**Task 1: Data Aggregation and Analysis**

Task Overview

**What you'll learn**

* Understand the importance of data engineering and analysis in a long-term data-focused vision.
* Gain experience in using real-world transactional data for analysis.
* Learn to perform data analysis and answer specific questions based on the dataset provided.

**What you'll do**

* Analyze the "CSV Data Set" of supermarket transactions to address key questions.
* Document the analysis process, including formulas used, using spreadsheet software.
* Submit the analyzed file for review.
* Here is the background information on your task
* CBA is embarking on a long-term vision with trusted partner InsightSpark, a data science specialist.
* This vision is focused around using the huge volumes of CBA’s transactional data, open-source data, and advanced data science to build a platform that will provide insights for businesses, governments, and investors in Australia.
* The data engineering team is responsible for building data engineering pipelines, running analysis on big data, and deploying algorithms into production at scale. This project’s success lies heavily on the data engineering team.

**Task 1: Data Aggregation and Analysis**

Here is your task

Your first task is to use the provided **“CSV Data Set”**(you can find it in the Resources) of transactions at a supermarket and analyse the data to answer some questions. We suggest you complete this task in Excel or an open-source alternative like LibreOffice Calc.

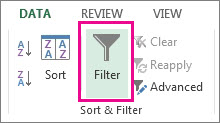
Task details and the answers you need to find are listed below in an email from the team lead:

|  |
| --- |
| *Hello,*  *I have provided you with a data set named “supermarket\_transactions.csv.” This data set contains three years of transactional data we collected from supermarkets across Australia.*  *InsightSpark wants to access our data to perform some analysis across Australian supermarkets. Before this can happen, I need you to perform some analysis on the provided CSV data set (a sample of the database) to answer the following questions:*   * *Across locations, how many apples were purchased in cash?* * *How much total cash was spent on these apples?* * *Across all payment methods, how much money was spent at the Bakershire store location by non-member customers?* |

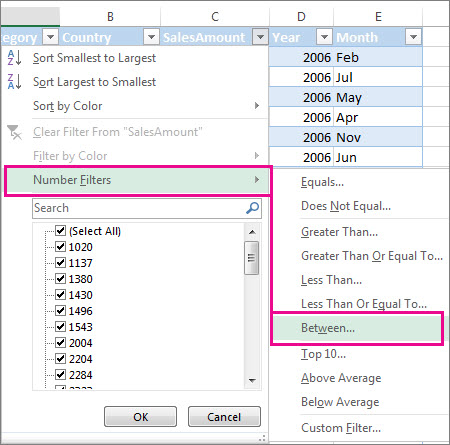
You should conduct your analysis in Excel or a spreadsheet program of your choice. In your file, document the formulas used to generate your values. When finished, please submit the analysed file for me to review before we share our findings with InsightSpark.

Filter a range of data

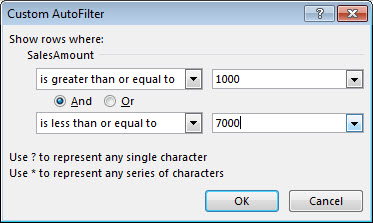
1. Select any cell within the range.
2. Select **Data**> **Filter**.



1. Select the column header arrow Filter arrow.
2. Select **Text Filters** or**Number Filters,**and then select a comparison, like **Between**.

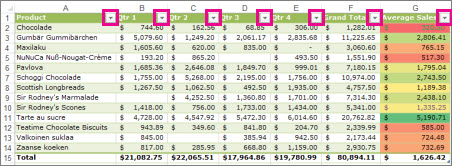


1. Enter the filter criteria and select **OK**.

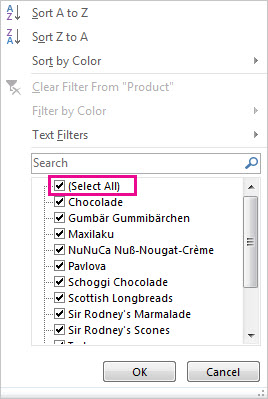


Filter data in a table

When you [put your data in a table](https://support.microsoft.com/en-us/office/create-and-format-tables-e81aa349-b006-4f8a-9806-5af9df0ac664), filter controls are automatically added to the table headers.



1. Select the column header arrow Filter drop-down arrow for the column you want to filter.
2. Uncheck **(Select All)** and select the boxes you want to show.



1. Select **OK**.

The column header arrow Filter drop-down arrow changes to a Applied filter icon**Filter** icon. Select this icon to change or clear the filter.

Related Topics

[Excel Training: Filter data in a table](https://support.microsoft.com/en-us/office/filter-data-in-a-range-or-table-7fbe34f4-8382-431d-942e-41e9a88f6a96)

[Guidelines and examples for sorting and filtering data by color](https://support.microsoft.com/en-us/office/guidelines-and-examples-for-sorting-and-filtering-data-by-color-b1bf3982-051d-49b8-8330-80e99c94365b)

[Filter data in a PivotTable](https://support.microsoft.com/en-us/office/filter-data-in-a-pivottable-cc1ed287-3a97-4e95-b377-ddfafe79fa8f)

[Filter by using advanced criteria](https://support.microsoft.com/en-us/office/filter-by-using-advanced-criteria-4c9222fe-8529-4cd7-a898-3f16abdff32b)

[Remove a filter](https://support.microsoft.com/en-us/office/clear-or-remove-a-filter-f147ac85-4fd6-4143-b52d-ce153250fed3)

Quick start: Sort data in an Excel worksheet

*Excel for Microsoft 365 Excel 2021 Excel 2019 Excel 2016*

When sorting information in a worksheet, you can rearrange the data to find values quickly. You can sort a range or table of data on one or more columns of data. For example, you can sort employees —first by department, and then by last name.

How to sort in Excel?

|  |  |
| --- | --- |
| Icon image | **Select the data to sort**  Select a range of tabular data, such as A1:L5 (multiple rows and columns) or C1:C80 (a single column). The range can include the first row of headings that identify each column.  Example of selected data to sort in Excel |

|  |  |
| --- | --- |
| Icon image | **Sort quickly and easily**   1. Select a single cell in the column you want to sort. 2. On the **Data** tab, in the **Sort & Filter** group, click A to Z command in Excel that sorts A to Z or smallest number to largest to perform an ascending sort (from A to Z, or smallest number to largest).   Sort buttons in the Sort & Filter group on the Data tab in Excel   1. Click Z to A command in Excel that sorts Z to A or largest number to smallest to perform a descending sort (from Z to A, or largest number to smallest). |

|  |  |
| --- | --- |
|  | **Sort by specifying criteria**  Use this technique to choose the column you want to sort, together with other criteria such as font or cell colors.   1. Select a single cell anywhere in the range that you want to sort. 2. On the **Data** tab, in the **Sort & Filter** group, click **Sort** to display the Sort popup window.   Sort command in the Sort & Filter group on the Data tab in Excel   1. In the **Sort by** dropdown list, select the first column on which you want to sort. 2. In the **Sort On** list, choose **Values**, **Cell Color**, **Font Color**, or **Cell Icon**. 3. In the **Order** list, choose the order that you want to apply to the sort operation—alphabetically or numerically, ascending or descending (that is, from A to Z (or Z to A) for text, or lower to higher, or higher to lower for numbers). |

Example Answer Explanation

The example answer is just one way to approach this problem, not the only solution. For this question, first, we opened the CSV file `supermarket\_transactions.csv` in Excel. Then we aggregated and analysed the data using filters and formulas to answer the following questions:

**Question 1: Across locations, how many apples were purchased in cash?**

To answer this question, we filtered the data sheet to include only rows where the  
`product\_name` is “apple” and where the `payment\_method` is “cash.” Then, we summed the  
`quantity` column to get an answer of **117 apples**.

