characterized by non-numerical description	The Category column in the super store data table	Page Link

21	Quantitative	Characterizes by numerical description	The Sales column in the super store data table	Page Link
22	Row	A horizontal element in a data table that records one unique entry describes by each of the fields	Row ID 1 in the super store data table	Page Link
23	Unique ID	A non-repeating entry that independently defines a row in a data table	CA-2016-152156Claire GuteFUR-BO-10001798 in the super store data table	Excel workbook cell B2

Example Analysis: Bringing it All Together

In this section, we'll talk about setting up your workbook for analysis in Excel, including best practices on structuring worksheets and font colors.

Then we'll perform an analysis answering the question: **which states have the highest profit?**

Setting Up the Workbook

Structuring your Excel workbook so that it is easy to understand is the hallmark of a strong, collaborative data analyst. It's important to follow a set of guidelines that are generally accepted among professionals to ensure (1) that others can review your work, and (2) that you can understand when revisiting the workbook later.

In general:

- Data sources should be held on independent worksheets.
- Work with outputs that constitute significant steps in the logic of the analysis should be held on independent worksheets.
- A conclusive summary page should be in the first tab, ALWAYS.
- The only hard-coded numbers should come from the source worksheets—every other cell should be a formula linking to those worksheets. The exception to this rule is single-cell inputs from a different source used only in one step of the analysis, such as the conversion rate between two currencies on a specific date.
- Where hard-coded numbers are required in a *presentation* page, they should be colored neon blue.
- Formula cells should be colored **black.**
- The gridlines should always be removed for presentation pages (View>Gridlines checkbox).

Data tables stored in an Excel worksheet should always:

- Begin in column A1.
- Be saved as a range, not a data table (Table > Convert to Range).
- Have the drop-down filter functionality active (Data > Filter).
- Be alone on a worksheet.

	A	B	C	D	E	F	G
1	Row ID	Order ID	Order Dat	Ship Date	Ship Mod	Customer	Customer
2	1	CA-2016-152	11/8/16	11/11/16	Second Class	CG-12520	Claire Gute
3	2	CA-2016-152	11/8/16	11/11/16	Second Class	CG-12520	Claire Gute
4	3	CA-2016-138	6/12/16	6/16/16	Second Class	DV-13045	Darrin Van H
5				10/18/			O'Donn
6				10/18/13			Sean O'Donn
7	6	CA-2014-115	6/9/14	6/14/14	Standard Cla	BH-11710	Brosina Hoff
8	7	CA-2014-115	6/9/14	6/14/14	Standard Cla	BH-11710	Brosina Hoff
9	8	CA-2014-115	6/9/14	6/14/14	Standard Cla	BH-11710	Brosina Hoff
10	9	CA-2014-115	6/9/14	6/14/14	Standard Cla	BH-11710	Brosina Hoff
11	10	CA-2014-115	6/9/14	6/14/14	Standard Cla	BH-11710	Brosina Hoff
12	11	CA-2014-115	6/9/14	6/14/14	Standard Cla	BH-11710	Brosina Hoff
13	12	CA-2014-115	6/9/14	6/14/14	Standard Cla	BH-11710	Brosina Hoff
14	13	CA-2017-114	4/15/17	4/20/17	Standard Cla	AA-10480	Andrew Alle
15	14	CA-2016-161	12/5/16	12/10/16	Standard Cla	IM-15070	Irene Maddo

Calculation sheets and pivot tables should always:

- Begin in cell B3 (down two and right 1 of cell A1) for presentation purposes.
- Where hard-coded numbers are required in a presentation page, they should be colored neon blue.
- Formula cells should be colored **black.**

Example Analysis

Which states have the highest profit?

To answer this question, we need to identify values at a specific level of detail for a single dimension. We also need to determine how many states we will show, which will be based on the findings. Essentially, this is simply a question of *sorting*.

