B126_1

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 2023.

Intellectual property

Contents

Introduction

The objective of this course is to explore the ways to visualise data sets, such as univariate and bivariate data, in Excel and to familiarise yourself with the functions used in Excel to explore the relationship between variables. Univariate data here refers to data that consists of one variable, and bivariate data refers to data that consists of two variables.

Figure 1 Making sense of visualising and exploring data

This course requires Microsoft Excel in some activities. Therefore, this course should be completed on a desktop or laptop rather than a mobile device. If you do not have access to Microsoft Excel, there are various other free options – such as Google Sheets, Apple Numbers or LibreOffice.

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

Before making any decision for a business, it is usually a good idea to get a clear picture of the data, which can provide you with an overview of the relevant information. It is good practice, therefore, to organise and present data in such ways that make it useful for decision making and problem solving. Microsoft Excel is used widely for data analysis in a professional context. Researchers and analysts alike use this tool for various applications in the real world, such as in business, medicine, academia, tax and auditing, marketing, accounting and finance. Moreover, it is flexible enough to be used with many types of data, irrespective of whether it is qualitative or quantitative data.

In this course, you will make extensive use of Excel spreadsheets. In this section you will 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 Using Excel

You can run Excel by double-clicking on the Excel icon on your desktop or laptop (or select **Excel 2013** or your version 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 to 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. You will be using this style, as opposed to the **R1C1** reference style, which labels both rows and columns with numbers. However, it is sometimes useful to think of a value or formula as being in a cell with a specific row number and column number.

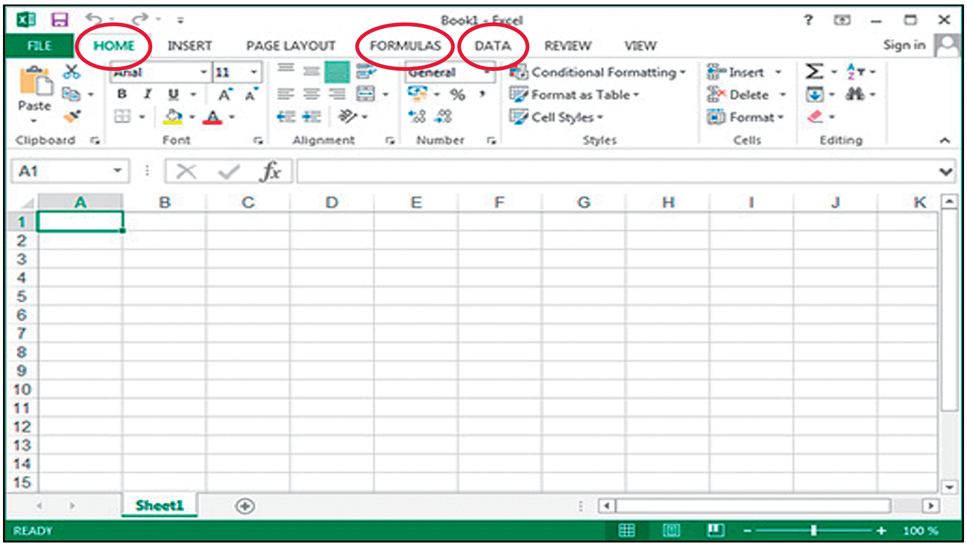


Figure 2 An empty worksheet in Excel

The latest versions of Excel, including Excel 2013, are organised using a **ribbon** interface, which consists of a collection of icons for each tab. The screenshot above in Figure 2 shows the **Home** ribbon with several icons for general editing such as font size, text alignment or cell styles. In this section, you will use the **Insert**, **Formulas** and **Data** tabs in Excel. You may click on each tab and familiarise yourself with the large main icons.

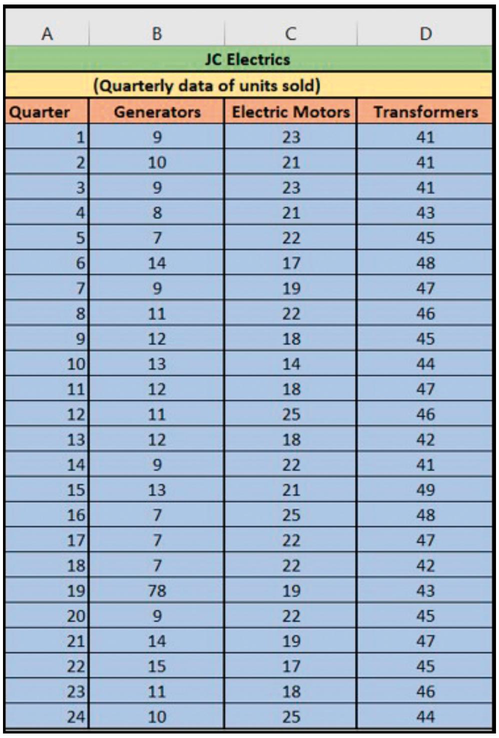
You should not be worried if you are not very familiar with Excel, or do not know how to use the various icons, tabs and functions yet. The aim of this course is to gradually build up your familiarity with the software.