**Question 2: How much total cash was spent on these apples?**

Here, we left the data sheet filtered to include only rows where the `product\_name` is “apple” and where the `payment\_method` is “cash.” Then, we summed the `total\_amount` column to get an answer of **$537.03**.

**Question 3: Across all payment methods, how much money was spent at the Bakershire store location by non-member customers?**

For this question, we cleared all the previous filter criteria. Then, we filtered the data sheet to include only rows where the `customer\_type` was non-member, and the `store` was Bakershire. Then, we summed the `total\_amount` column to get an answer of **$2,857.51**.

**Task 2: Data Anonymisation**

Task Overview

**What you'll learn**

* Understand the importance of data anonymization in protecting sensitive customer information.
* Learn various techniques for anonymizing data while preserving its utility.
* Comprehend the risks associated with linkage attacks on anonymized data.

**What you'll do**

* Anonymize the provided data set, ensuring the protection of personal details.
* Utilize anonymization techniques such as removing unnecessary columns, masking identifying information, and categorizing sensitive figures.
* Submit the anonymized data as a CSV file.

# Here is the background information on your task

* For CBA and InsightSpark to continue with the data science vision, the data scientists at InsightSpark need to combine customer data with the transactional data from the supermarkets.
* The customer data will come from the CBA mobile app. However, we cannot provide them with the raw data until it has been suitably anonymised to protect the identity of individuals in the data set.
* It is your job to design a privacy pipeline to execute on the data set. This pipeline must anonynmise the data set while preserving useful information.
* Data privacy is crucial in preventing linkage attacks. This attack is a way of recovering personal information about people from an anonymised data set. It’s a powerful technique that can be executed with even the smallest personal information. Review the following [resource](https://www.schneier.com/blog/archives/2007/12/anonymity_and_t_2.html) for more information about linkage attacks.
* Here is your task
* This task focuses on anonymising a data set of sensitive customer information. You must preserve the utility and valuable information within the dataset while anonymising any personal or sensitive information. You can complete this exercise with any tool you wish, but we recommend using Excel or a spreadsheet program of your choice. You only need to submit the anonymised data set as a CSV file for your deliverable. Should you need help, visit the Data Anonymisation links in the Resources section.
* Further details about this task are given below in an email from the team lead:

|  |
| --- |
| *Hello,*  *I have provided you with a data set named “****mobile\_customers.csv****.” This contains information about customers that have signed up for the mobile app in the last three years.*  *We need you to anonymise this data to hide personal details while preserving any useful information for the data scientists at InsightSpark.*  *Here are some examples of how you may anonymise a data set:* *You could remove columns that don’t provide helpful information for analysis (e.g., names or credit card numbers).* *You could mask any columns that can identify an individual (e.g., passport numbers or mobile numbers).* *You could categorise personal figures (e.g., age and income) into a bracket rather than a specific number.*  *First, research the different techniques available for anonymising a data set. Then, edit the data set to create an anonymised data set as a CSV file. When finished, please submit this CSV file for me to review before we share it with InsightSpark.* |

* Once you’ve completed your anonymisation, submit your CSV file below.

# Overview of formulas in Excel

**Important:**The calculated results of formulas and some Excel worksheet functions may differ slightly between a Windows PC using x86 or x86-64 architecture and a Windows RT PC using ARM architecture. [Learn more about the differences](https://support.microsoft.com/en-us/office/calculation-differences-between-pc-and-windows-rt-computers-858c27e6-7843-4a8c-a5aa-d7b1cbd021d5).

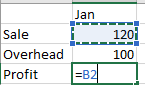
**Important:**In this article we discuss XLOOKUP and VLOOKUP, which are similar. Try using the new [XLOOKUP](https://support.microsoft.com/en-us/office/xlookup-function-b7fd680e-6d10-43e6-84f9-88eae8bf5929) function, an improved version of VLOOKUP that works in any direction and returns exact matches by default, making it easier and more convenient to use than its predecessor.

Create a formula that refers to values in other cells

1. Select a cell.
2. Type the equal sign =.

**Note:**Formulas in Excel always begin with the equal sign.

1. Select a cell or type its address in the selected cell.



1. Enter an operator. For example, – for subtraction.
2. Select the next cell, or type its address in the selected cell.



1. Press Enter. The result of the calculation appears in the cell with the formula.

See a formula

When a formula is entered into a cell, it also appears in the **Formula bar.**

Formula Bar

* To see a formula in the formula bar, select a cell.

See formula bar

Enter a formula that contains a built-in function

1. Select an empty cell.
2. Type an equal sign = and then type a function. For example, =SUM for getting the total sales.
3. Type an opening parenthesis (.
4. Select the range of cells, and then type a closing parenthesis).

range

1. Press Enter to get the result.

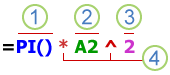
Formulas in-depth

You can browse through the individual sections below to learn more about specific formula elements.

**The parts of an excel formula**

A formula can also contain any or all of the following: **functions**, **references**, **operators**, and **constants**.

**Parts of a formula**



1. **Functions**: The [PI() function](https://support.microsoft.com/en-us/office/pi-function-264199d0-a3ba-46b8-975a-c4a04608989b) returns the value of pi: 3.142...

2. **References**: A2 returns the value in cell A2.

3. **Constants**: Numbers or text values entered directly into a formula, such as 2.

4. **Operators**: The ^ (caret) operator raises a number to a power, and the \* (asterisk) operator multiplies numbers.

**Using Constraints In Excel Formulas**

A constant is a value that is not calculated; it always stays the same. For example, the date 10/9/2008, the number 210, and the text "Quarterly Earnings" are all constants. An expression or a value resulting from an expression is not a constant. If you use constants in a formula instead of references to cells (for example, =30+70+110), the result changes only if you modify the formula. In general, it's best to place constants in individual cells where they can be easily changed if needed, then reference those cells in formulas.

**Using References In Excel Formulas**

A reference identifies a cell or a range of cells on a worksheet, and tells Excel where to look for the values or data you want to use in a formula. You can use references to use data contained in different parts of a worksheet in one formula or use the value from one cell in several formulas. You can also refer to cells on other sheets in the same workbook, and to other workbooks. References to cells in other workbooks are called links or external references.

* **The A1 reference style**

By default, Excel uses the A1 reference style, which refers to columns with letters (A through XFD, for a total of 16,384 columns) and refers to rows with numbers (1 through 1,048,576). These letters and numbers are called row and column headings. To refer to a cell, enter the column letter followed by the row number. For example, B2 refers to the cell at the intersection of column B and row 2.

| **To refer to** | **Use** |
| --- | --- |
| The cell in column A and row 10 | A10 |
| The range of cells in column A and rows 10 through 20 | A10:A20 |
| The range of cells in row 15 and columns B through E | B15:E15 |
| All cells in row 5 | 5:5 |
| All cells in rows 5 through 10 | 5:10 |
| All cells in column H | H:H |
| All cells in columns H through J | H:J |
| The range of cells in columns A through E and rows 10 through 20 | A10:E20 |

* **Making a reference to a cell or a range of cells on another worksheet in the same workbook**

In the following example, the [AVERAGE function](https://support.microsoft.com/en-us/office/average-function-047bac88-d466-426c-a32b-8f33eb960cf6) calculates the average value for the range B1:B10 on the worksheet named Marketing in the same workbook.