Steps:

1. Pivot the data to isolate the State dimension in the Row field and place the profit measure in the Values field with a SUM aggregation
2. Sort from highest to lowest
3. Compare the rankings over time by placing the Order date in the column field
4. Determine how many states to isolate based on their % contribution to the total
5. Create a line graph
6. Revise as needed to best answer the question, remembering that we always want to compare horizontally (over time) and vertically (among the items themselves)

First I've pivoted the data source to get an idea of the data in question.

Row Labels	Sum of Profit2	Sum of Profit	Sum of Profit2	Sum of Profit	Sum of Profit2	Sum of Profit	Sum of Profit2	Sum of Profit	Total Sum of Profit2	Total Sum of Profit
	Column Labels									
	2014		2015		2016		2017			
California	12,637.95	25.50%	14,371.26	23.36%	20,005.72	24.63%	29,366.46	31.65%	76,381.39	26.79%
New York	13,748.94	27.74%	19,277.58	31.34%	16,620.32	20.47%	24,357.07	26.25%	74,003.92	25.96%
Washington	6,607.28	13.33%	5,328.72	8.66%	4,209.88	5.18%	17,256.78	18.60%	33,402.65	11.72%
Michigan	1,817.11	3.67%	5,165.80	8.40%	8,992.52	11.07%	8,487.76	9.15%	24,463.19	8.58%
Indiana	868.15	1.75%	1,990.13	3.24%	10,385.13	12.79%	5,139.53	5.54%	18,382.94	6.45%
Virginia	5,954.67	12.01%	2,985.79	4.85%	7,602.32	9.36%	1,806.01	1.95%	18,348.80	6.44%
Georgia	1,491.71	3.01%	4,763.91	7.74%	3,546.43	4.37%	6,447.98	6.95%	16,250.04	5.70%
Kentucky	2,309.40	4.66%	2,818.35	4.58%	1,216.66	1.50%	4,751.72	5.12%	11,096.13	3.89%
Minnesota	6,237.36	12.58%	1,721.48	2.80%	404.47	0.50%	2,459.88	2.65%	10,823.19	3.80%
Delaware	1,338.59	2.70%	1,608.87	2.62%	976.71	1.20%	6,053.20	6.52%	9,977.37	3.50%

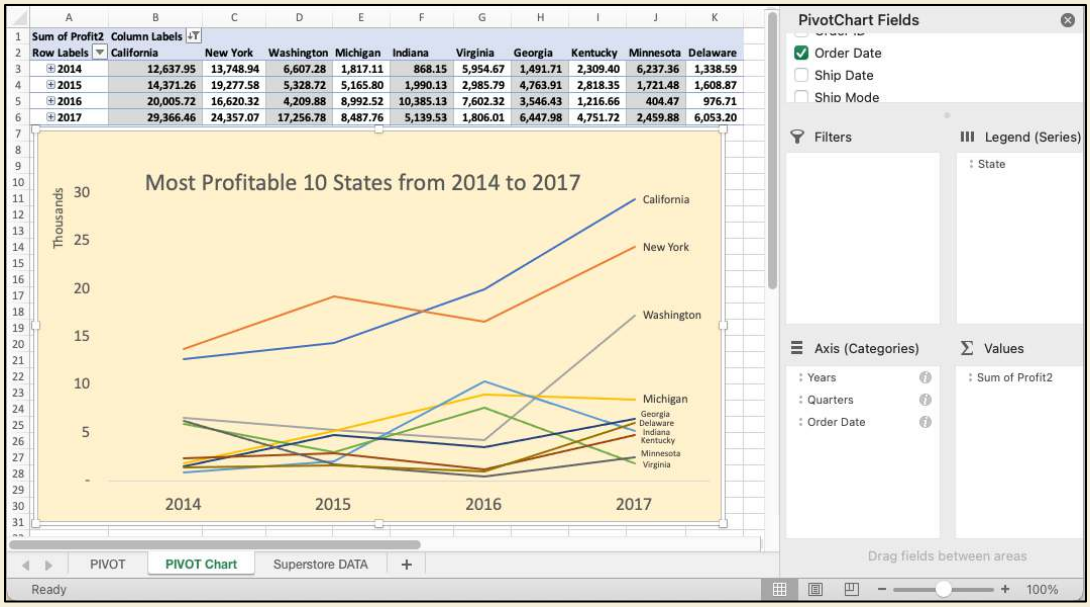
(If you're wondering how I retrieved the percent of total, it's a functionality in Excel Pivot Table. Right click the Sum of Profit Measure in the Values field and select Field Settings. From there, navigate to Show Data As > Drop Down > % of Column Total. To add the sum of profit field again, simply drag it to the Values field again.)

