OpenLearn

The Open University

B126_3

Data analysis: visualisations in Excel

This item contains selected online content. It is for use alongside, not as a replacement for the module website, which is the primary study format and contains activities and resources that cannot be replicated in the printed versions.

About this free course

This free course is an adapted extract from the Open University course B126 *Business data analytics and decision making* - www.open.ac.uk/courses/modules/b126.

This version of the content may include video, images and interactive content that may not be optimised for your device.

You can experience this free course as it was originally designed on OpenLearn, the home of free learning from The Open University –

There you'll also be able to track your progress via your activity record, which you can use to demonstrate your learning.

First published 2025.

Intellectual property

Contents

Introduction

The objectives of this course are to explore ways to visualise data sets in Excel, familiarise yourself with Excel functions and explore the relationship(s) between variables.

Figure 1 Making sense of visualising and exploring data

Please note: this course will require the use Microsoft Excel or a similar program.

This OpenLearn course is an adapted extract from the Open University course B126 *Business data analytics and decision making*.

Learning outcomes

After studying this course, you should be able to:

- explore the functionalities of Excel that are used for problem solving in a business context
- demonstrate the numeracy skills required for gathering and organising data for decision making related to a specific problem
- use graphical techniques (histograms and scatter diagrams) to provide a visual summary of available data
- recognise data presentation and communication techniques used in a range of traditional and electronic media
- describe the relationship between two variables (independent and dependent variables).

1 Excel spreadsheets: a tool for organising and starting to process data

Before making any business decision, managers need to see a clear picture of their data, which provides them with all the relevant information. Therefore, it is good practice to organise and present data in a way that makes it useful for decision making and problem solving. Microsoft Excel is the most attractive and useful tool for data analysis. Researchers and analysts alike use this tool for various applications in the real world, such as in business, medicine, academia, logistics, operations, transportation, tax and auditing, marketing, accounting and finance. Moreover, it is flexible enough to be used with all types of data, irrespective of whether the data is qualitative or quantitative.

In this section, you will make extensive use of Excel spreadsheets. Therefore, your task is to learn and familiarise yourself with the basics of using Excel. This will enhance your analytical skills as well as your employability skills.

This section briefly explains the various features and functions of Excel that are used by researchers and data analysts to explore, organise and analyse data.

Excel for OU students

If you are currently studying with the OU as a fee-paying student, you have free access to Microsoft Office 365. This includes the spreadsheet software Excel. For this you need to go to the OU Computing Guide and scroll down to 'Microsoft Office 365'. Click on the link. If you have not done so, then you should follow the instructions to sign up to get access for your free version of the software. If you have already installed Microsoft Excel on your laptop, then you may prefer to use your own version. Although earlier versions of Excel are not significantly different, the layout of some tabs and menus may vary slightly.

1.1 Accessing Microsoft Excel

There are many other options, of course, but being an OU student you have free access to Microsoft 365 that includes the spreadsheet software Excel. If you have already installed Microsoft Excel on your laptop, then you may prefer to use your own version. Although earlier versions of Excel are not significantly different, the layout of some tabs and menus may vary slightly.

- You can run Excel by double-clicking on the Excel icon on your desktop or laptop (or select **Excel** from your list of programs).
- Excel will open with a clean, new worksheet called **Book 1** that contains only one worksheet (**Sheet1**). You can add more sheets by clicking on the plus sign at the bottom of the worksheet or spreadsheet in case one sheet becomes too small or too cluttered. This will help you organise and manage your large sets of data, or a variety of data.
- Add a new worksheet by clicking on the plus sign.
- The new sheet **Sheet2** looks exactly like **Sheet1**. Each individual cell is labelled according to the **A1** reference style. This means that columns are labelled with letters from A to XFD and rows with number 1 to 1048576.

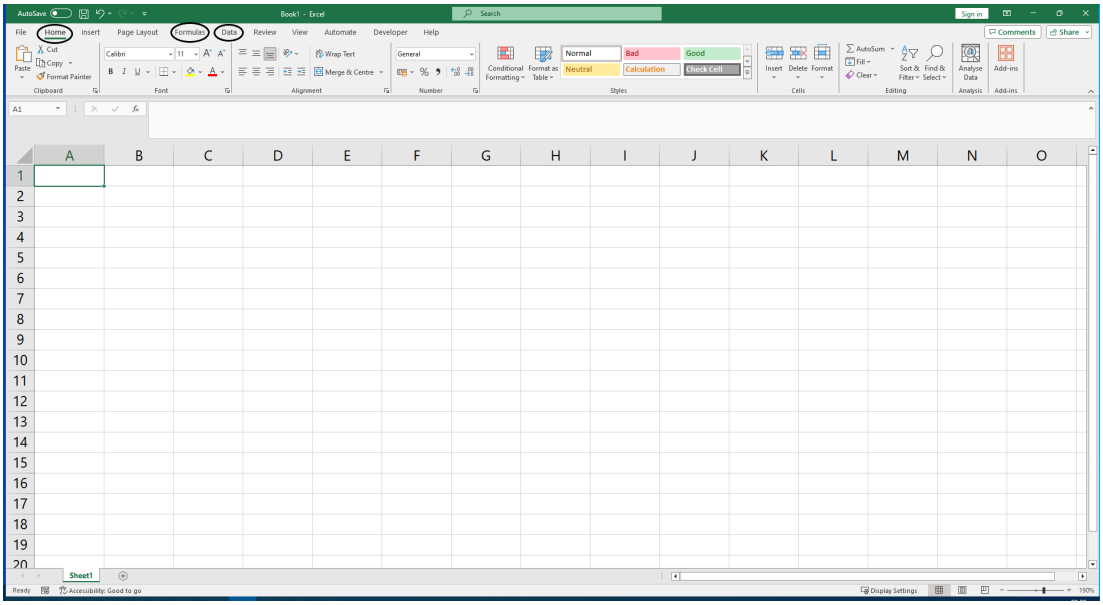


Figure 2 An empty worksheet in Excel

The latest versions of Excel are organised using a **ribbon** interface, which consists of a collection of icons for each tab. The screenshot above shows the **Home** ribbon with several icons for general editing such as font size, text alignment or cell styles. In this session, you will use the **Home**, **Formulas** and **Data** tabs in Excel. You may click on each tab and familiarise yourself with the large main icons.

1.2 Opening an Excel file

Excel stores the data in files that contain one or multiple worksheets. When you open a worksheet by double-clicking the Excel icon, it will usually display the data contained in the last active sheet.

- Download the file: JC Electrics template. JC Electrics is a company manufacturing heavy machinery, such as generators, electric motors and transformers.
- Click on **File**, then **Open** and navigate to the folder in which you saved this file and open the file by double-clicking it.
- Navigate to the first sheet, labelled ‘JC Electrics 1’.

You should now see quarterly data of units sold for JC Electrics in four columns. The screenshot below shows a spreadsheet with these columns labelled as: ‘Quarter’, ‘Generators’, ‘Transformers’ and ‘Electric Motors’.

A	B	C	D
JC Electrics			
(Quarterly data of units sold)			
Quarter	Generators	Electric Motors	Transformers
1	9	23	41
2	10	21	41
3	9	23	41
4	8	21	43
5	7	22	45
6	14	17	48
7	9	19	47
8	11	22	46
9	12	18	45
10	13	14	44
11	12	18	47
12	11	25	46
13	12	18	42
14	9	22	41
15	13	21	49
16	7	25	48
17	7	22	47
18	7	22	42
19	78	19	43
20	9	22	45
21	14	19	47
22	15	17	45
23	11	18	46
24	10	25	44

Figure 3 Data in Excel

1.3 Adding the Data Analysis ToolPak in Excel

In the next section of this unit, you are going to learn how to add the Data Analysis ToolPak in Excel and what its purpose is.

Data Analysis ToolPak is a Microsoft Excel add-in function that provides you with a set of statistical tools for analysing data. This add-in comes with Excel during installation. However, if it has not already been loaded or activated on the machine you are using, you need to do so to proceed further.

To load the Data Analysis ToolPak:

- Open the Excel worksheet (this will already be opened if you have opened ‘JC Electrics template’ file).
- Click on the **File** tab, then click on **Options**, and select **Add-Ins**.
- In the **Manage** box, select Excel **Add-ins** as shown in the screenshot below and click **Go (see Figure 4).**
- A new box will appear, and in the **Add-Ins** dialog box, select the **Analysis Tool Pak** check box, and then click **OK**.

The Data Analysis icon should now be available on the **Data** ribbon in the **Analysis** group. Start of Figure