1. Refers to the worksheet named Marketing

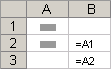
2. Refers to the range of cells from B1 to B10

3. The exclamation point (!) separates the worksheet reference from the cell range reference

**Note:**If the referenced worksheet has spaces or numbers in it, you need to add apostrophes (') before and after the worksheet name, like ='123'!A1 or ='January Revenue'!A1.

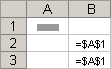
* **The difference between absolute, relative and mixed references**
  1. **Relative references**    A relative cell reference in a formula, such as A1, is based on the relative position of the cell that contains the formula and the cell the reference refers to. If the position of the cell that contains the formula changes, the reference is changed. If you copy or fill the formula across rows or down columns, the reference automatically adjusts. By default, new formulas use relative references. For example, if you copy or fill a relative reference in cell B2 to cell B3, it automatically adjusts from =A1 to =A2.

**Copied formula with relative reference**



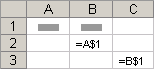
* 1. **Absolute references**    An absolute cell reference in a formula, such as $A$1, always refer to a cell in a specific location. If the position of the cell that contains the formula changes, the absolute reference remains the same. If you copy or fill the formula across rows or down columns, the absolute reference does not adjust. By default, new formulas use relative references, so you may need to switch them to absolute references. For example, if you copy or fill an absolute reference in cell B2 to cell B3, it stays the same in both cells: =$A$1.

**Copied formula with absolute reference**



* 1. **Mixed references**    A mixed reference has either an absolute column and relative row, or absolute row and relative column. An absolute column reference takes the form $A1, $B1, and so on. An absolute row reference takes the form A$1, B$1, and so on. If the position of the cell that contains the formula changes, the relative reference is changed, and the absolute reference does not change. If you copy or fill the formula across rows or down columns, the relative reference automatically adjusts, and the absolute reference does not adjust. For example, if you copy or fill a mixed reference from cell A2 to B3, it adjusts from =A$1 to =B$1.

**Copied formula with mixed reference**



* **The 3-D reference style**

**Conveniently referencing multiple worksheets**    If you want to analyze data in the same cell or range of cells on multiple worksheets within a workbook, use a 3-D reference. A 3-D reference includes the cell or range reference, preceded by a range of worksheet names. Excel uses any worksheets stored between the starting and ending names of the reference. For example, =SUM(Sheet2:Sheet13!B5) adds all the values contained in cell B5 on all the worksheets between and including Sheet 2 and Sheet 13.

* 1. You can use 3-D references to refer to cells on other sheets, to define names, and to create formulas by using the following functions: SUM, AVERAGE, AVERAGEA, COUNT, COUNTA, MAX, MAXA, MIN, MINA, PRODUCT, STDEV.P, STDEV.S, STDEVA, STDEVPA, VAR.P, VAR.S, VARA, and VARPA.
  2. 3-D references cannot be used in array formulas.
  3. 3-D references cannot be used with the intersection operator (a single space) or in formulas that use implicit intersection.

**What occurs when you move, copy, insert, or delete worksheets**    The following examples explain what happens when you move, copy, insert, or delete worksheets that are included in a 3-D reference. The examples use the formula =SUM(Sheet2:Sheet6!A2:A5) to add cells A2 through A5 on worksheets 2 through 6.

* 1. **Insert or copy**    If you insert or copy sheets between Sheet2 and Sheet6 (the endpoints in this example), Excel includes all values in cells A2 through A5 from the added sheets in the calculations.
  2. **Delete**     If you delete sheets between Sheet2 and Sheet6, Excel removes their values from the calculation.
  3. **Move**    If you move sheets from between Sheet2 and Sheet6 to a location outside the referenced sheet range, Excel removes their values from the calculation.
  4. **Move an endpoint**    If you move Sheet2 or Sheet6 to another location in the same workbook, Excel adjusts the calculation to accommodate the new range of sheets between them.
  5. **Delete an endpoint**    If you delete Sheet2 or Sheet6, Excel adjusts the calculation to accommodate the range of sheets between them.
* **The R1C1 reference style**

You can also use a reference style where both the rows and the columns on the worksheet are numbered. The R1C1 reference style is useful for computing row and column positions in macros. In the R1C1 style, Excel indicates the location of a cell with an "R" followed by a row number and a "C" followed by a column number.

| **Reference** | **Meaning** |
| --- | --- |
| R[-2]C | A relative reference to the cell two rows up and in the same column |
| R[2]C[2] | A relative reference to the cell two rows down and two columns to the right |
| R2C2 | An absolute reference to the cell in the second row and in the second column |
| R[-1] | A relative reference to the entire row above the active cell |
| R | An absolute reference to the current row |

When you record a macro, Excel records some commands by using the R1C1 reference style. For example, if you record a command, such as selecting the **AutoSum** button to insert a formula that adds a range of cells, Excel records the formula by using R1C1 style, not A1 style, references.

You can turn the R1C1 reference style on or off by setting or clearing the **R1C1 reference style**check box under the **Working with formulas** section in the **Formulas** category of the **Options** dialog box. To display this dialog box, select the **File** tab.

# An In-Depth Guide to Data Masking

Applications have become more advanced than ever. They're not just used to get or share information. Applications today are capable…

Applications have become more advanced than ever. They’re not just used to get or share information. Applications today are capable of doing almost everything—ordering products, writing an essay for you, completing financial transactions, and much more. To be able to do all this, applications need to deal with sensitive data.

Sensitive data attracts malicious actors who try to steal your data. So, when sensitive data is involved, you need to take care of data security. One very effective yet simple technique is data masking. In this post, we’ll discuss why data masking is important and then see some examples of using it.

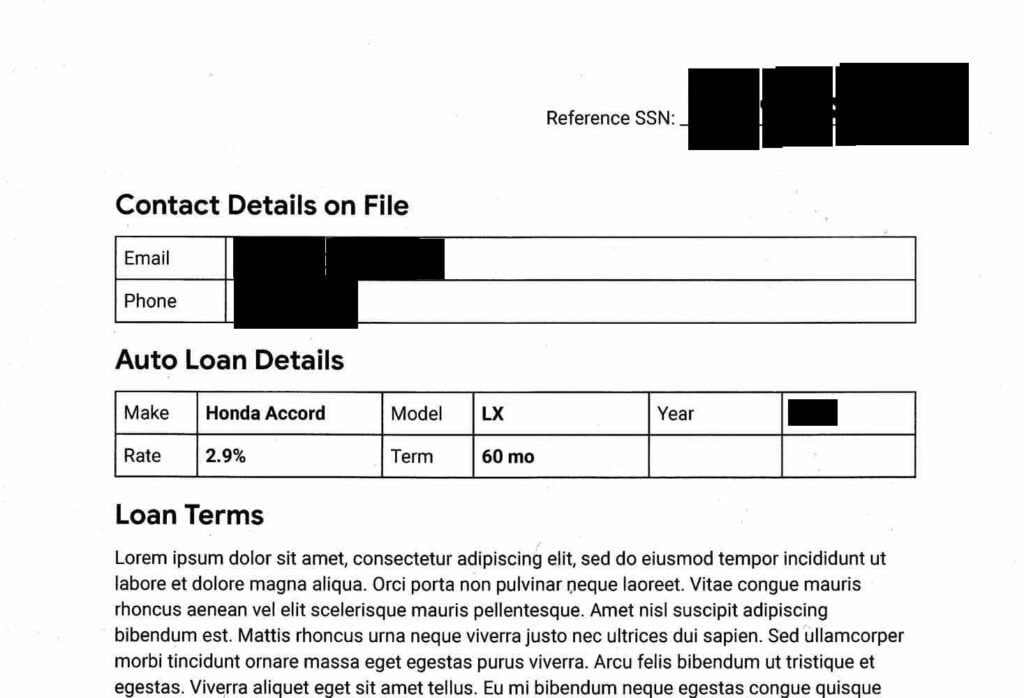
What is Data Masking?