As you can see, over four years more than 60% of the total sales are in California, New York, and Washington! This is already an insight. It seems like a reasonable response to the question "which states have the highest profit?" can be answered with the top 10 states.

Now we need to address the element of time on the answer. We can see in the image above that while the total of Indiana is higher than the total of Virginia, this was not the case in all years. To communicate this, we can simply show the first-year ranking and the last year ranking. Then, we can use a line graph to show this evolution.

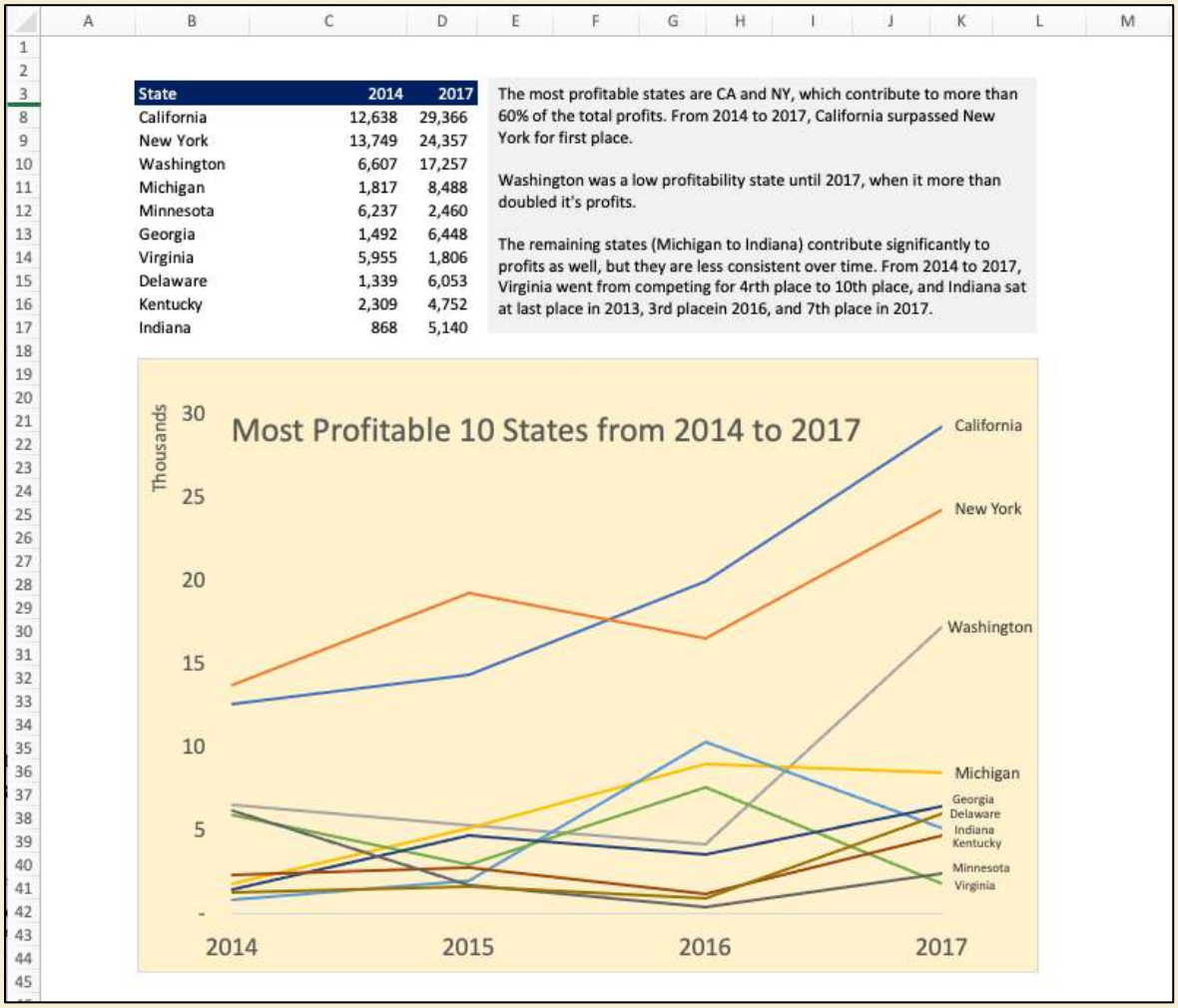
To do this, filter the top 10 states by highlighting them and right-clicking > Keep Selected Items (you can filter for them as well). I then remove the percent of total value from the value field and remove the grand totals by highlighting them and right-clicking > remove grand total. Then I select a cell in the PivotTable and navigate to Insert > Line > Line. PivotTables automatically place the rows on the x-axis, so we need to switch them by right-clicking the chart > Select Chart Data… > Switch Row/Column. This will change our PivotTable as well.