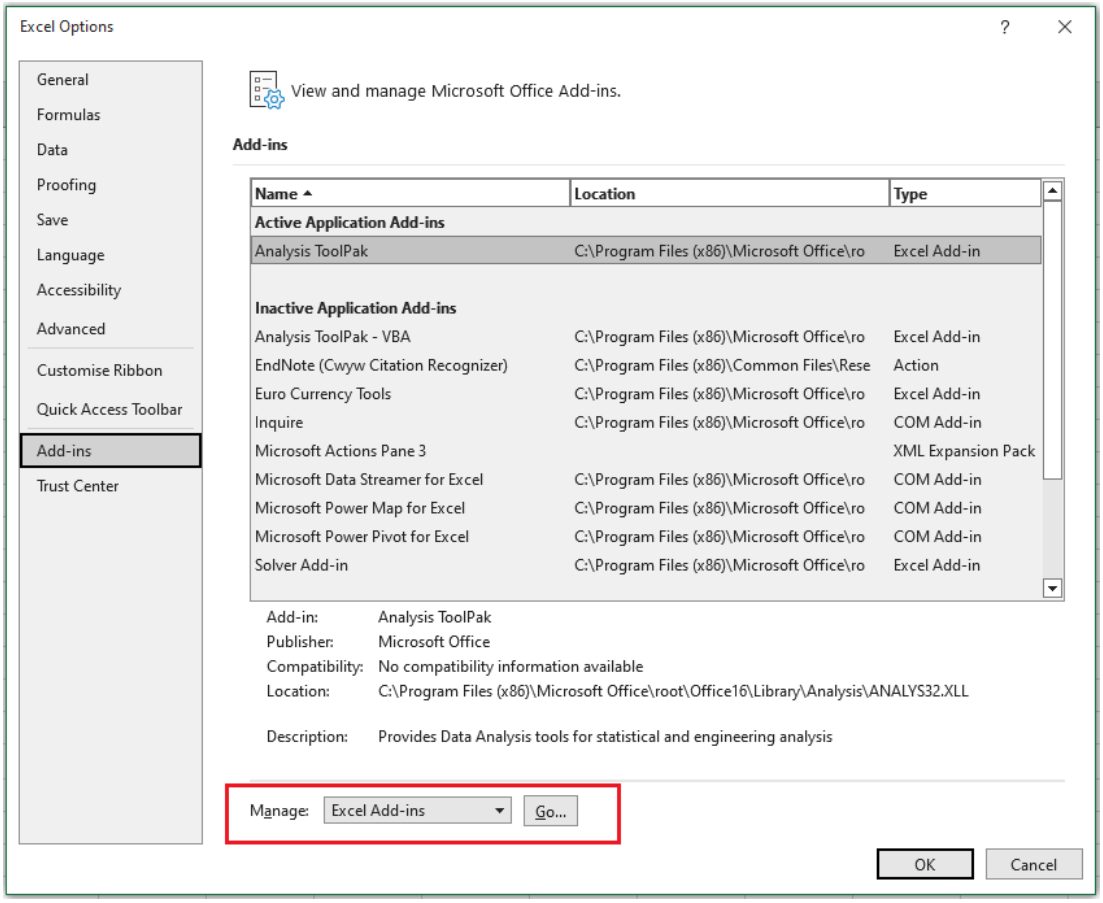


Figure 4 Activating the Data Analysis ToolPak

The Data Analysis icon should now be available on the **Data** ribbon in the **Analysis** group.

1.4 Decimal points and dates

If you are using a version of Excel other than the English (UK) one, you may have noticed the different date format and decimal point. In addition, one of the Excel's habits of automatically adjusting the format of cell contents can sometimes produce unwanted or incorrect results. Therefore, it is useful to learn how things can be corrected and adjusted in Excel.

In Excel spreadsheet, click on **File,** select **Options** and then select the **Advanced tab**; the results of this are shown below.

If you untick the box 'Use system separators', you can enter alternative symbols to use instead.

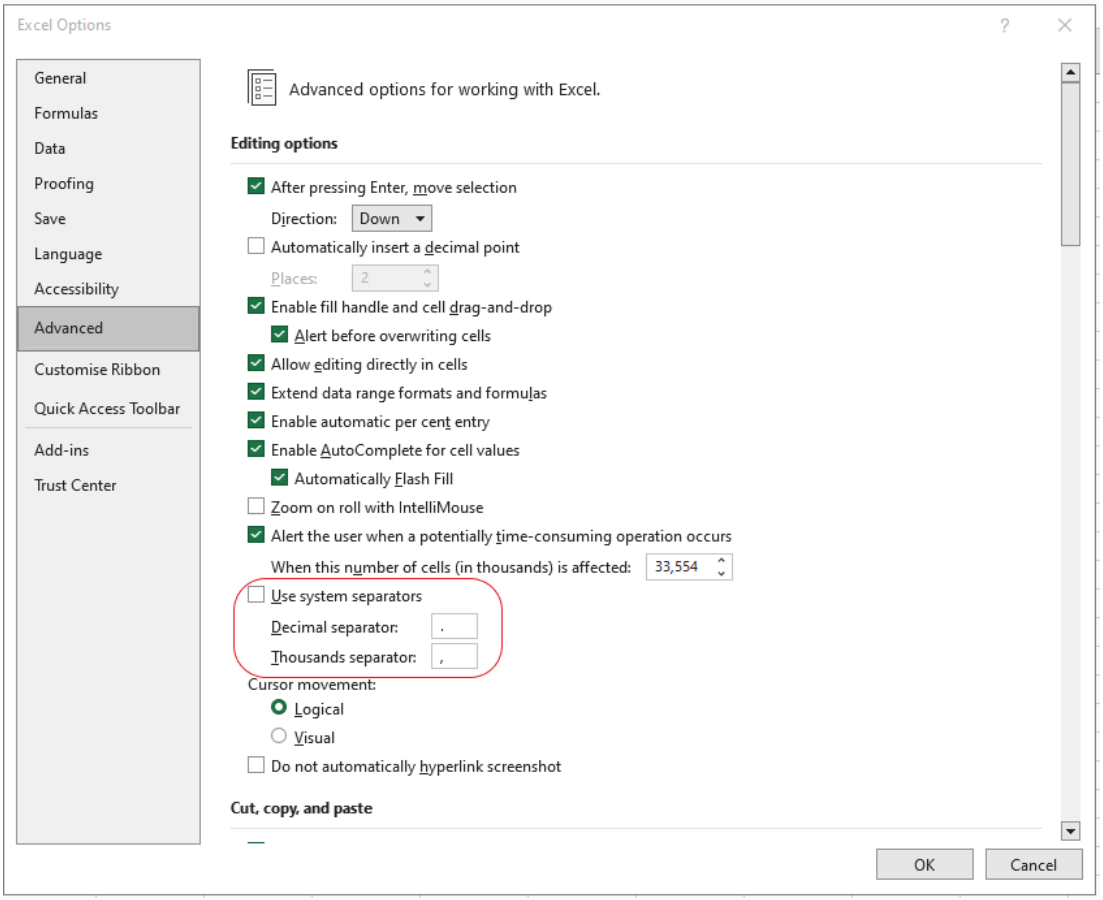


Figure 5 Advanced options in Excel

You can adjust the date format by selecting a cell that contains a date, then clicking on **Format** in the **Home** ribbon under **Cells** and **Format Cells …** or by right-clicking on the cells and selecting **Format Cells …** from the context menu.

The figure below shows you the ‘Format’ dialog for an English version of Excel installed on a computer. You can produce the desired date format; for example, **DD/MM/YYYY** by selecting the option and then clicking **OK**.

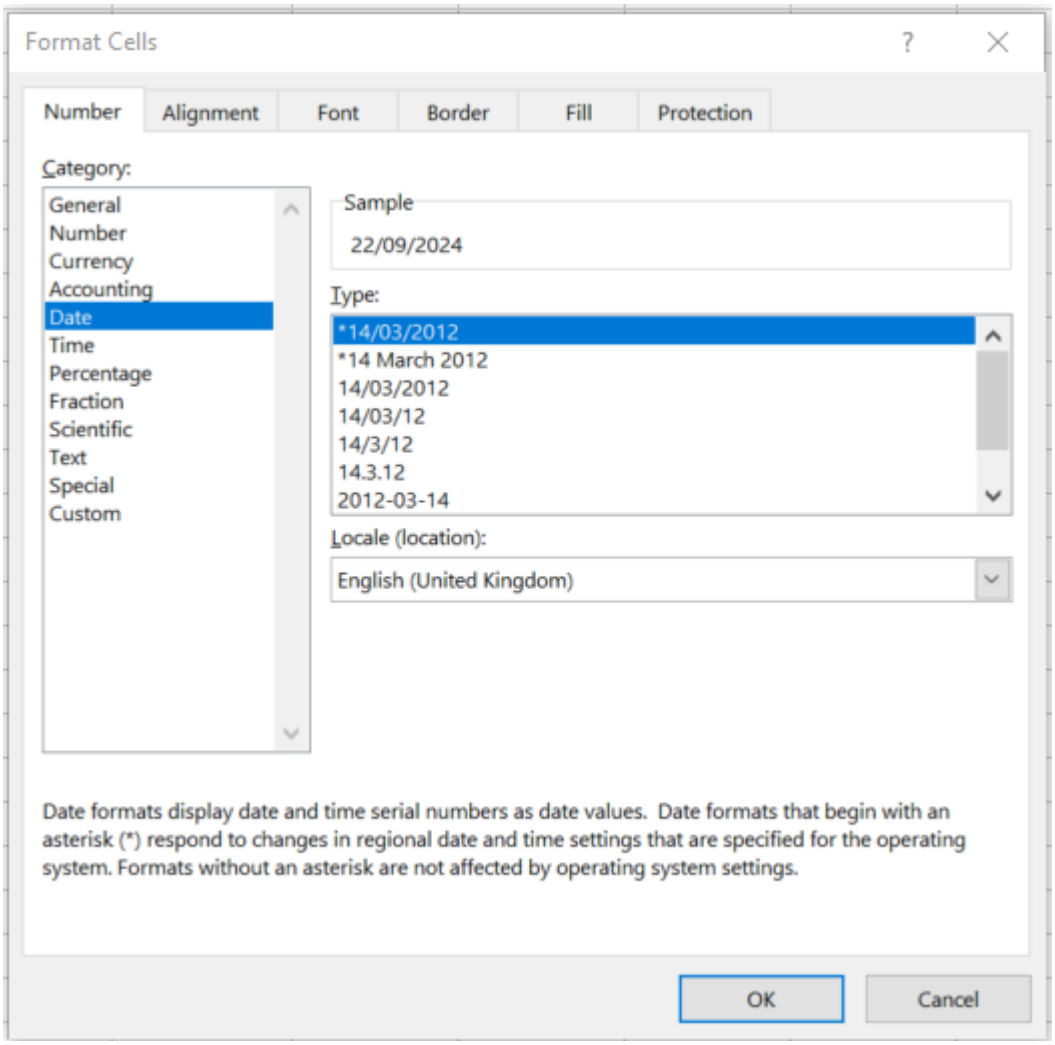


Figure 6 Changing the date format

1.5 Using shortcut keys in Excel

One useful task in Excel is to learn to use shortcut keys. This will speed up your Excel projects, especially when working under pressure in a professional environment.

Some shortcut keys are listed below. Try some of these out while reading the list.

Table 1 Navigating inside and between worksheets

Arrow keys	Move around the spreadsheet.
Page Down/ Page Up	Move screen down or up.
Ctrl + Arrow keys	Move to the edge of a region. This is useful for navigating large blocks of data, particularly with the Ctrl + Shift + Arrow selection functionality.
Home	Move to the beginning of a row.

Ctrl + Home	Move to the beginning of a worksheet. This is most useful if you have multiple worksheets and want to prepare a nice-looking workbook, by cycling through all worksheets pressing Ctrl + Page Down and Ctrl + Home for each sheet, which quickly puts the cursor in the upper-left corner.