Data masking hides sensitive data by replacing it, or a part of it, with fake data. The idea is to maintain the structure of actual data without providing sensitive data.

For example, many applications store users’ contact numbers. If the user forgets their password to log in, the application will send a one-time password (OTP) to the user’s phone, and the user can then reset their password. To get an OTP, the user provides their username or email address and verifies the number to which the OTP will be sent. But here’s the problem: If user Tony has forgotten his password, he would provide his email address, confirm the phone number, and get OTP. But what if some other user, say Steve, knew Tony’s email address? If Steve used the forgot password option and entered Tony’s email address, then when the app tries to confirm Tony’s phone number, Steve would see the number.

This is a privacy breach. To avoid such situations, we can show a part of the actual data. If the phone number has 10 digits, we can display the last 3 digits and mask the first 7: XXXXXXX452. This way, Tony can verify his phone number, but Steve can’t get it.

Data redaction is another technique for hiding sensitive data. Redaction means blacking out all the sensitive data. Here’s a screenshot from [Google Cloud](https://cloud.google.com/dlp/docs/redacting-sensitive-data-images) that shows what data redaction looks like.



Why is Data Masking Important?

In a word, data masking is important for data security. But let’s dive a little deeper and understand why it’s beneficial.

Insider Threat

And insider threat comes from inside an organization. It can be an employee with malicious intent or an employee who’s not following security practices. Data is stored in databases, and some employees of the company will have access to all data. For example, the customer support people may have access to users’ credit card numbers and SSN. But you need to control what data they can see and what is hidden. Customer service employees might need access to this database to help customers, but it’s not secure to view all of the customers’ data. In such cases, data masking can help secure customers’ data and protect privacy.

Data Breach

If some malicious actor gets access to a system’s database, they could copy the data to a different location and demand a ransom. Using data masks can help avoid such situations because it’s useless if the malicious actor gets hold of masked data. Data masking will not prevent a data breach, but it will limit the effects of a data breach.

Production Database Protection

In this case, the malicious actor need not get access to the database itself. They could get access to data from the client-side of an application or when data is moving from the server to the client. The example of Tony and Steve mentioned above demonstrates how poor practices can lead to illegitimate data access.

Compliance

There are many regulations for data security that require data masking to be compliant. The [General Data Protection Regulation](https://gdpr-info.eu/) (GDPR) is one example created to provide data security. But even if there’s no data incident at your organization, if you aren’t compliant with regulations, you might get into legal issues and end up paying fines. Data masking helps you prevent this situation.

Now that you understand why data masking is important let’s see where we commonly use it.

When Do We Use Data Masking?

We use data masking in two main cases: testing and production.

Test Data Management

To build a robust application, developers should know what kind of data they’re dealing with, its structure, and how the application behaves when dealing with real data. But neither developers nor testers need to see the actual sensitive data. They both can work with data that is similar to the actual data. That’s where data masking comes in. By using masked data, developers and testers can build and test products without worrying about privacy issues. They can also cover all the cases to build and test the product without risking a privacy breach.

Most of us use automated testing tools today because they’re fast and accurate. To use those tools, we don’t need to create a separate database of masked data. Tools like [Testim](https://www.testim.io/) provide all the important features needed for testing applications, along with a custom step to mask sensitive data (black it out) before taking a screenshot. This way, the tool tests the application and generates reports as usual, but the results include no sensitive data.

Applications Using Sensitive Data

While many applications deal with sensitive data, it’s not always necessary for end-users to see it. For example, e-commerce applications can let users save credit card details to prevent entering them every time they purchase something. Once they’ve saved the card details, they don’t need to see the complete details for every transaction. Just seeing part of the details, say the last 4 digits of the card, is enough to identify which card they’re paying with. That’s an instance where data masking can provide protection.

Why do we need to display part of the details and not mask everything? Imagine that a user needs to pay with a specific card and have multiple cards in their profile. If you mask all details, they wouldn’t know which card they’re paying with. But displaying the last 4 digits helps them select or identify a card. Hence, this is a common data masking practice in applications.

Now let’s get to the hands-on part and see some examples of data masking in action.

Data Masking Examples

First, here’s a simple Python example to demonstrate how data masking works and how to implement it in a database.

Here’s the Python code:

users = [{'Name':'Tony', 'Country': 'France', 'Card Number':'3542-7583-7228-3788'},

{'Name':'Steve','Country':'Austria', 'Card Number':'3881-8829-5554-4875'},

{'Name':'Peter', 'Country':'Spain', 'Card Number':'8445-5556-9621-9962'}]

print("Name\t\tCountry\t\tCard Number")

for user in users:

print (user['Name']+"\t\t"+user['Country']+"\t\t"+user['Card Number'])

First, I create a list of dictionaries with user names, countries, and card numbers. Then I use a loop to display these details. I’m not masking any data here yet. You can use any Python IDE, CLI, or online Python interpreters to execute this. I’m using the CLI. Let’s look at the output of this code.

As you can see, the card number, which is sensitive information, is clearly visible. Now let’s change the code slightly to mask the card details.

users = [{'Name':'Tony', 'Country': 'France', 'Card Number':'3542-7583-7228-3788'},

{'Name':'Steve','Country':'Austria', 'Card Number':'3881-8829-5554-4875'},

{'Name':'Peter', 'Country':'Spain', 'Card Number':'8445-5556-9621-9962'}]

print("Name\t\tCountry\t\tCard Number")

for user in users:

card\_number = 'XXXXXXXXXXXXXX'+user['Card Number']

[-4:]

print (user['Name']+"\t\t"+user['Country']+"\t\t"+card\_number)

I’ve changed all characters except the last 4 digits of the card number to ‘X,’ which acts as a mask in this code. Let’s look at the output.

This shows how data masking prevents sensitive information from being completely visible. In this case, data was hard-coded in the code itself, which is never the case in practical applications. Data is fetched from the database. Now let’s see how we can use a data mask while getting data from the database.