This is what the final output looks like after some formatting (we won't look at formatting here because it's a huge topic and outside the scope of this book).



Notice that I have stored the superstore data in one workbook, the original pivot in a separate workbook, and the chart editions in a third workbook. This is the proper structure because it shows significant steps in separate tabs, making it easier to understand for you and anyone who wants to review.

Now we're ready to create the presentation page. We already concluded that we would only take the top 10 states because these represent >60% of the total, and the other 40 states share small portions of the remaining 30%. We also concluded that it's important to show evolution by including the first year and last year values, as well as a chart. Here's what my presentation page looks like:



State	2014	2017
California	12,638	29,366
New York	13,749	24,357
Washington	6,607	17,257
Michigan	1,817	8,488
Minnesota	6,237	2,460
Georgia	1,492	6,448
Virginia	5,955	1,806
Delaware	1,339	6,053
Kentucky	2,309	4,752
Indiana	868	5,140

The most profitable states are CA and NY, which contribute to more than 60% of the total profits. From 2014 to 2017, California surpassed New York for first place.

Washington was a low profitability state until 2017, when it more than doubled it's profits.

The remaining states (Michigan to Indiana) contribute significantly to profits as well, but they are less consistent over time. From 2014 to 2017, Virginia went from competing for 4rth place to 10th place, and Indiana sat at last place in 2013, 3rd placein 2016, and 7th place in 2017.

As you can see, this presentation clearly shows the evolution of the top 10 most profitable states and directly answers the question "which states are most profitable?" while adding insights about horizontal performance (over time). It also includes a textual description of the findings.

Conclusion: How to Become a Master

This introduction to data analysis e-book shows what you need to get started as a data analyst. At its core, data analysis boils down to manipulating unique IDs, applying the 6 aggregation functions, taking averages and related statistical functions, and applying correlations. By now you have seen that PivotTables are a key functionality to execute these techniques as well, and 5 charts are all you need to visualize most results.

But what's next? What's the difference between a beginner and a master?

The obvious element is that masters use more advanced techniques. We saw a near comprehensive list of techniques under the data analysis types section, which include advanced mathematical techniques such as clustering, artificial intelligence techniques such as fuzzy logic, and qualitative analyses such as word coding.

In addition, masters also learn tools that enable them to execute these techniques, just like we used Microsoft Excel to execute simple aggregations and correlations. Some of these tools include coding languages such as Python, R, and SQL, or even VBA within Excel. Other tools include data visualization and management software such as Tableau, PowerBI.