Table 2 Data selection

Shift + Space	Select the entire row at the cursor position.
Ctrl + A	Select the entire worksheet or the data-containing area. Press Ctrl + A a second time to select the entire worksheet.
Shift + Arrow key	Extend the selection by one cell. This is one of the most useful shortcuts.
Ctrl + Shift + Arrow key	Extend the selection to the last cell with content in row or column. You can do this with the Page Up/Down keys.
Esc	Cancel selection.

Table 3 Editing

Ctrl + C	Copy active selection.
Ctrl + X	Cut active selection. Think carefully about whether you want to copy or cut a selection before pasting in each situation, because cell references in copied selections will point to other cells and not the original references when pasted.
Ctrl + V	Paste active selection.
Ctrl + Z	Undo last action.
Ctrl + Y	Redo last action.
Alt + Enter	Start a new line in the same cell when entering text.
Ctrl + D	Copy above cell down.

Table 4 Formulas and special functions

Ctrl + Shift + Enter	Enter an array formula. Must have a range selected first. (This is shown here only for reference and will be explained later.)
Shift + F3	Display the 'Insert Function' dialog.
F4	When editing a cell reference (e.g. 'H5'), pressing F4 makes this reference absolute (e.g., '$H$5'). Pressing F4 repeatedly makes only row or column absolute.
F9	Force re-calculation of worksheets. It can be used to calculate part of a formula, when selecting part of formula and pressing F9, this evaluates the selected part.
Shift + F9	Calculate the active worksheet.
Ctrl + S	Save the current workbook. Extremely useful for the occasional power outage or computer crash.

1.6 Use of Excel spreadsheets

Excel spreadsheets are used to manage, organise and present data in a systematic way. They can be used in many ways in different fields; for example, in finance, they are mostly used to present and analyse data such as accounting transactions (e.g., sales, payments, cost, forecasting and budgeting). Spreadsheets are also used to design templates of financial statements and survey results.

Some other uses are listed below:

- In schools and universities, spreadsheets are often used to manage student data in areas such as their grade performance, attendance or personal biography.
- In hospitals, spreadsheets are used to manage patient data, like their personal information, details of their illness or details of the medicines they use.
- Data is often exported from more complicated computer systems, such as manufacturing, financial or marketing systems, to allow managers or analysts to manipulate the data once it has been created, and to carry out forecasts, simulations and 'what-if' scenarios.

Excel also has many formatting options (borders, colour highlighting), to allow you to draw attention to aspects of your figures and tables.

Activity 1: Test your Excel knowledge

Allow approximately 20 minutes to complete this activity

Choose an answer to each of the questions below, and then hit 'Next' to move on to the next section. You can check your answer to each question as you go.

What is the shortcut key to copy a cell in an Excel spreadsheet?

- ☐ a) Ctrl + C
- ☐ b) Ctrl + F
- ☐ c) Shift + F3

What is the shortcut key to save the work in an Excel spreadsheet?

- ☐ a) Ctrl + V
- ☐ b) Ctrl + C
- ☐ c) Ctrl + S

What is the shortcut key to cancel the selection in the sheet or cell?

- ☐ a) Ctrl + Alt + Delete
- ☐ b) Esc
- ☐ c) F12

What is the shortcut key to display the insert function dialog?

- ☐ a) Shift + F3
- ☐ b) Ctrl + Z
- ☐ c) Shift + F2

In the next section, you will learn how to present and summarise a univariate (one variable only) data set in a table and graphical form.

2 Univariate data visualisation

In practice, there are two ways to visualise data in Excel. These are:

- tabular form
- graphical form.

While presenting data in Excel, it is important to know the features of data. If your data is **univariate** (that is, the data consists of many observations for only one variable) then you can either use a frequency table or a histogram to present the data and get an idea of its features. In the JC Electrics example, if the analysis is only carried out for 'Generators' (therefore only for one variable), this will be seen as a univariate analysis.

However, if your data is **bivariate (**that is, the data consists of two variables (an independent variable and a dependent variable)) and you need to know the relationship between these two variables, then, for example, you can use either a contingency table or scatter diagram to present the data and get an idea of its structure. You will learn about bivariate data visualisation later in this session.

The next section will briefly explain frequency tables.

2.1 Frequency tables

Before learning how to make a frequency tables and histograms in Excel, you first need to know what frequency distribution is, and why we need histograms.

The frequency of a value is the number of times that value appears in a data set. A frequency distribution table displays the pattern of frequencies of a variable in a tabular form. It gives the information of how many times each value of a variable occurs in a data set. A frequency distribution table is an effective way to present and organise the collected raw data so that all its features are summarised in a tabular form. The first step that a researcher or analyst must take with collected raw data is to organise and present the data in a way that is more meaningful, and easy to digest.

Frequency distribution tables are also called frequency tables, and in practice both terms are used interchangeably. In short, a frequency table gives you a snapshot of how your data is distributed and spread out.

A frequency distribution table has two columns: Column A and Column B. Column A presents the outcome of the values and Column B presents the frequency of the outcomes. We can understand this better with the example below.

Anna is an analyst, and she works Monday and Tuesday in a hospital. On Wednesday and Thursday, she works in a small accounting firm. On Friday, Saturday and Sunday she works in a bank.

Now you can display this data through a frequency distribution table, as shown in Figure 7.

A	B
Types of organisation	**No. of days**
hospital	2
accounting firm	2
bank	3

Figure 7 Displaying the data in the frequency distribution table

This table gives you an idea of how many days Anna works in each different organisation.

In the next section, you will learn various types of frequency distribution table.
In the next section, you will learn about various types of frequency distribution table.

2.2 Types of frequency distribution

There are four types of frequency distribution table:

- Ungrouped frequency distribution tables
- Grouped frequency distribution tables
- Relative frequency distribution tables
- Cumulative frequency distribution tables

Before describing each type of frequency distribution table, you need to know the difference between **ungrouped data** and **grouped data**.

In simple terms, **ungrouped data** is raw data that has not been categorised or grouped in categories. For example, a manager in a firm knows that 100 employees work in their firm; this is raw data because it does not tell you how many employees work in each department (e.g. buying, production, sales, distribution). However, if you have raw data that is categorised, it is defined as **grouped data**. For example, if this manager knows that 20 employees work in buying department, 30 in production and 40 in sales and 10 in distribution, then it means that the data is organised in such a way that it provides a clear indication of how many employees work in each department.

2.2.1 Concepts involved in frequency tables

The following terms are frequently used in frequency distribution:

Class interval or class limit: The lowest and the highest value defined for a class or group are called class limits. The lowest value is called the lower class limit and the highest value is called the upper class limit of that class. In the example below (see Table 5), the lower class limits are 7, 9, 11, 13, 15, and the upper limits are 8, 10, 12, 14 and 16.

Table 5 Class intervals for JC Electrics (Generators)

Class intervals	
7–8	
9–10	
11–12	
13–14	
15–16	

Midpoint or class mark: This is the average of a class interval, and is obtained by dividing the sum of upper and lower class limits by 2. Thus, the class mark of the interval 7– 8 is 7.5, as

$$(7+8)/2=7.5$$

Range: this is the difference between the maximum value and the minimum value of the data set. For example, the data set of a student's score in a maths quiz contains the maximum value 30 and the minimum value 8. Hence, to calculate the range, you must calculate $30\text{-}8=22$. In the JC Electrics, Generators example, the range is the difference between 15 (as the maximum value sold) and 7 (as the minimum value sold), hence the range is 8.

2.2.2 Ungrouped frequency distribution tables

When you are summarising small amounts of data, then it is better to organise and represent it in an ungrouped frequency distribution table. This is a type of distribution that shows how many times each individual value occurs in a data set.

For example, say you are interested to know how many plants people have in their homes. A survey gives the following figures as number of plants that 18 different people have in their homes:

Number of plants = 1, 5, 2, 2, 3, 3, 5, 5, 1, 1, 1, 3, 4, 4, 2, 3, 3, 3

To answer the question, first you need to see the frequency of each value in the data. Value 1 occur 4 times, so you can describe it as four people having one plant. Then, you do the same for the rest of the values, so: three people have two plants, six people have three plants, two people have four plants, and three people have five plants.

In the following activity, you will learn how to make ungrouped frequency tables in Excel.

Activity 2: How to make an ungrouped frequency table in Excel and a Bar chart

Allow approximately 55 minutes to complete this activity

In this activity, you will learn how to make an ungrouped frequency table in Excel, by following the instructions that are given below. Once you have produced the ungrouped frequency table in Excel, you may need to compare it with our final output by clicking 'Reveal discussion'. This will help you to see whether you have produced the accurate ungrouped frequency table or not.

- Open the Excel file JC Electrics (template). Select the worksheet named 'Ungrouped frequency table'. This sheet contains only the quarterly data of numbers of generators sold from the raw data in worksheet 'JC Electrics 1'.
- Label **Column B** using the variable name (such as 'Generators') and label **Column C** 'Frequency'.
- Find the minimum and maximum value in the data. In this example **=MAX(A5:A28)**, which is 15, and **=MIN(A5:A28)**, which is 7.
- Calculate the range: **(MAX – MIN)**, so, which is 8.
- Arrange the values in Column A from Smallest to Largest order. Select the values **(A5:A28)** in Column A, click **Data** in the toolbar and then click **Sort**, select **Continue with current selection** and **press Enter**. See below Figure 8, which shows how your information should be displayed.