Data Masking in Database

I’ll be using MariaDB for this example, which comes pre-installed with the Kali Linux operating system. I start the service by running the following command:

service mysql start

First, let’s create a database and add data.

mysql -u root

create database sample;

use sample;

create table users(Name varchar(10),Country varchar(10), Card\_Number varchar(20));

insert into users values ('Tony','France','3542-7583-7228-3788'),('Steve','Austria','3881-8829-5554-4875'),('Peter','Spain','8445-5556-9621-9962');

select \* from users;

You can see that when I use the select function, all data is clearly visible. Usually, applications use a select query to fetch data from a database. How can we change this to get masked data? MySQL database has many [built-in functions](https://www.mysql.com/products/enterprise/masking.html) for data masking. I’ll use a simple string-replace approach. The query is as follows:

SELECT Name,Country,CONCAT(REGEXP\_REPLACE (LEFT(Card\_Number,16), '.', 'X'),RIGHT(Card\_Number,4)) from users;

The logic here is simple. We fetch data for all columns, but for the card number, we replace the first sixteen characters with ‘X’ and concatenate the remaining four characters to the replaced string. The output looks something like this:

That’s how simple masking is!

Data Masking Best Practices

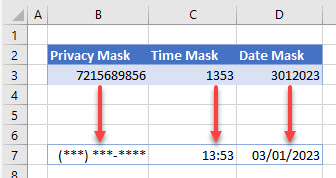
There are various approaches to data masking, and we need to follow the most secure approaches. We’ve gone through different aspects of data masking and learned how important and easy it is. I’ll conclude with some best practices for data masking.

1. Find and mask all sensitive data. If you have different databases and places where you store sensitive data, find and mask all of them.
2. Mask data at the origin. If you think masking data only on the client-side is enough, you’re wrong. Various tools can intercept traffic before it reaches the client application. If you’re masking data after it reaches the client, malicious actors can still get hold of unmasked data from the network.
3. Use irreversible data masking techniques. The whole point of data masking is to protect sensitive data. If users can convert masked data back to original data, there’s no point in masking it. For example, masking digits with alphabetic characters at the associated positions (1->a, 2->b, 3->c, etc.) is not a secure approach because users can reverse it.

In conclusion, data masking is a simple yet effective technique for data security. If your organization deals with sensitive data, data masking is a must.

# How to Mask Data Input in Excel & Google Sheets

*This tutorial demonstrates how to mask data input in Excel and Google Sheets.*



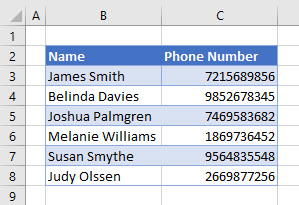
What does it mean to “mask” data in Excel?

* Do you want to [**hide sensitive information**](https://www.automateexcel.com/how-to/mask-data-input/#privacy)?
* Or, do you want to apply a **number format** that can [**simplify data entry**](https://www.automateexcel.com/how-to/mask-data-input/#format)?

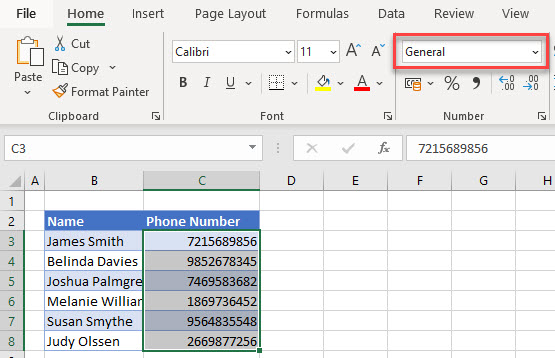
[Click here](https://www.automateexcel.com/how-to/mask-data-input/#GoogleSheets) to jump to the **Google Sheets** walkthrough.

**Mask Data for Privacy**

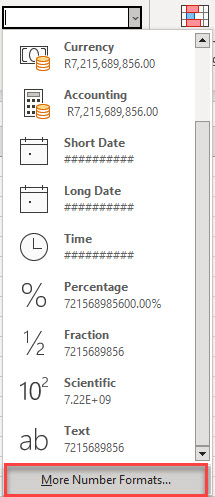
If you wish to mask data for privacy reasons, you can represent the data by a series of hashes or other [symbols](https://www.automateexcel.com/how-to/insert-sign-symbol/). Consider the following list of names and [phone numbers](https://www.automateexcel.com/how-to/format-phone-numbers). Say you want to mask the phone numbers so passersby can’t see them.



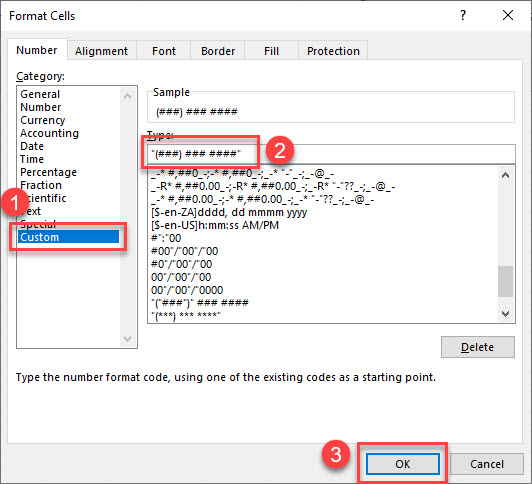
1. [Highlight](https://www.automateexcel.com/how-to/select-multiple-cells/) the phone number data cells, and then in the **Ribbon**, go to **Home > Number**. Click the **Number Format** drop down.



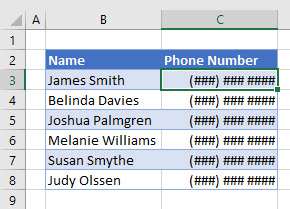
1. Scroll down to the bottom and click **More Number Formats…**



1. Click **Custom**, and then type in the input mask inside double quotes. In this example, for phone numbers represented by hashes, that’s **“(###) ### ####”**. Click **OK** to apply the data mask.



1. Your data is now formatted with a privacy mask.



**Mask Data Input**

When entering data such as [time](https://www.automateexcel.com/how-to/data-validation-date-format/) or [date](https://www.automateexcel.com/formulas/change-date-format/), it can be time-consuming to enter preformatted values including [symbols](https://www.automateexcel.com/how-to/signs-symbols-mean-in-formulas/) like “**:**” or “**/**“. Therefore, you can create an input mask that allows a user to enter [numbers only](https://www.automateexcel.com/how-to/limit-restrict-cell-value/) and then [automatically format](https://www.automateexcel.com/how-to/autoformat-automatic-formatting/) them.

ส่วนบนของฟอร์ม

**Try our AI Formula Generator**

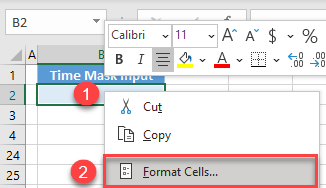
Generate

ส่วนล่างของฟอร์ม

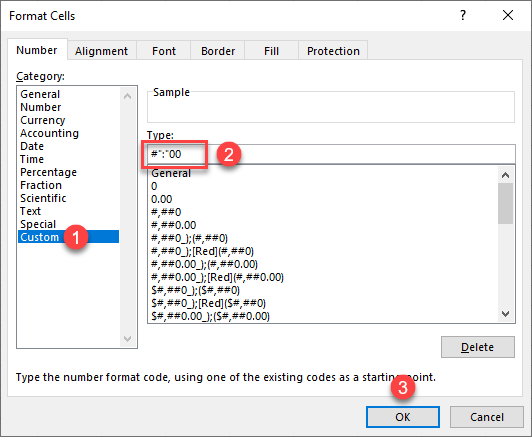
**Time**

For example, if only numbers for a time are inserted, you want Excel to automatically format them as HH:MM. Say you want to create an input mask for a time in cell B2.