These are tools that you learn when you become comfortable with the contents in this book. These tools and techniques all work with the same basic topics we discussed, but the complexity and syntax changes. That's why it's important to have very strong fundamentals.

You can always go faster

Even if you think to yourself now "I understand these concepts and I'm ready to take the next step" ask yourself how fast you can execute what you know.

Good analysts are fast – really fast. They can churn out analyses with big data tables and answer many questions in a day's time. You should always push yourself to be faster. Consider learning to use Excel without a mouse, and maximize.

Once you can use Excel to answer a variety of questions with minimal thought about where buttons are located or what shortcut to use, then you're ready to take the next step. At www.AnalystAnswers.com, you'll find loads of free content to help you progress, and we'll soon have advanced courses to help you drive even more insights from data.

Practice, practice, practice

As with any skill, data analysis requires practice. To become even faster, you'll need to practice, practice, practice answering questions with data tables. To help you get started, I've included below real-world business case question below to conclude the book.

Practice Business Cases

What regions generate the most sales?

Hint: this question is asking for a level of detail sorting assessment of regions, just like the example analysis above. You can see my analysis in the attached workbook.

Are customers buying higher quantities after their first purchase?

Hint: this question is asking for a comparison in user purchase behavior between their first and subsequent purchases. An intuitive approach would involve finding MIN order date at the level of detail of the member, establishing average quantity size, then identifying average quantity size for all subsequent orders. Be careful, because you'll need to focus only on returning customers.

You can see my analysis in the attached workbook and an explanation in the answers section.

Which segment is the most profitable?

Hint: this question is not asking which segment has the highest overall profit (SUM of profit). It's asking which segment has the highest profit margin – that is, per unit of sales, which one has the highest relative profit. You can see my analysis in the attached workbook.

Answers: Section Quizzes & Business Cases

Section Quiz 1 Answers

1. What is data?
   Data is organized information.
2. What is the basic structure of all data?
   A data table.
3. What is a data table?
   A collection of columns and rows that describe a series of elements (rows) with numerical and non-numerical information (columns).
4. What special trait distinguishes a data table from other collections of columns and rows?
   A data table always has a unique ID, aka primary key, that provides a non-repeating identity to each row.
5. What is an aggregation?
   The combination of information in multiple rows with 1 of the 6 aggregation functions (SUM, etc.), wherein qualitative information is always counted and quantitative data can be aggregated with any of the aggregation functions.
6. What is raw data?
   An unaggregated form of a data table that shows the information as it was collected, making it the most detailed, or granular, information available.
7. What are the six aggregation functions?
   SUM, AVERAGE, MIN, MAX, COUNT, and COUNT DISTINCT.
8. What is the difference between a data set and a data table?
   A data set is a collection of one or more data tables or other data objects, whereas a data table is only a collection of rows and column. When a data set includes only one data table, the terms are interchangeable.

Section Quiz 2 Answers

1. What is data analysis?
   Critically thinking about organized information to answer targeted, unambiguous, and unbiased questions by manipulating unique IDs, applying the 6 aggregation functions, taking averages and related statistical functions, and applying correlations.
2. What is the starting point of any analysis? And what are its three key components?
   A good question that is targeted, unambiguous, and unbiased.
3. What is the data cycle? How does analysis fit into it?
   The data cycle includes a question, data collection, data cleaning, data processing, **data analysis**, data interpretation, and an answer. Data analysis is performed after data processing.
4. What are data types, methods, and techniques?
   Data types are categories such a quantitative and qualitative. Methods are additional categories that show how each data type breaks down, such as classification and forecasting. Techniques are the practical applications of those methods such as cluster analysis, regressions, and moving averages.
5. What four techniques constitute the job of a data analyst?
   Manipulating unique IDs, applying the 6 aggregation functions, applying averages, and applying correlations.

Section Quiz 3 Answers

1. What function returns the largest value in an array?
   =MAX()
2. What function returns a specified value when an error message occurs on the base calculation?
   =IFERROR()
3. What function computes the sum of an array based on criteria in the same or another array?
   =SUMIF()
4. What function combines the values of multiple cells into one?