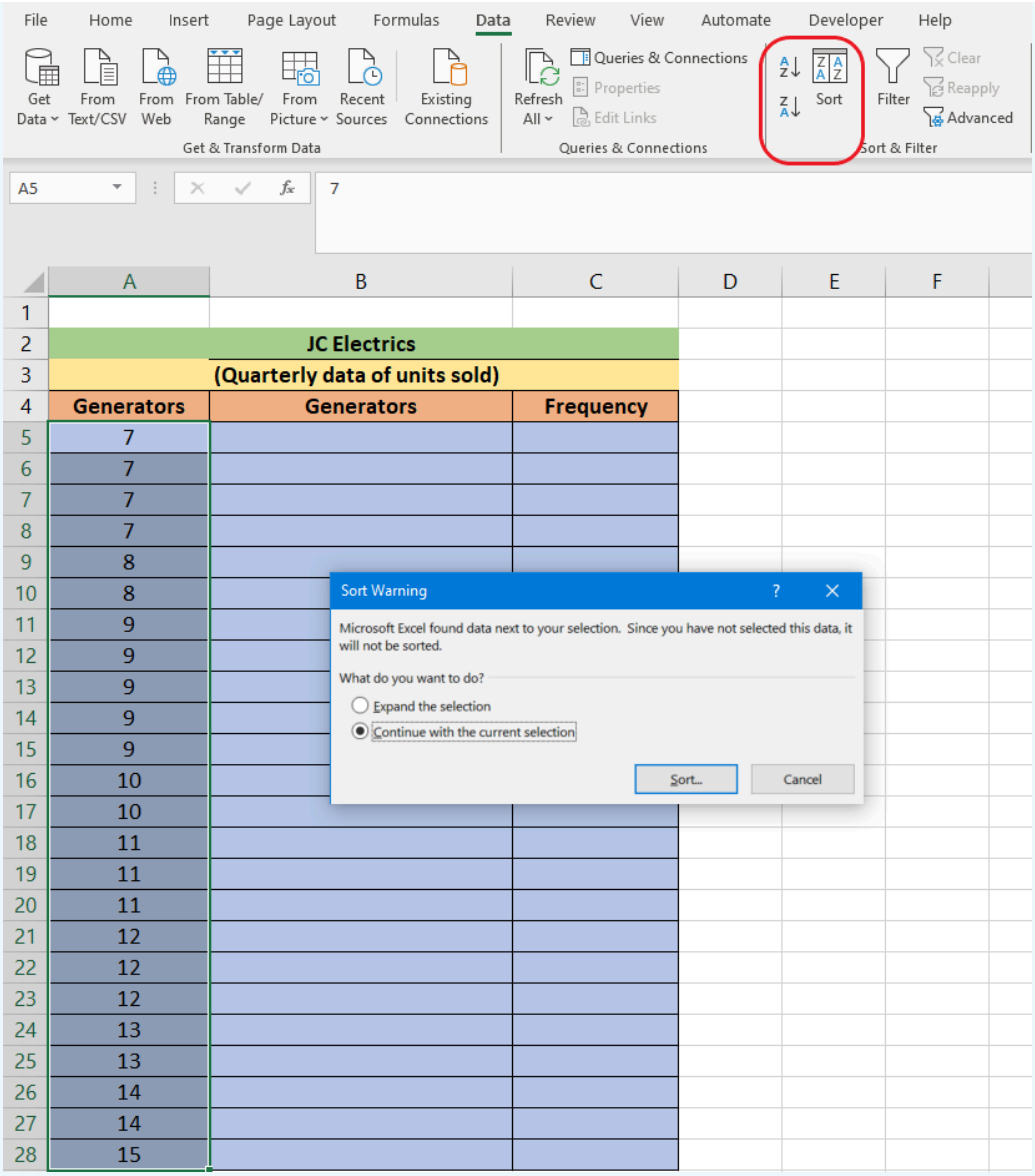


Figure 8 Arranging the data from ascending to descending order

- To count the number of quarters in which 7 units of Generators were sold, you need to calculate the frequency in Column C. Type **=COUNTIF (Range, value)**. For example, **=COUNTIF (A5:A28,7).** This calculation will return a value of 4, therefore there are 4 quarters when only 7 units of generators were sold.
- In Column B insert the different values of units sold such as: 7, 8, 9, 10, 11, 12, 13, 14 and 15 as they appear in Column A (see Figure 8).
- To continue calculating the frequency in Column C, the formula in cell C5 can now be adjusted to =COUNTIF($A$5:$A$28, B5). This is the same formula as above; however, this is now presented in a format that will allow you to copy the formula up to Cell C13. As noted previously (see Table 4), to lock a cell or a group of cells you can use function F4. In your formula, this will show the $ in front of the rows or columns you prefer to keep locked in your calculations.
- The copied formula in Cell C13, should be =COUNTIF($A$5:$A$28, B13). In column C, you will now have the calculated frequencies for all units sold.

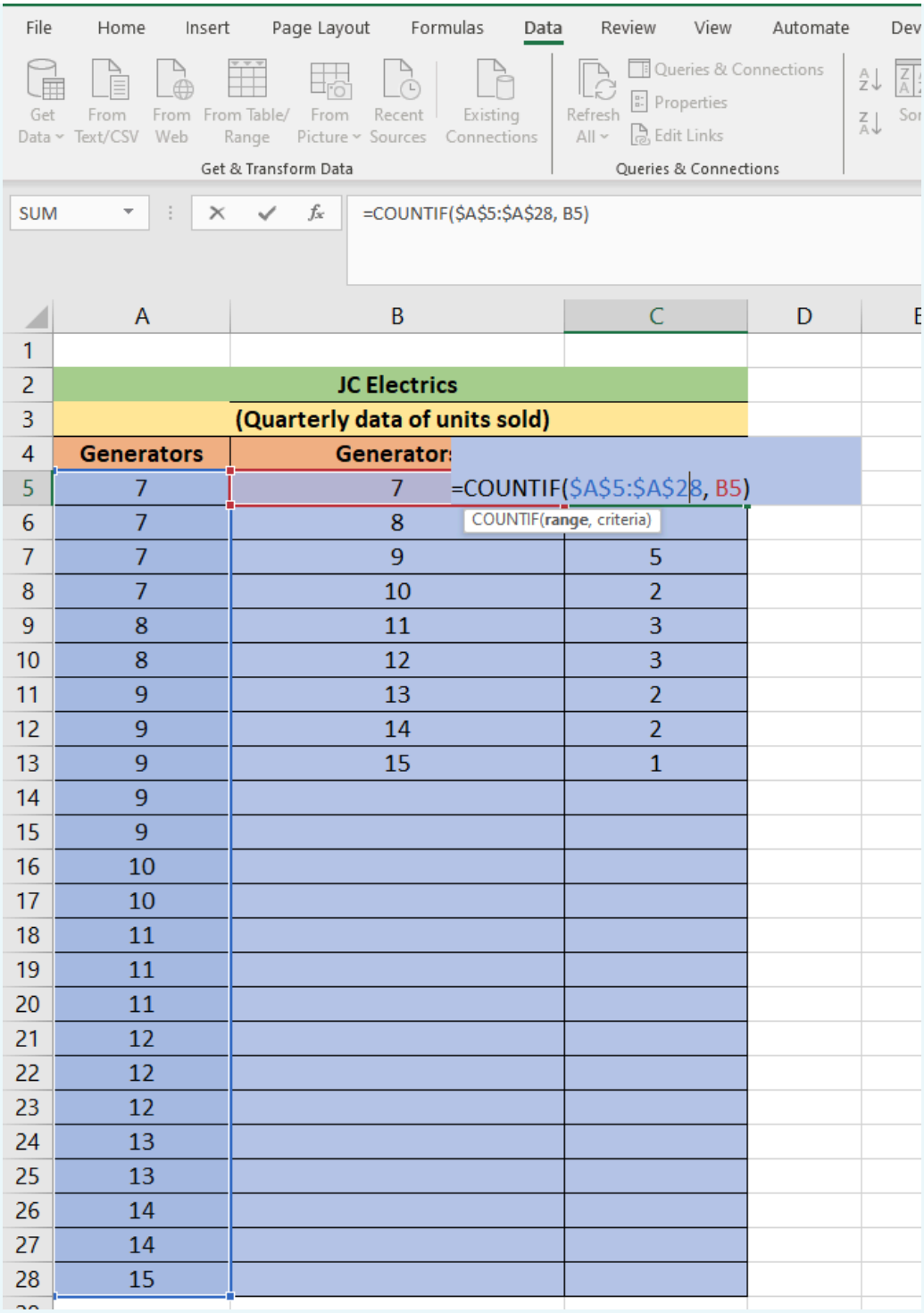


Figure 9 Calculating the frequency of ungroup data

- Check your answers in the Excel file: JC Electrics (solved).
- Optional: As further practice, you can now use the data for the other two products from JC Electrics 1 (Electric motors and Transformers) to identify the frequency of their products sold.
- You can check your answers in for solved files: JC Electrics (solved) Electric motors and JC Electrics (solved) Transformers.

Discussion

	B	C
2	JC Electrics	
3	(Quarterly data of units sold)	
4	Value	Frequency
5	7	4
6	8	2
7	9	5
8	10	2
9	11	3
10	12	3
11	13	2
12	14	2
13	15	1

Figure 10 Ungrouped frequency table

Producing the correct ungrouped frequency table provides you with information about the frequency of Generators sold with a particular value for the units sold. For example, 7 Generators are sold during 4 quarters, 8 Generators are sold during 2 quarters, 9 generators are sold during 5 quarters and so on...

As is mentioned above, ungrouped frequency tables are useful when you have a small set of data, and you want to easily observe the frequency of each value in the data set. However, if you have a large data set then a grouped frequency distribution table is the best option. You will learn about these in the next section.

2.2.3 Grouped frequency distribution tables

When you are summarising large masses of raw data, it is useful to represent the data in groups. The groups are commonly known as classes or class intervals. You might then want to determine the number of values belonging to each class or class interval; this is called *class frequency.* A tabular arrangement of data by class together with the corresponding class frequencies is called a *grouped frequency distribution table*. This is a more efficient way to find the trends within the data, but there is a possibility that the grouping process may sacrifice some of the original detail of the data.

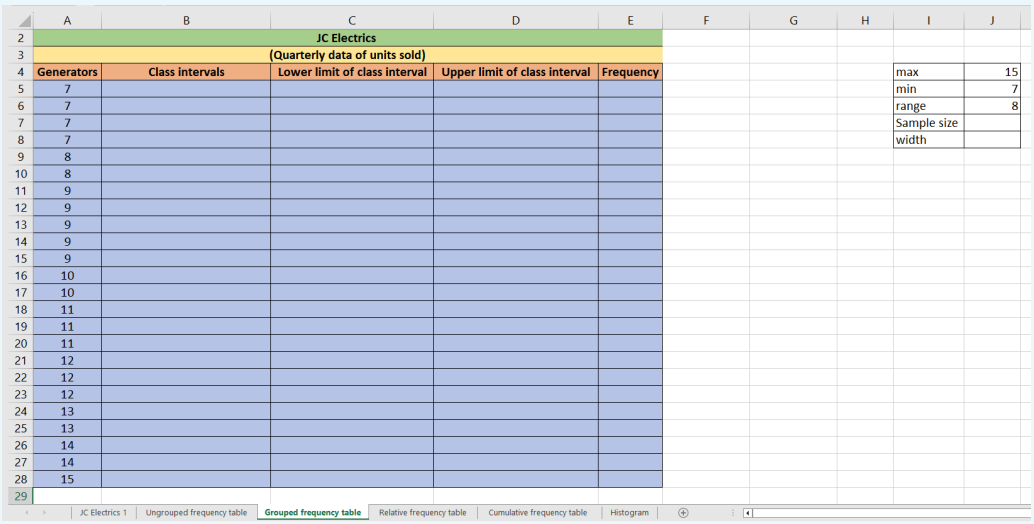
In the following activity, you will learn how to make a grouped frequency table in Excel.