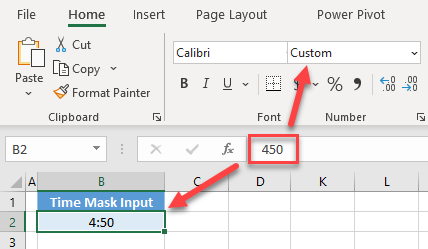
1. Right-click the cell where you want to create a mask input (here, cell B2), and choose **Format Cells…**



* 1. In the [Format Cells window](https://www.automateexcel.com/shortcuts/format-cells-dialog-box-keyboard-shortcut/), choose **Number > Custom**, enter **#”:”00** in the **Type** box, and click **OK**.



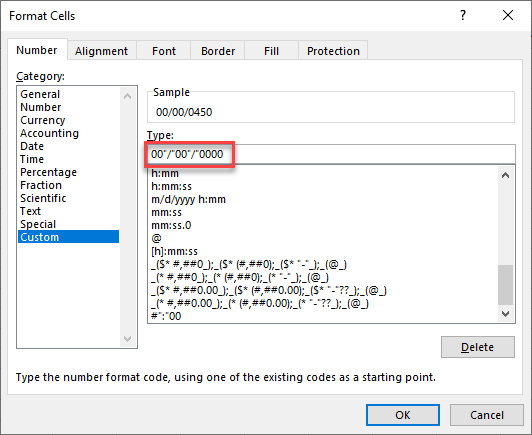
Now, you can **enter numbers only for the time** (here, **450**), and Excel formats it as **4:50**.



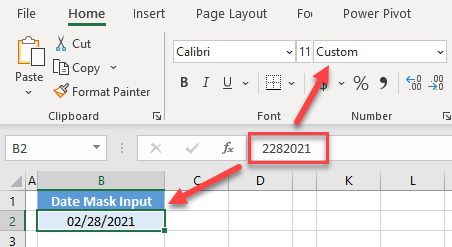
As you can see, this cell has a custom format. Therefore, it won’t be recognized as a time format by Excel.

**Date**

In a similar way, you can also mask data input for dates. In this case, each entry should be only numbers (for example, **02282021**), and you want Excel to display it as a date (**02/28/2021**). To do this, repeat the steps above, but for Step 2, type in the following: **00″/”00″/”0000**.



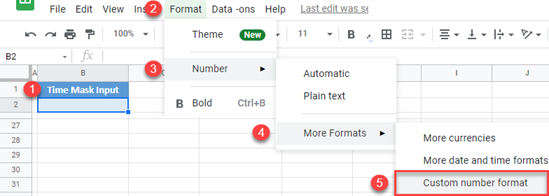
Now, you can enter **02282021** in cell B2, and it is displayed as **02/28/2021**.



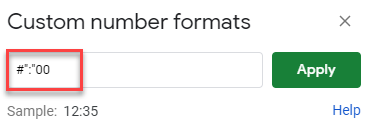
**Mask Data Input in Google Sheets**

You can also mask data input for **times** in Google Sheets.

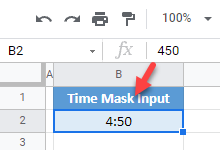
1. Select the cell where you want to create a mask input (here, cell B2), and in the **Menu**, go to **Format > Number > More Formats > Custom number format**.



1. In the number format box, **enter #”:”00** and click **Apply**.



The result is the same as in Excel. If you enter 450 in cell B2, it is formatted as 4:50.



Using the steps above, you can also mask data input for **dates** by entering **00″/”00″/”0000** as a custom number format. Or mask for **privacy** with, for example, the format **“(###) ### ####”**.

Top ten ways to clean your data

*Excel for Microsoft 365 Excel 2021 Excel 2019 Excel 2016 Excel 2013*

Misspelled words, stubborn trailing spaces, unwanted prefixes, improper cases, and nonprinting characters make a bad first impression. And that is not even a complete list of ways your data can get dirty. Roll up your sleeves. It is time for some major spring-cleaning of your worksheets with Microsoft Excel.

**The Basics Of Cleaning Your Data**

You don't always have control over the format and type of data that you import from an external data source, such as a database, text file, or a Web page. Before you can analyze the data, you often need to clean it up. Fortunately, Excel has many features to help you get data in the precise format that you want. Sometimes, the task is straightforward and there is a specific feature that does the job for you. For example, you can easily use Spell Checker to clean up misspelled words in columns that contain comments or descriptions. Or, if you want to remove duplicate rows, you can quickly do this by using the **Remove Duplicates** dialog box.

At other times, you may need to manipulate one or more columns by using a formula to convert the imported values into new values. For example, if you want to remove trailing spaces, you can create a new column to clean the data by using a formula, filling down the new column, converting that new column's formulas to values, and then removing the original column.

The basic steps for cleaning data are as follows:

1. Import the data from an external data source.
2. Create a backup copy of the original data in a separate workbook.
3. Ensure that the data is in a tabular format of rows and columns with: similar data in each column, all columns and rows visible, and no blank rows within the range. For best results, use an Excel table.
4. Do tasks that don't require column manipulation first, such as spell-checking or using the **Find and Replace** dialog box.
5. Next, do tasks that do require column manipulation. The general steps for manipulating a column are:
   1. Insert a new column (B) next to the original column (A) that needs cleaning.
   2. Add a formula that will transform the data at the top of the new column (B).
   3. Fill down the formula in the new column (B). In an Excel table, a calculated column is automatically created with values filled down.
   4. Select the new column (B), copy it, and then paste as values into the new column (B).
   5. Remove the original column (A), which converts the new column from B to A.

To periodically clean the same data source, consider recording a macro or writing code to automate the entire process. There are also a number of external add-ins written by third-party vendors, listed in the [Third-party providers](https://support.microsoft.com/en-us/office/top-ten-ways-to-clean-your-data-2844b620-677c-47a7-ac3e-c2e157d1db19#bmthird_party_providers) section, that you can consider using if you don't have the time or resources to automate the process on your own.

| **More information** | **Description** |
| --- | --- |
| [Fill data automatically in worksheet cells](https://support.microsoft.com/en-us/office/fill-data-automatically-in-worksheet-cells-74e31bdd-d993-45da-aa82-35a236c5b5db) | Shows how to use the **Fill** command. |
| [Create and format tables](https://support.microsoft.com/en-us/office/create-and-format-tables-e81aa349-b006-4f8a-9806-5af9df0ac664)  [Resize a table by adding or removing rows and columns](https://support.microsoft.com/en-us/office/resize-a-table-by-adding-or-removing-rows-and-columns-e65ae4bb-e44b-43f4-ad8b-7d68784f1165)  [Use calculated columns in an Excel table](https://support.microsoft.com/en-us/office/use-calculated-columns-in-an-excel-table-873fbac6-7110-4300-8f6f-aafa2ea11ce8) | Show how to create an Excel table and add or delete columns or calculated columns. |
| [Create a macro](https://support.microsoft.com/en-us/office/automate-tasks-with-the-macro-recorder-974ef220-f716-4e01-b015-3ea70e64937b) | Shows several ways to automate repetitive tasks by using a macro. |