=CONCATENATE()

5. What function returns the number of characters in a cell?
   =LEN()
6. What function returns the position of a cell within an array?
   =MATCH()
7. What combination of functions returns the value of a cell at a specified row and column location in an array, where the location coordinates are the result of a nested function?
   INDEX+MATCH
8. What do PivotTables do?
   PivotTables create grouped values that aggregate the individual items of a more extensive table within one or more discrete (non-repeating) categories.

Section Quiz 4 Answers

5. In simple terms, what does an average tell us about a data set?
   It tells us what the central point of the set is.
6. What is variance, and what is its key limitation?
   Variance tells us what the average distance is of each point to the average in the series. To keep the distances positive, it squares the distances, which means that it's key limitation is difficulty comparing the result to the magnitude of values in the series.
7. What is standard deviation?
   Standard deviation is the square root of variance, which allows us to compare the average distance of each point in the magnitude of the values in the series.
8. When should you use variance, and when should you use standard deviation?
   Use variance to compare two separate data series and use standard deviation to understand a single data series.

Section Quiz 5 Answers

7. What is covariance?
   Covariance is a measure of the distance from the mean of two variables (series) in their respective sets that tells us what the direction and strength of the relationship is between those two variables.
8. What makes covariance different from variance?
   Covariance looks at two variables, whereas variance looks at a single variable. Covariance therefore tells us the direction and strength of a relationship, while variance tells us the "spread" within a single variable.
9. What is a correlation?
   A benchmarked measure of the strength and direction of a relationship of two variables from -1 to 1, where >0.7 is considered strong, 0.5<>0.7 is medium, and <0.5 is weak.
10. Can correlations be negative?
    Yes.
11. What is different about the correlation formula compared to variance, standard deviation, and covariance?
    All of these formulas are fractions. The denominator in the correlation equation is values from the series themselves, whereas the denominator in the other equations is a count of the values in the series.
12. What are correlations useful for?
    They're useful for determining variables that have strong relationships with a benchmark.