Activity 3: How to make a grouped frequency distribution table in Excel

Allow approximately 45 minutes to complete this activity

In this activity, you will produce a grouped frequency table in Excel by following the instructions given below. Once you have produced the grouped frequency table in Excel, you can check your answer by clicking ‘Reveal discussion’.

- Open the Excel file JC Electrics template.
- Open the worksheet named ‘Grouped frequency table’. This is the third worksheet in the file JC Electrics template.
- In Column A you already have the data for Generators sorted.
- The tables for the following column are indicated as: Column B: Class intervals; Column C: Lower limit of class intervals, Column D: Upper limit of class intervals, Column E: Frequency and last column,
- Calculate the range. As mentioned before, this is the difference between the maximum and minimum value in the data set. For Generators, you can do this either by entering the formula **=MAX (**A5:A28**)-MIN (**A5:A28**),** or by calculating this on a separate section in the spreadsheet. For example, insert the formula =MAX(A5:A28) in cell J4. Insert formula =MIN(A5:A28) in cell J5. Calculate the rage in cell J6 by inserting the formula =J4-J5.



	A	B	C	D	E	F	G	H	I	J
2	JC Electrics									
3	(Quarterly data of units sold)									
4	Generators	Class intervals	Lower limit of class interval	Upper limit of class interval	Frequency				max	15
5	7								min	7
6	7								range	8
7	7								Sample size	
8	7								width	
9	8									
10	8									
11	9									
12	9									
13	9									
14	9									
15	9									
16	10									
17	10									
18	11									
19	11									
20	11									
21	12									
22	12									
23	12									
24	13									
25	13									
26	14									
27	14									
28	15									
29										

JC Electrics 1 | Ungrouped frequency table | Grouped frequency table | Relative frequency table | Cumulative frequency table | Histogram

Figure 11 Calculating the range

- Decide the class interval width. There are no firm rules on how to choose the width. However, the following formula is the most common method to calculate the width:

  MathJax failure: MathML - MathML must be formed by a single element

- You can round this value to a whole number or a number that is convenient to add (such as multiple of 10). For example, the width calculated in the given data set is 1.63, so will be taken as **2** (see Figure 12). The =SQRT(number) is the formula in Excel for calculating a square root.

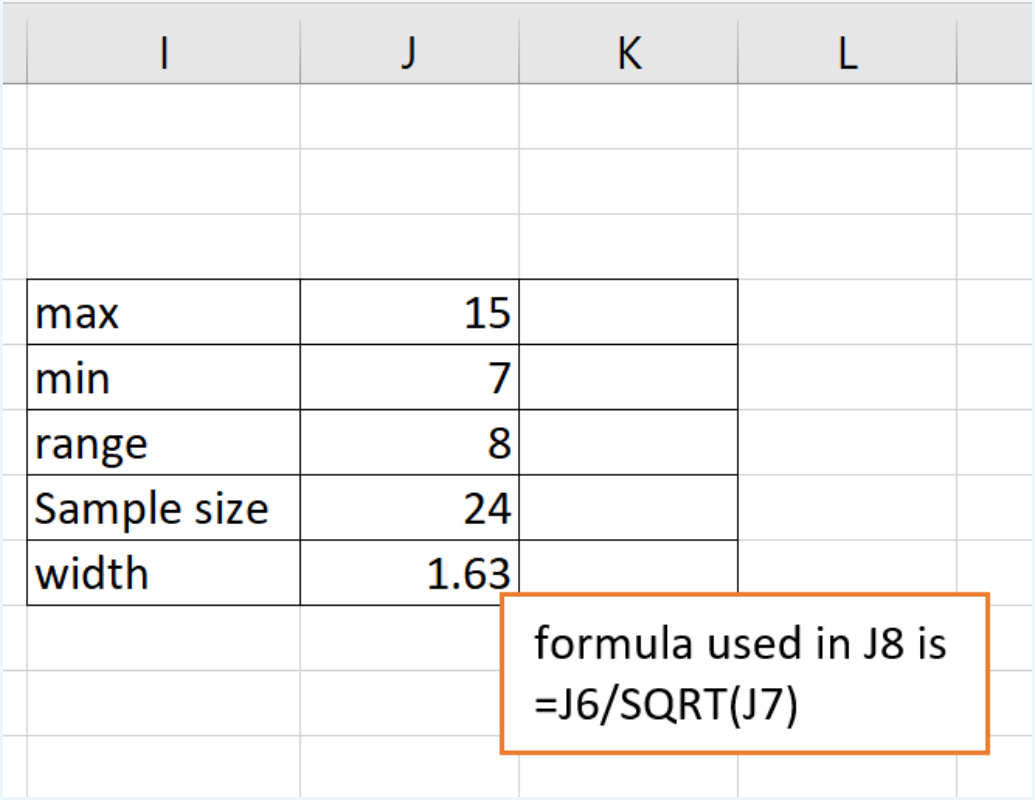


Figure 12 Calculating the width

- Decide the number of groups or class intervals into which data is to be distributed. Each class interval is defined by a lower limit and an upper limit. The lower limit of first class interval is the lowest value in the data set, in our case 7. Add the class interval width to find the lower limit of the next class interval. Keep adding the interval width to calculate more class intervals until you exceed the highest value. For example, in the given data set, you determined the class intervals width equals 2, so you should make the class intervals as 7_8, 9_10, 11_12, 13_14, 15_16.
- This means that the first class interval has lower limit of 7 and upper limit of 8. See Figure 13 below.

A	B	C	D
JC Electrics			
(Quarterly data of units sold)			
Generators	Class intervals	Lower Limit of class interval	Upper Limit of class interval
7	7_8	7	8
7	9_10	9	10
7	11_12	11	12
7	13_14	13	14
8	15_16	15	16
8			
9			
9			
9			
9			
9			
10			
10			
11			
11			
11			
12			
12			
12			
13			
13			
14			
14			
15			

Figure 13 Making the class intervals

- Next step is to calculate the frequency. This will be displayed in the section E5:E9. In cell E5 insert the formula =FREQUENCY(A5:A28,D5:D9). This will return the values 6 in the first cell followed by 7, 6, 4, 1 and 0.
- As you see, the Upper limit of class interval is selected here, therefore the returned value 6 indicates, that 6 times up to 8 units of Generators were sold. The next calculation in cell E6 shows a value of 7, therefore there 7 times Generators were sold with up to 10 units, but above 8 and so on.
- Check your answers in the Excel file: JC Electrics (solved).
- Optional: As further practice, you can now use the data for the other two products from JC Electrics 1 (Electric motors and Transformers) to identify the class intervals, lower limit of class interval, upper limit of class interval and frequency.
- You can check your answers in for solved files: JC Electrics (solved) Electric motors and JC Electrics (solved) Transformers.

Discussion

E5 =FREQUENCY(A5:A28,D5:D9)

	A	B	C	D	E	F
1						
2	JC Electrics					
3	(Quarterly data of units sold)					
4	Generators	Class intervals	Lower limit of class interval	Upper limit of class interval	Frequency	Cumulative frequency
5	7	7_8	7	8	6	
6	7	9_10	9	10	7	
7	7	11_12	11	12	6	
8	7	13_14	13	14	4	
9	8	15_16	15	16	1	
10	8				0	
11	9					
12	9					
13	9					
14	9					
15	9					
16	10					
17	10					
18	11					
19	11					
20	11					
21	12					
22	12					
23	12					
24	13					
25	13					
26	14					
27	14					
28	15					

Figure 14 Calculating the frequency by using an array formula

Figure 14 shows the results of using the array formula to calculate frequency. It is important to be aware than any error entered may result in an incorrect grouped frequency table and provide false information about the business.

As mentioned above, a grouped frequency table is the best option to visualise the frequency of values in a large data set. However, if you are interested to know the proportion of a particular value in relation to the total number of values in the data set, then a relative frequency distribution table is the better option. In the next session, you will learn how to produce a relative frequency table in Excel.

2.2.4 Relative frequency distribution tables

Relative frequency distribution is another type of frequency distribution. This type of distribution tells us the proportion of each value or class interval of a variable in a sample. In other words, relative frequency distribution describes the number of times a particular value occurs in relation to the total number of values. You can use this type of frequency distribution for any type of variable when you are more interested in comparing frequencies than the actual number of observations.

For example, Team A has won 6 football games from a total of 12 football games played. The frequency of winning is 6 and the relative frequency of winning is 50% (i.e. 60/12=0.5).

You will learn how to make a relative frequency table in Excel in the following activity.

Activity 4: How to make a relative frequency table in Excel

Allow approximately 45 minutes to complete this activity

In this activity, you will build a relative frequency table using the ungrouped frequency distribution table from Activity 2. Use the ungrouped frequency table that you made in the previous activities.

Use the file JC Electrics template, sheet Relative frequency table. In Column D insert the name Relative Frequency. Data in Columns A, B and C is the same as calculated for the Ungrouped frequency table. In Column C, Cell C14 insert the formula =SUM(C5:C13). This will calculate the sum of all frequencies calculated. The returned value is 24.

In Column D, Cell D5 insert formula =C5/$C$14. The returned value is 0.166666667. If you prefer to see this value in percentages, you can select the cells in column D, right mouse click, select Format Cells, followed by Percentages. For a decimal place of value 2, it will return a value for the relative frequency as 16.66%. If this value is rounded up, the relative frequency in this case is 17%.

You can use the same formula to continue calculate the relative frequency.

Once you have made the relative frequency distribution table in Excel, check your answer by clicking ‘Reveal discussion’ below.