**Spell Checking**

You can use a spell checker to not only find misspelled words, but to find values that are not used consistently, such as product or company names, by adding those values to a custom dictionary.

| **More information** | **Description** |
| --- | --- |
| [Check spelling and grammar](https://support.microsoft.com/en-us/office/check-spelling-and-grammar-in-office-5cdeced7-d81d-47de-9096-efd0ee909227) | Shows how to correct misspelled words on a worksheet. |
| [Use custom dictionaries to add words to the spelling checker](https://support.microsoft.com/en-us/office/add-or-edit-words-in-a-spell-check-dictionary-56e5c373-29f8-4d11-baf6-87151725c0dc) | Explains how to use custom dictionaries. |

**Removing Dublicate Rows**

Duplicate rows are a common problem when you import data. It is a good idea to filter for unique values first to confirm that the results are what you want before you remove duplicate values.

| **More information** | **Description** |
| --- | --- |
| [Filter for unique values or remove duplicate values](https://support.microsoft.com/en-us/office/filter-for-unique-values-or-remove-duplicate-values-ccf664b0-81d6-449b-bbe1-8daaec1e83c2) | Shows two closely-related procedures: how to filter for unique rows and how to remove duplicate rows. |

**Finding And Replacing Text**

You may want to remove a common leading string, such as a label followed by a colon and space, or a suffix, such as a parenthetic phrase at the end of the string that is obsolete or unnecessary. You can do this by finding instances of that text and then replacing it with no text or other text.

| **More information** | **Description** |
| --- | --- |
| [Check if a cell contains text (case-insensitive)](https://support.microsoft.com/en-us/office/check-if-a-cell-contains-text-case-insensitive-7bb505c7-2815-4a7a-9544-57a5f0dcd551)  [Check if a cell contains text (case-sensitive)](https://support.microsoft.com/en-us/office/check-if-a-cell-contains-text-case-sensitive-7f302a1e-5a3b-421c-b00a-615c89bc39ba) | Show how to use the **Find** command and several functions to find text. |
| [Remove characters from text](https://support.microsoft.com/en-us/office/split-text-into-different-columns-with-functions-49ec57f9-3d5a-44b2-82da-50dded6e4a68) | Shows how to use the **Replace** command and several functions to remove text. |
| [Find or replace text and numbers on a worksheet](https://support.microsoft.com/en-us/office/find-or-replace-text-and-numbers-on-a-worksheet-0e304ca5-ecef-4808-b90f-fdb42f892e90) | Show how to use the **Find** and **Replace** dialog boxes. |
| [FIND, FINDB](https://support.microsoft.com/en-us/office/find-findb-functions-c7912941-af2a-4bdf-a553-d0d89b0a0628)  [SEARCH, SEARCHB](https://support.microsoft.com/en-us/office/search-searchb-functions-9ab04538-0e55-4719-a72e-b6f54513b495)  [REPLACE, REPLACEB](https://support.microsoft.com/en-us/office/replace-replaceb-functions-8d799074-2425-4a8a-84bc-82472868878a)  [SUBSTITUTE](https://support.microsoft.com/en-us/office/substitute-function-6434944e-a904-4336-a9b0-1e58df3bc332)  [LEFT, LEFTB](https://support.microsoft.com/en-us/office/left-leftb-functions-9203d2d2-7960-479b-84c6-1ea52b99640c)  [RIGHT, RIGHTB](https://support.microsoft.com/en-us/office/right-rightb-functions-240267ee-9afa-4639-a02b-f19e1786cf2f)  [LEN, LENB](https://support.microsoft.com/en-us/office/len-lenb-functions-29236f94-cedc-429d-affd-b5e33d2c67cb) [MID, MIDB](https://support.microsoft.com/en-us/office/mid-midb-functions-d5f9e25c-d7d6-472e-b568-4ecb12433028) | These are the functions that you can use to do various string manipulation tasks, such as finding and replacing a substring within a string, extracting portions of a string, or determining the length of a string. |

**Changing The Case of Text**

Sometimes text comes in a mixed bag, especially when the case of text is concerned. Using one or more of the three Case functions, you can convert text to lowercase letters, such as e-mail addresses, uppercase letters, such as product codes, or proper case, such as names or book titles.

| **More information** | **Description** |
| --- | --- |
| [Change the case of text](https://support.microsoft.com/en-us/office/change-the-case-of-text-01481046-0fa7-4f3b-a693-496795a7a44d) | Shows how to use the three Case functions. |
| [LOWER](https://support.microsoft.com/en-us/office/lower-function-3f21df02-a80c-44b2-afaf-81358f9fdeb4) | Converts all uppercase letters in a text string to lowercase letters. |
| [PROPER](https://support.microsoft.com/en-us/office/proper-function-52a5a283-e8b2-49be-8506-b2887b889f94) | Capitalizes the first letter in a text string and any other letters in text that follow any character other than a letter. Converts all other letters to lowercase letters. |
| [UPPER](https://support.microsoft.com/en-us/office/upper-function-c11f29b3-d1a3-4537-8df6-04d0049963d6) | Converts text to uppercase letters. |

**Removing Spaces and Non Printing Characters From Text**

Sometimes text values contain leading, trailing, or multiple embedded space characters (Unicode character set values 32 and 160), or nonprinting characters (Unicode character set values 0 to 31, 127, 129, 141, 143, 144, and 157). These characters can sometimes cause unexpected results when you sort, filter, or search. For example, in the external data source, users may make typographical errors by inadvertently adding extra space characters, or imported text data from external sources may contain nonprinting characters that are embedded in the text. Because these characters are not easily noticed, the unexpected results may be difficult to understand. To remove these unwanted characters, you can use a combination of the TRIM, CLEAN, and SUBSTITUTE functions.

| **More information** | **Description** |
| --- | --- |
| [CODE](https://support.microsoft.com/en-us/office/code-function-c32b692b-2ed0-4a04-bdd9-75640144b928) | Returns a numeric code for the first character in a text string. |
| [CLEAN](https://support.microsoft.com/en-us/office/clean-function-26f3d7c5-475f-4a9c-90e5-4b8ba987ba41) | Removes the first 32 nonprinting characters in the 7-bit ASCII code (values 0 through 31) from text. |
| [TRIM](https://support.microsoft.com/en-us/office/trim-function-410388fa-c5df-49c6-b16c-9e5630b479f9) | Removes the 7-bit ASCII space character (value 32) from text. |
| [SUBSTITUTE](https://support.microsoft.com/en-us/office/substitute-function-6434944e-a904-4336-a9b0-1e58df3bc332) | You can use the SUBSTITUTE function to replace the higher value Unicode characters (values 127, 129, 141, 143, 144, 157, and 160) with the 7-bit ASCII characters for which the TRIM and CLEAN functions were designed. |

**Fixing Number And Signs**

There are two main issues with numbers that may require you to clean the data: the number was inadvertently imported as text, and the negative sign needs to be changed to the standard for your organization.