Section Quiz 6 Answers

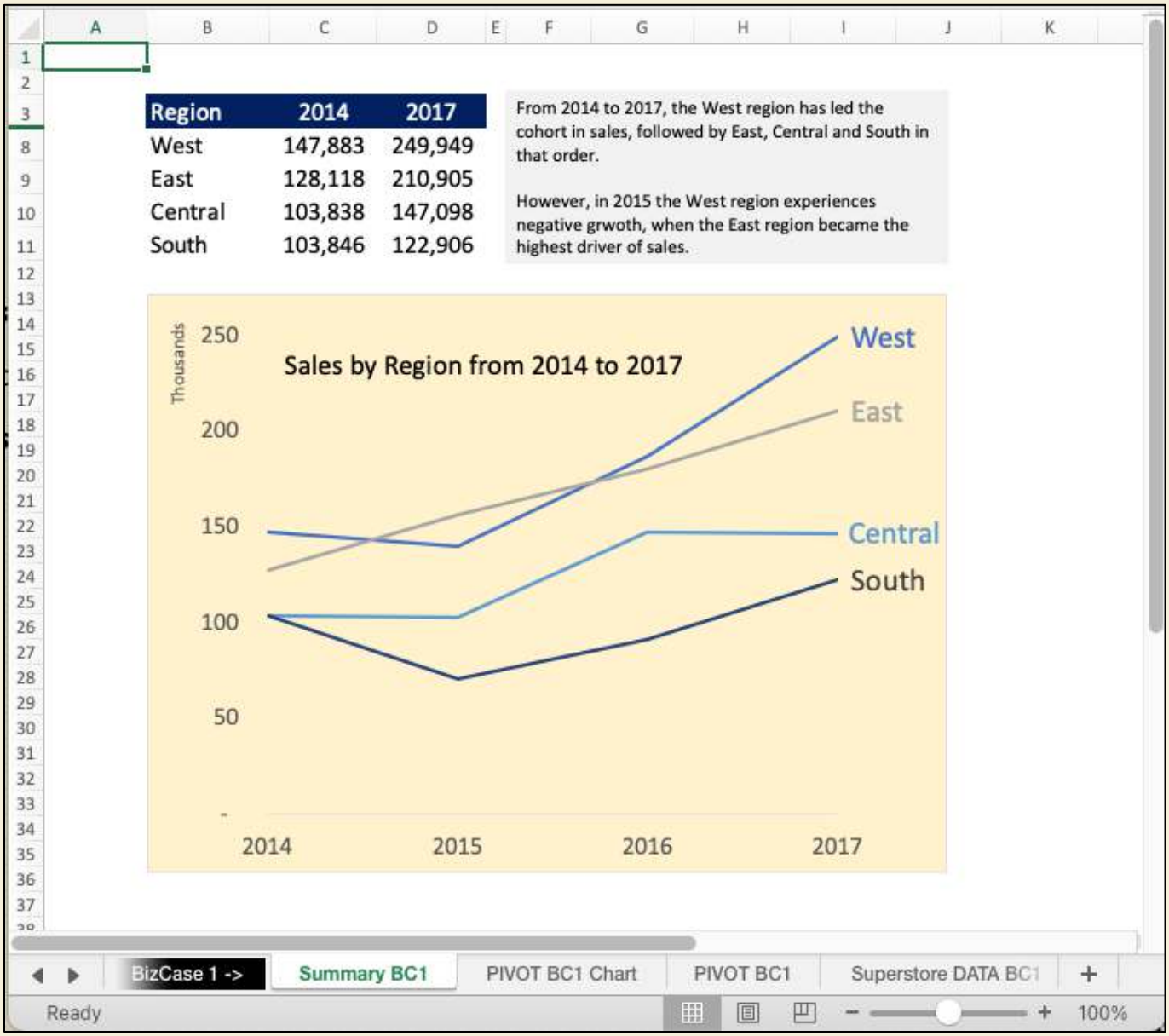
6. Which chart would you use to show change in values for one dimension over time?
   Line graph.
7. Which chart would you use to show the relationship between two variables?
   Scatter Plot.
8. Which chart would you use to compare the values of several instances of a dimension?
   Bar or column chart.
9. Which chart would you use to visually emphasize the magnitude of change over time?

Area chart.