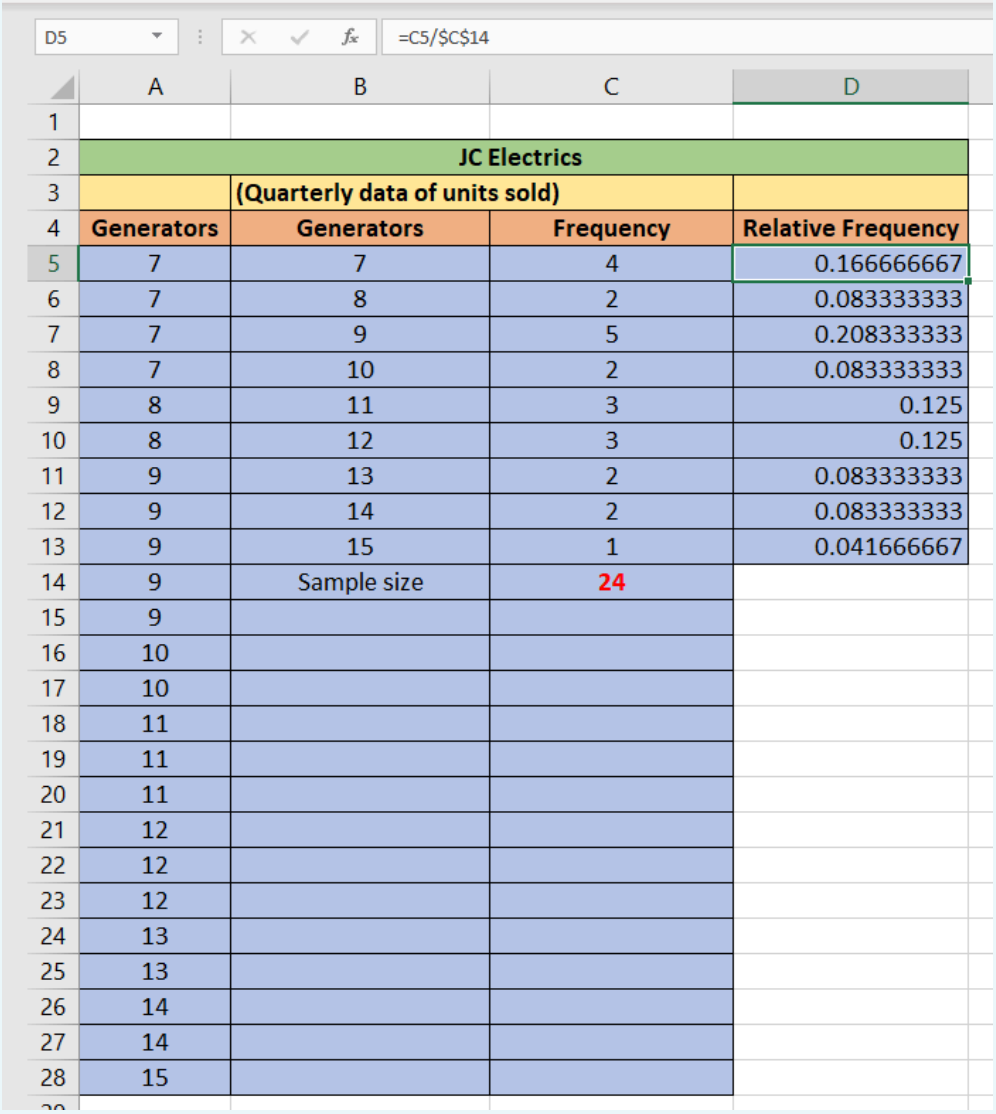
The ungrouped frequency distribution table consists of three columns. Column A is labelled ‘Generators’, Column B is labelled ‘Generators’ and Column C is labelled ‘Frequency’. Add a fourth column to the table for the Relative Frequency.

To calculate the relative frequencies, you need to divide each frequency by the sample size (frequency / sample size). You can calculate the sample size by taking the sum of all the frequencies in Column C, which is 24.

- Check your answers in the Excel file JC Electrics (solved).

Optional: As further practice, you can now use the data for the other two products from JC Electrics 1 (Electric motors and Transformers) to identify the relative frequency of products sold. You can check your answer in the Excel files: JC Electrics (solved) Electric motors and JC Electrics (solved) Transformers.

Discussion



	A	B	C	D
1				
2	JC Electrics			
3		(Quarterly data of units sold)		
4	Generators	Generators	Frequency	Relative Frequency
5	7	7	4	0.166666667
6	7	8	2	0.083333333
7	7	9	5	0.208333333
8	7	10	2	0.083333333
9	8	11	3	0.125
10	8	12	3	0.125
11	9	13	2	0.083333333
12	9	14	2	0.083333333
13	9	15	1	0.041666667
14	9	Sample size	24	
15	9			
16	10			
17	10			
18	11			
19	11			
20	11			
21	12			
22	12			
23	12			
24	13			
25	13			
26	14			
27	14			
28	15			

Figure 15 Calculating the relative frequency of ungrouped data

We calculate the relative frequency table whenever we are interested in knowing what proportion or percentage of the observations in a data set take on a certain value or fall within a range of values. Any error in the calculation may lead a business to make the wrong decision based on faulty information.

In the next section, you will learn how to make cumulative frequency distribution tables in Excel.

2.2.5 Cumulative frequency distribution tables

Cumulative frequency distribution is the fourth type of frequency distribution table. This type of frequency distribution can be used for **ordinal** or **quantitative variables**, especially when you want to understand how often observations fall below certain values.

For example, Company A sells 250 books in the first week, 150 books in the second week and 400 books in the third week. The cumulative number of books sold in the second week by Company A is 400 books (250 books in the first week + 150 books in the second week). The cumulative number in the third week is 800 books (250 books in the first week + 150 books in the second week + 400 books in the third week).

In the following activity, you will learn how to build a cumulative frequency distribution table in Excel.

Activity 5: How to make a cumulative frequency distribution table in Excel

Allow approximately 45 minutes to complete this activity

In this activity, you will build a cumulative frequency distribution table using the grouped frequency distribution table in Activity 3. Once you have built the cumulative frequency distribution table, you can check your answer by clicking 'Reveal discussion' below.

Use the data from the grouped frequency distribution table from Activity 3 in the template provided in JC Electrics template file. This table consists of six columns. Column A is labelled Generators, Column B is labelled Class intervals, Column C is labelled Lower limit of class interval, Column D is labelled Upper limit of class interval and Column E is labelled Frequency. Column F considers the cumulative frequency. The cumulative frequency is calculated by adding each frequency from a frequency distribution table to the sum of its predecessors. The last value will always be equal to the total for all observations, since all frequencies will already have been added to the previous total. Therefore, the expected final answer is 24.

To calculate the cumulative frequency, in Cell F5 enter formula = =E5, this is the same value of the first frequency. Now, in Cell F6, to calculate the cumulative frequency, you will add the new frequency value to the previous one. In Cell F6, enter the formula =F5+E6. In Cell F7, the formula will be =F6+E7. Continue this step up to Cell F9.

- Check your answers in the Excel file: JC Electrics (solved).
- Optional: As further practice, you can now use the data for the other two products from JC Electrics 1 (Electric motors and Transformers) to identify the Cumulative frequency of products sold.

Discussion

	A	B	C	D	E	F
1						
2	JC Electrics					
3	(Quarterly data of units sold)					
4	Generators	Class intervals	Lower limit of class interval	Upper limit of class interval	Frequency	Cumulative frequency
5	7	7_8	7	8	6	6
6	7	9_10	9	10	7	13
7	7	11_12	11	12	6	19
8	7	13_14	13	14	4	23
9	8	15_16	15	16	1	24
10	8				0	
11	9					
12	9					
13	9					
14	9					
15	9					
16	10					
17	10					
18	11					
19	11					
20	11					
21	12					
22	12					
23	12					
24	13					
25	13					
26	14					
27	14					
28	15					
29						
30						

JC Electrics 1 | Ungrouped frequency table | Grouped frequency table | Relative frequency table | Cumulative frequency table

Figure 16 Calculating the cumulative frequency

A cumulative frequency table is used to determine the number of observations that lie above or below a certain value in a data set. Any error in the table may not reflect the reality of the data.

In the next section, you will learn how to draw histograms in Excel.

2.3 Histograms: a graphical visualisation of frequency tables

A histogram is a very popular visualisation tool to summarise the distribution of continuous or discrete data. In a histogram, the variable is divided into intervals called ‘bins’. You then count the number of observations in each bin. The horizontal *x*-axis displays the ‘bins’ and vertical *y*-axis displays the number of observations in each bin. Histograms can help you to see whether the data is clustered around certain values or whether there are many small or many large values. A typical histogram in Excel looks like the following bar chart. In a correct histogram, there are no spaces or gaps between the bars.

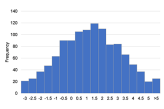


Figure 17 A typical histogram in Excel

In the following activity, you will learn how to plot a histogram in Excel.

Activity 6: Using Excel to draw a histogram

Allow approximately 60 minutes to complete this activity

In this activity, you will learn how to produce a histogram in Excel by following the instructions that are given below. Once you have produced the histogram in Excel, you can check your answer by clicking 'Reveal discussion' below.

- Open the Excel file called JC Electrics template, sheet Histogram which contains the quarterly data of number of generators sold. Column A contains information about the number of generators sold per quarter.
- Use the same data and table format as calculated in the Group frequency table.
- There are many ways to calculate the width of the bin in Excel. One of the easiest ways to calculate this is similar as the width of the bin or class intervals. As previously indicated, this would be the range divided by square root of sample size, which was 1.63.
- Click on 'Data Analysis' in the 'Data' ribbon. This will bring up a list of some of the statistical analyses that you can perform in Excel.
- Select 'Histogram' and click 'OK'.
- Specify the input range as A5:A28 and the bin range as D5:D9. Excel automatically allocates a $ sign for each selection as in Figure 18.
- Tick the box 'Chart Output' and specify the output location as H12, as shown below.

A	B	C	D	E
		JC Electrics		
		(Quarterly data of units sold)		
Generators	Class intervals	Lower Limit of class interval	Upper Limit of class interval	Frequenc
7	7_8	7	8	6
7	9_10	9	10	7
7	11_12	11	12	6
7	13_14	13	14	4
8	15_16	15	16	1
8				
9				
9				
9				
9				
9				
10				
10				
11				
11				
11				
12				
12				
12				
13				

Figure 18 A frequency distribution table in Excel

Click 'OK'. Excel will put the histogram next to your frequency table.