| **More information** | **Description** |
| --- | --- |
| [Convert numbers stored as text to numbers](https://support.microsoft.com/en-us/office/convert-numbers-stored-as-text-to-numbers-40105f2a-fe79-4477-a171-c5bad0f0a885) | Shows how to convert numbers that are formatted and stored in cells as text, which can cause problems with calculations or produce confusing sort orders, to number format. |
| [DOLLAR](https://support.microsoft.com/en-us/office/dollar-function-a6cd05d9-9740-4ad3-a469-8109d18ff611) | Converts a number to text format and applies a currency symbol. |
| [TEXT](https://support.microsoft.com/en-us/office/text-function-20d5ac4d-7b94-49fd-bb38-93d29371225c) | Converts a value to text in a specific number format. |
| [FIXED](https://support.microsoft.com/en-us/office/fixed-function-ffd5723c-324c-45e9-8b96-e41be2a8274a) | Rounds a number to the specified number of decimals, formats the number in decimal format by using a period and commas, and returns the result as text. |
| [VALUE](https://support.microsoft.com/en-us/office/value-function-257d0108-07dc-437d-ae1c-bc2d3953d8c2) | Converts a text string that represents a number to a number. |

**Merging And Splitting Columns**

A common task after importing data from an external data source is to either merge two or more columns into one, or split one column into two or more columns. For example, you may want to split a column that contains a full name into a first and last name. Or, you may want to split a column that contains an address field into separate street, city, region, and postal code columns. The reverse may also be true. You may want to merge a First and Last Name column into a Full Name column, or combine separate address columns into one column. Additional common values that may require merging into one column or splitting into multiple columns include product codes, file paths, and Internet Protocol (IP) addresses.

| **More information** | **Description** |
| --- | --- |
| [Combine first and last names](https://support.microsoft.com/en-us/office/combine-first-and-last-names-6a2c1222-837d-482e-94f0-a38a4effb1b5)  [Combine text and numbers](https://support.microsoft.com/en-us/office/combine-text-and-numbers-a32c8e0e-90a2-435b-8635-5dd2209044ad)  [Combine text with a date or time](https://support.microsoft.com/en-us/office/combine-text-with-a-date-or-time-cef6a66c-8176-470c-ba85-4b030405dfbf)  [Combine two or more columns by using a function](https://support.microsoft.com/en-us/office/combine-text-from-two-or-more-cells-into-one-cell-81ba0946-ce78-42ed-b3c3-21340eb164a6) | Show typical examples of combining values from two or more columns. |
| [Split text into different columns with the Convert Text to Columns Wizard](https://support.microsoft.com/en-us/office/split-text-into-different-columns-with-the-convert-text-to-columns-wizard-30b14928-5550-41f5-97ca-7a3e9c363ed7) | Shows how to use this wizard to split columns based on various common delimiters. |
| [Split text into different columns with functions](https://support.microsoft.com/en-us/office/split-text-into-different-columns-with-functions-49ec57f9-3d5a-44b2-82da-50dded6e4a68) | Shows how to use the LEFT, MID, RIGHT, SEARCH, and LEN functions to split a name column into two or more columns. |
| [Combine or split the contents of cells](https://support.microsoft.com/en-us/office/distribute-the-contents-of-a-cell-into-adjacent-columns-ce6b3ffd-2ed9-4ee7-a7cc-86ef70121290) | Shows how to use the CONCATENATE function, & (ampersand) operator, and Convert Text to Columns Wizard. |
| [Merge cells or split merged cells](https://support.microsoft.com/en-us/office/merge-and-unmerge-cells-5cbd15d5-9375-4540-907f-c673a93fcedf) | Shows how to use the **Merge Cells**, **Merge Across**, and **Merge and Center** commands. |
| [CONCATENATE](https://support.microsoft.com/en-us/office/concatenate-function-8f8ae884-2ca8-4f7a-b093-75d702bea31d) | Joins two or more text strings into one text string. |

**Transforming And Rearranging Columns And Rows**

Most of the analysis and formatting features in Office Excel assume that the data exists in a single, flat two-dimensional table. Sometimes you may want to make the rows become columns, and the columns become rows. At other times, data is not even structured in a tabular format, and you need a way to transform the data from a nontabular to a tabular format.

| **More information** | **Description** |
| --- | --- |
| [TRANSPOSE](https://support.microsoft.com/en-us/office/transpose-function-ed039415-ed8a-4a81-93e9-4b6dfac76027) | Returns a vertical range of cells as a horizontal range, or vice versa. |

**Reconciling Table Data By Joining And Matching**

Occasionally, database administrators use Office Excel to find and correct matching errors when two or more tables are joined. This might involve reconciling two tables from different worksheets, for example, to see all records in both tables or to compare tables and find rows that don't match.

| **More information** | **Description** |
| --- | --- |
| [Look up values in a list of data](https://support.microsoft.com/en-us/office/look-up-values-in-a-list-of-data-c249efc5-5847-4329-bfee-ecffead5ef88) | Shows common ways to look up data by using the lookup functions. |
| [LOOKUP](https://support.microsoft.com/en-us/office/lookup-function-446d94af-663b-451d-8251-369d5e3864cb) | Returns a value either from a one-row or one-column range or from an array. The LOOKUP function has two syntax forms: the vector form and the array form. |
| [HLOOKUP](https://support.microsoft.com/en-us/office/hlookup-function-a3034eec-b719-4ba3-bb65-e1ad662ed95f) | Searches for a value in the top row of a table or an array of values, and then returns a value in the same column from a row you specify in the table or array. |
| [VLOOKUP](https://support.microsoft.com/en-us/office/vlookup-function-0bbc8083-26fe-4963-8ab8-93a18ad188a1) | Searches for a value in the first column of a table array and returns a value in the same row from another column in the table array. |
| [INDEX](https://support.microsoft.com/en-us/office/index-function-a5dcf0dd-996d-40a4-a822-b56b061328bd) | Returns a value or the reference to a value from within a table or range. There are two forms of the INDEX function: the array form and the reference form. |
| [MATCH](https://support.microsoft.com/en-us/office/match-function-e8dffd45-c762-47d6-bf89-533f4a37673a) | Returns the relative position of an item in an array that matches a specified value in a specified order. Use MATCH instead of one of the LOOKUP functions when you need the position of an item in a range instead of the item itself. |
| [OFFSET](https://support.microsoft.com/en-us/office/offset-function-c8de19ae-dd79-4b9b-a14e-b4d906d11b66) | Returns a reference to a range that is a specified number of rows and columns from a cell or range of cells. The reference that is returned can be a single cell or a range of cells. You can specify the number of rows and the number of columns to be returned. |

**Third-Party Providers**

The following is a partial list of third-party providers that have products that are used to clean data in a variety of ways.

**Note:**Microsoft does not provide support for third-party products.

| **Provider** | **Product** |
| --- | --- |
| Add-in Express Ltd. | [Ultimate Suite for Excel, Merge Tables Wizard, Duplicate Remover, Consolidate Worksheets Wizard, Combine Rows Wizard, Cell Cleaner, Random Generator, Merge Cells, Quick Tools for Excel, Random Sorter, Advanced Find & Replace, Fuzzy Duplicate Finder, Split Names, Split Table Wizard, Workbook Manager](https://www.ablebits.com/downloads/index.php) |
| Add-Ins.com | [Duplicate Finder](https://www.add-ins.com/duplicate_finder.htm) |
| AddinTools | [AddinTools Assist](http://www.addintools.com/english/assist/) |
| WinPure | [ListCleaner Lite](https://www.winpure.com/lite.html) [ListCleaner Pro](https://www.winpure.com/pro.html) [Clean and Match 2007](http://www.winpure.com/cleanmatch.html) |