10. Which chart would you use to show the additions and reductions in a balance over time?
    Waterfall chart.

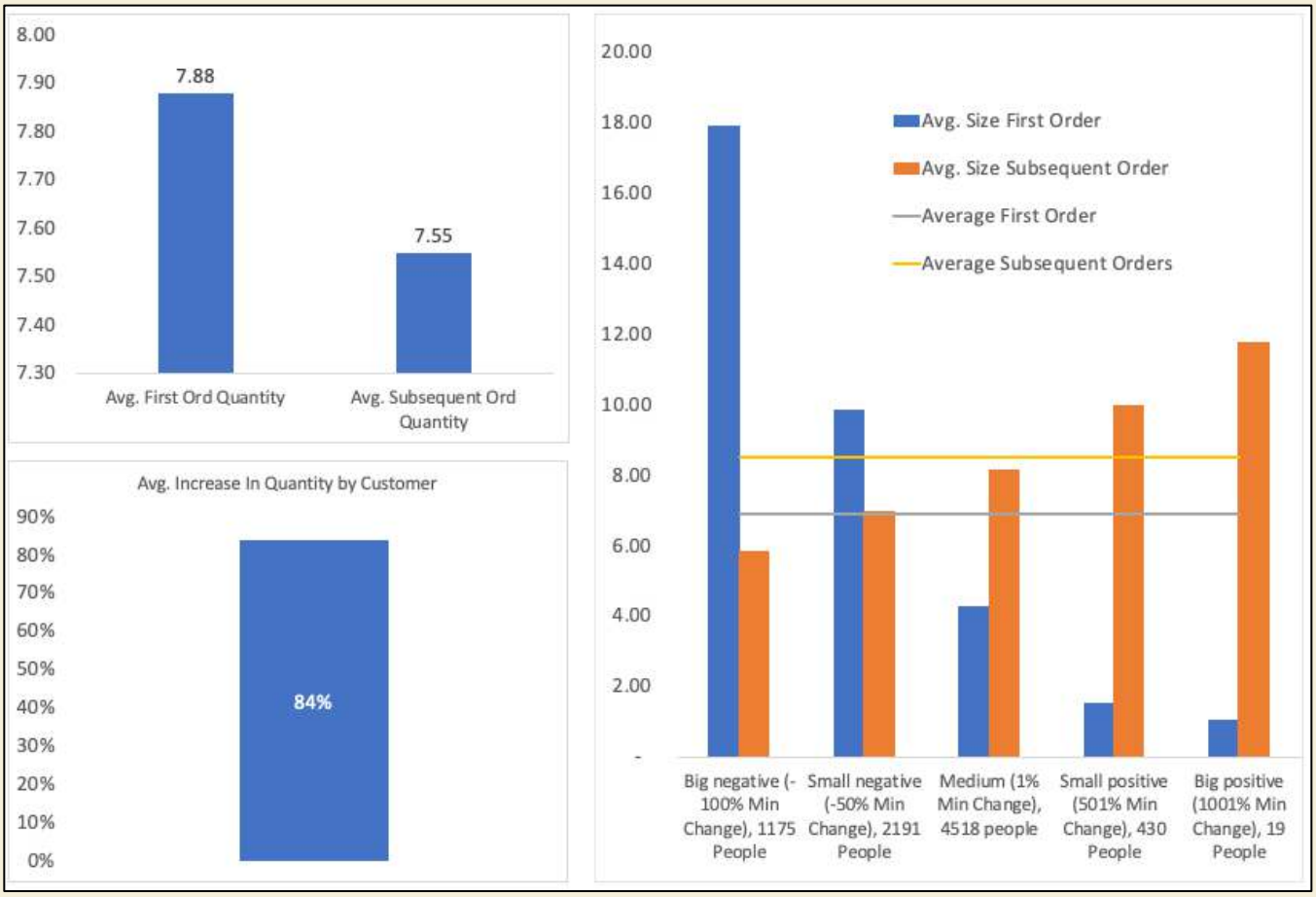
Answer: What regions generate the most sales?

(Analysis available in Excel workbook)



Answer: Are customers buying higher quantities after their first purchase?

(Analysis available in Excel workbook)



This analysis is a bit tricky and should be reserved for a time when you're more comfortable with the concepts of this book. But let's tackle it to get an idea what advanced cases look like. The answer to the question is yes, customers are buying higher quantities after their first purchase on average by roughly 84% (bottom left graph).

That said, the average quantity purchased per order is substantially the same (7.55 on subsequent orders versus 7.88 on first orders, upper left graph).

The reason why the average increase in quantity is high at 84% while the average overall order size remains the same is a question of starting point. There were many customers, 4,518 to be exact, whose % increase in average quantity was between 1% and 500%. You can see this in the middle column in the right-side graph. However, their first order quantities were 4 and subsequent order quantities were closer to the average at 7.9.

In other words, they were starting from a lower position and regressed to the mean.

Likewise, there were 1,175 customers whose change was at least -100% in size, and 2,191 customers whose change was at least -50. This means they were fewer than the 4,518 customers who % increased, so the average % change favored the latter. However, the customers who bought smaller quantities regressed to the mean as well, so the average quantities purchased remained consistent.

What does this tell us? It's insightful because there is potential to target reductions and promotions to the customers who grow (avg. first purchases of 4, 2, and 1), creating customer loyalty and increasing the chance they will purchase greater quantities in later periods.

Answer: Which segment is the most profitable?

(Analysis available in Excel workbook)