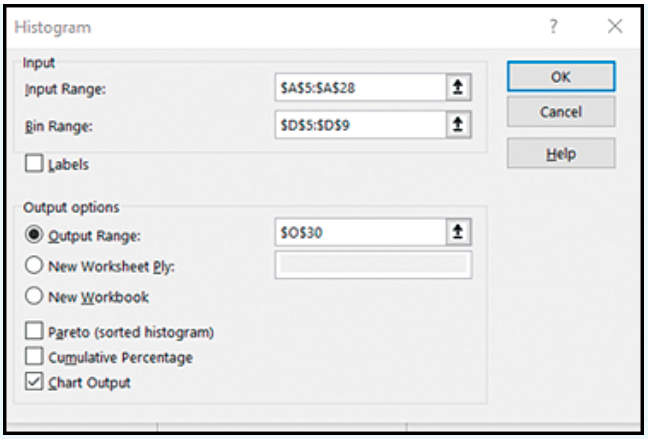


Figure 19 Histogram dialog box in Excel

Click ‘OK’. Excel will put the histogram next to your frequency table.

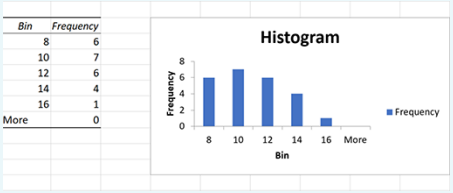


Figure 20 Histogram of number of units sold (Generators)

- To remove the space between the bars, right click a bar, click **Format Data Series**, and change the Gap Width to 0%.
- To add borders, right click a bar, click **Format Data Series**, click the **Fill & Line** icon, click **Border**, and select a colour.
- Now click ‘Reveal discussion’ to compare what you have made against the answer.

Discussion

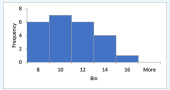


Figure 21 Histogram showing units sold of generators

Working through the steps given above should return the completed histogram shown in Figure 20.

2.4 Frequency density

Frequency density is defined as the frequency per unit of the data in each class. Frequency density is calculated by dividing the frequency by the class width (the class width is the difference between the upper limit of the class interval and the lower limit of

the class interval). Frequency density allows for a meaningful comparison of different classes where the class widths may not be equal.

$$\text{Frequency density} = \frac{\text{Frequency}}{\text{Class width}}$$

The frequency density **gives us the ratio of the frequency of a class to its width**. Frequency density is used to plot a frequency density histogram; here, we plot frequency density instead of frequency on the *y*-axis. Frequency density gives us the total area of bars and tells us about the frequency in the histogram (rather than the height).

We calculate frequency density when we have a set of grouped data that consists of unequal widths of class intervals. For example, see the following Excel worksheet, which gives us information about the ages of people playing cricket.

	A	B	C	D
1	**Age**	**Frequency**	**Class width**	**Frequency density**
2	$0 \leq x < 10$	10		
3	$10 \leq x < 15$	9		
4	$15 \leq x < 20$	6		
5	$20 \leq x < 30$	4		
6	$30 \leq x < 50$	7		

Figure 22 Information about the age of people playing football in an Excel file

To calculate the frequency densities:

- In Column C, find the width of the class intervals by finding the difference of upper and lower bounds/limits. (For example, and so on.)

	A	B	C	D
1	**Age**	**Frequency**	**Class width**	**Frequency density**
2	$0 \leq x < 10$	10	10	
3	$10 \leq x < 15$	9	5	
4	$15 \leq x < 20$	6	5	
5	$20 \leq x < 30$	4	10	
6	$30 \leq x < 50$	7	20	

Figure 23 Calculating the class width

- Then, in Column D, divide the frequency of each class interval by its width following the Frequency density formula given above. Even though the frequency in the first age bracket is higher, because the interval is twice as large as in the second age bracket, it could be misleading. From looking at the frequency density column, we can see that cricket is most popular in the 10-15 age category, with a frequency density of 1.8.

A	B	C	D
Age	Frequency	Class width	Frequency density
$0 \leq x < 10$	10	10	1
$10 \leq x < 15$	9	5	1.8
$15 \leq x < 20$	6	5	1.2
$20 \leq x < 30$	4	10	0.4
$30 \leq x < 50$	7	20	0.35

Figure 24 Calculating the frequency density

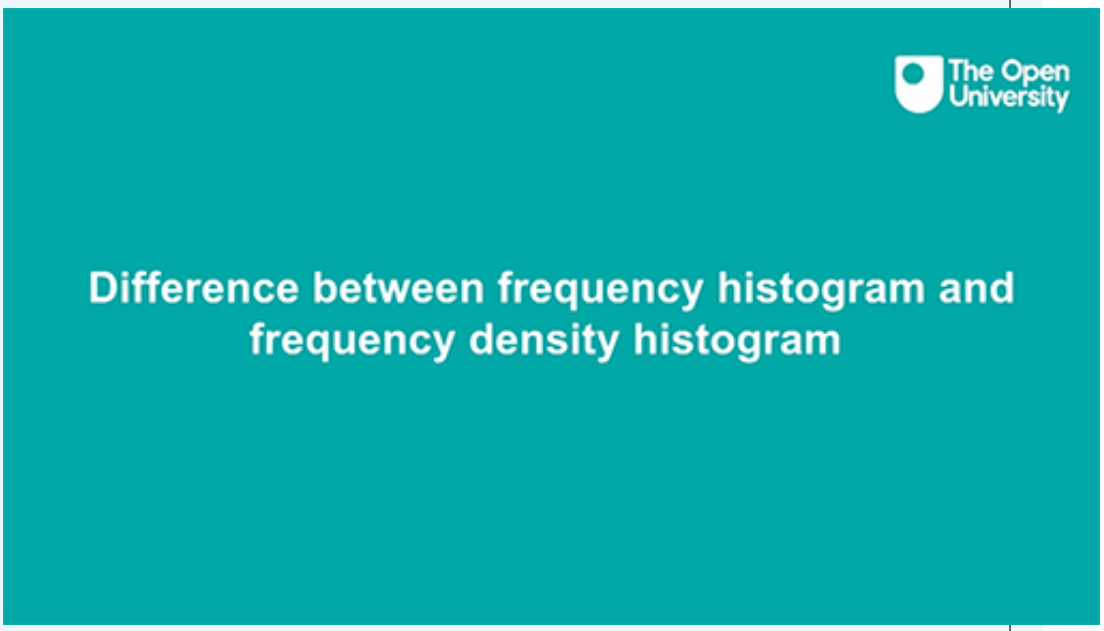
In the activity below, you will test your knowledge of the difference between a frequency density histogram and a frequency histogram.

Activity 7: The difference between frequency histograms and frequency density histograms

Allow approximately 35 minutes to complete this activity

Watch the video below and note down in the box the difference between frequency histograms and frequency density histograms.

Video content is not available in this format.
Video 1



Provide your answer...

3 Bivariate data

Bivariate data refers to an instance in which two separate variables are examined and compared. For example, a performance manager may be interested to know how long (time) it takes employees to complete their assigned task (output). In other words, the interest is to measure the efficiency of the employees. In this example, the performance manager is examining two variables: time and output.

Bivariate data is collected to explore the relationship between two variables and then uses this relationship to inform future decisions. One of the main aims of the researcher is to find out whether changes in one variable might be caused by changes in the other variable. This type of research involves two basic types of variables: independent variables and dependent variables.

Independent variables: An independent variable is **one that stands alone and is not changed by the other variable you are trying to measure**. The researcher changes the independent variable to see the effect it will have on the dependent variable.

Dependent variables: A dependent variable is **the one that changes because of independent variable manipulation**. It is the outcome you are interested in measuring, and it 'depends' on your independent variable. In statistics, dependent variables are also called response variables (as they respond to a change in another variable).

For example, say a researcher is interested to know if mature students' performance in a maths class changes based on the time of the class. To answer this question, the researcher measures mature students' performance in a morning class and an evening class. The study finds that mature students perform better in the evening class than in the morning class.

What are the independent and dependent variables in the example above? The independent variable is the time of the class, and the dependent variable is mature students' performance in maths, as it might change in relation to the independent variable.

In the next activity, you will expand your knowledge of bivariate data.

Activity 8: Bivariate data

Allow approximately 30 minutes to complete this activity

Watch the following video and make notes on bivariate data in the free response box below.

Video content is not available in this format.
Video 2

Provide your answer...

Bivariate data can be visualised using contingency tables and scatter diagrams. In the next section you will learn about contingency tables

3.1 Contingency tables

A contingency table is called a cross table or two-way-table and counts observations for each unique combination of values in two variables. It is a table of data in which the row entries tabulate the data according to one variable and the column entries tabulate the data according to another variable.

The tool best suited for a data set will depend on the variable's scale of measurement. The most important distinction here is whether a variable is discrete or continuous. If it is discrete, a convenient method to summarise bivariate data is a contingency table. If it is continuous, a scatter diagram can be more useful in visualising the data set.

Contingency tables are used in statistics to understand the relationship between categorical variables. For example, say we want to summarise the following sample of firms regarding their industry sector and size:

Table 7 Sample of firms

Firm	Sector	Employees
1	Technology	<50
2	Food	50+
3	Technology	<50

4	Food	<50
5	Food	<50
6	Food	50+
7	Technology	<50
8	Technology	<50
9	Technology	50+
10	Food	50+

See the following cross table, which elegantly summarises the information of our sample data.

Table 8 Cross table for sample firms

Sector	<50 Employees	50+ Employees	Total firms
Technology	4	1	5
Food	2	3	5
Total	6	4	10

The cross table shows that there are 2 firms in the food manufacturing sector that have less than 50 employees. However, the cell 50+ shows that there are 3 firms in the food manufacturing sector that have more than 50 employees. The sum of the total food manufacturing firm is 5 which is the 50% of the grand total.

Contingency tables vary in size and type because the size of the contingency table depends on the number of different ways in which data is categorised.

There is no formula to draw a contingency table in Excel. However, analysts use a PivotTable to build contingency tables. A PivotTable is considered a powerful statistical tool to summarise bivariate and multivariate data sets in an Excel spreadsheet or database table and obtain the desired report. This tool does not actually change the spreadsheet or database itself; it simply pivots or turns the data to view it from different perspectives. Researchers and analysts use PivotTables especially when they have large amounts of data that would be time consuming to calculate by hand. A PivotTable can perform a few data processing functions such as identifying sums, averages, ranges or outliers. It then arranges this information in a simple and meaningful way that draws attention to key values.

3.2 Scatter diagrams

A scatter diagram is another way to visualise a quantitative bivariate data set. This is a two-dimensional diagram or graph with one variable on the *x*-axis (the independent variable) and the other variable on the *y*-axis (the dependent variable). We then plot the corresponding point on the diagram.

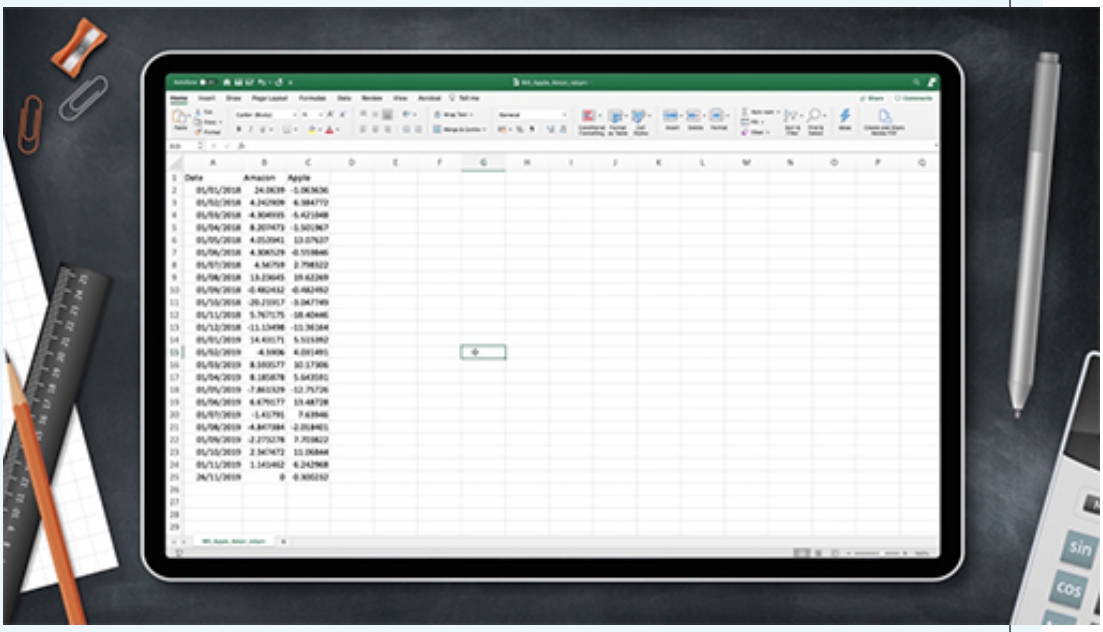
In the next activity, you will learn how to produce a scatter diagram in Excel.

Activity 9: Drawing scatter diagrams

Allow approximately 60 minutes to complete this activity

In this activity you will learn how to make an ungrouped frequency table in Excel, either by watching the screencast or by following the instructions that are given below.

Video content is not available in this format.
Video 3



Look at the following example, which shows a data set relating to the temperature on days in June, and the number of Pepsi drinks sold in a small shop.

Table 9 Temperature and number of drinks sold

Temperature (X)	12	14	15	17	22	13	20	23
Pepsi (Y)	12	16	16	19	32	10	24	40

Now we draw the scatter diagram (see file: Pepsi data.xlsx). In the example above, temperature is the independent variable and number of Pepsis is the dependent variable, as the sale of Pepsi is affected by changes in the temperature. Hence, we plot temperature on the *x*-axis and Pepsi on the *y*-axis.

To find out if there is any relationship between variable X (Temperature) and variable Y (Pepsi), execute the following steps in Excel.

- Select the range A1:B9

	A	B
1	X (Temperature)	Y (Pepsi)
2	12	12
3	14	16
4	15	16
5	17	19
6	22	32
7	13	10
8	20	24
9	23	40

Figure 25 Spreadsheet of Pepsi sold and Temperature

- On the **Insert** tab, in the **Charts** group, click the **Scatter** symbol.

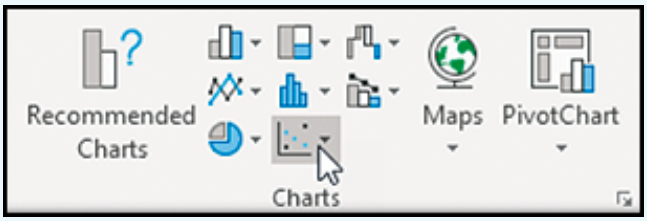


Figure 26 How to select the Scatter symbol

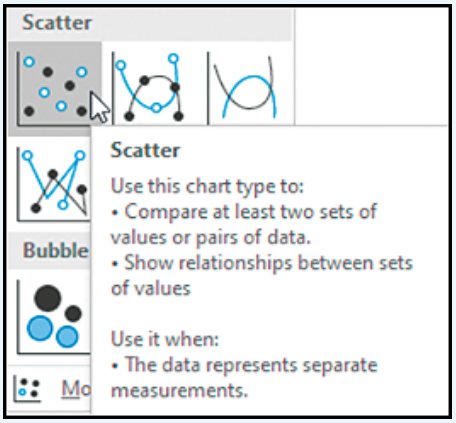


Figure 27 How to select the Scatter chart type

- Optional – returning to our example JC Electrics, draw a scatter diagram for Generators

Discussion

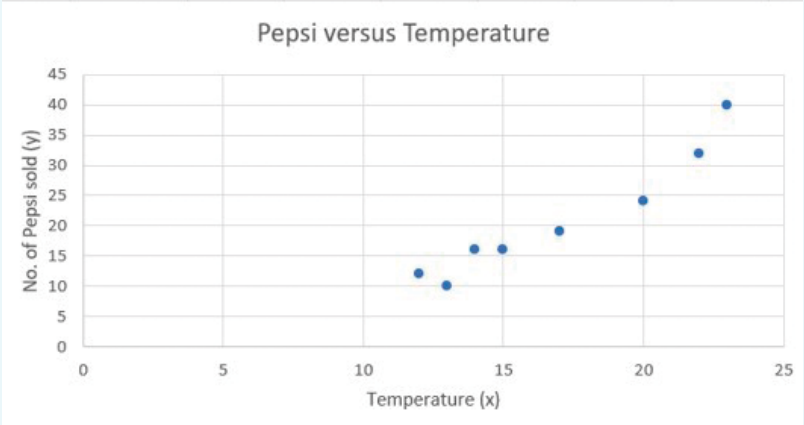


Figure 28 Temperature versus Pepsi sales

The final image shows the positive relationship between the number of Pepsis sold and the temperature. This shows that as the temperature increases, sales of Pepsi also increase.

For JC Electrics Generators, a scatter diagram for Generators will show as in Figure 28. This presents a slight overall increase in sales, with peaks in sales in quarter 6 with a value of 14, and quarter 10 and 15 with a value of sales of 13 generators, as well as in quarter 21 with a value of 14 and quarter 22 with a value of 15.

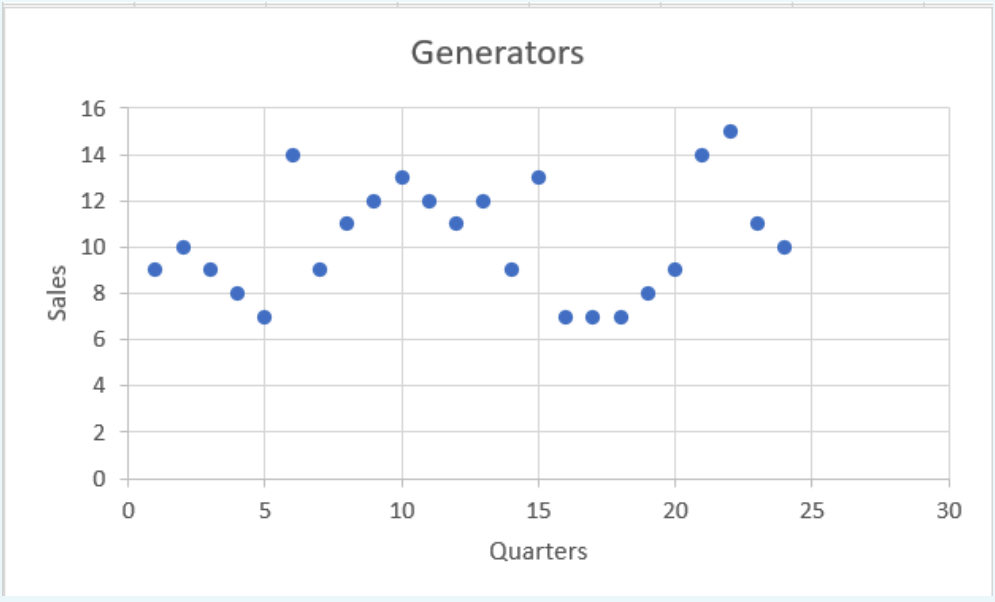


Figure 29 Generator sales scatter diagram

4 Conclusion

You have started to familiarise yourself with the spreadsheet software Excel, which is widely used in workplaces, and useful in many different fields and contexts, such as business, medicine, marketing, tax and auditing, accounting and finance.

You have also studied the basics of data analysis. The focus here was on the several ways to visualise and summarise data using tools available in Microsoft Excel, such as frequency tables, histograms, and scatter diagrams or plots. The main objective of data analysis and statistical modelling is to help make more evidence-based decisions. The various data visualisation tools studied in this session are only the first step toward starting the decision-making process using data.

A second OpenLearn course on data analysis, Data analysis: hypothesis testing, is now also available should you wish to take your studies further.

This OpenLearn course is an adapted extract from the Open University course B126 *Business data analytics and decision making*.

Acknowledgements

This free course was written by the B126 Open University course team.

Don't miss out

If reading this text has inspired you to learn more, you may be interested in joining the millions of people who discover our free learning resources and qualifications by visiting The Open University – www.open.edu/openlearn/free-courses.