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. For this course, you will be using a pre-populated worksheet for JC Electrics, a company manufacturing heavy machinery, such as generators, transformers and electric motors. Download the file at the link below. You can then save it to your computer so you can return to it at different points in the course. The file will also be linked at future points in the course.

- Download the file JC Electrics.
- Click on **File**, then **Open** and navigate to the folder in which you saved this file and open the file by double-clicking it.

By opening the worksheet, you should see quarterly data of units sold by 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

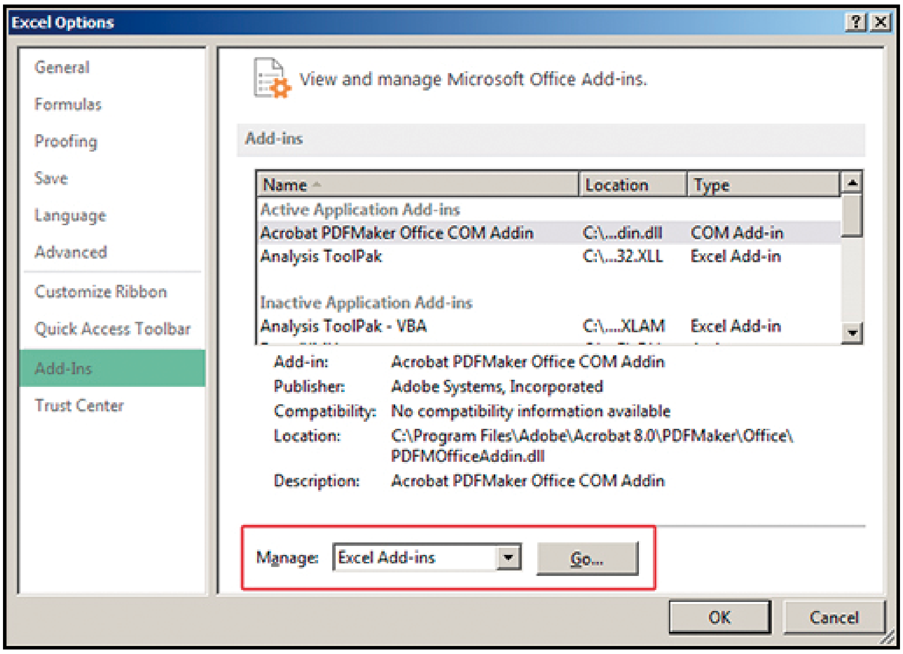
1.3 Adding the Data Analysis ToolPak in Excel

In this section, you are going to learn how to add the Data Analysis ToolPak in Excel and what its purpose is. If you are using another spreadsheet software, you may skip this section. Alternatively, similar functionality is usually available within the software, and all tasks within the Excel ToolPak can be carried out using other functions in Excel as well.

The Data Analysis ToolPak is a Microsoft Excel add-in function that provides you with a set of statistical tools for analysing data efficiently and effectively. 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 load or activate it before you can use it.

To load the Data Analysis ToolPak:

- Open the Excel worksheet.
- 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**.
- In the **Add-Ins** dialog box, select the **Analysis Tool Pak** check box, and then click **OK**.



View and manage Microsoft Office Add-ins.

Add-ins

Name	Location	Type
Active Application Add-ins		
Acrobat PDFMaker Office COM Addin	C:\...din.dll	COM Add-in
Analysis ToolPak	C:\...32.XLL	Excel Add-in
Inactive Application Add-ins		
Analysis ToolPak - VBA	C:\....XLAM	Excel Add-in

Add-in: Acrobat PDFMaker Office COM Addin
Publisher: Adobe Systems, Incorporated
Compatibility: No compatibility information available
Location: C:\Program Files\Adobe\Acrobat 8.0\PDFMaker\Office\PDFMOfficeAddin.dll
Description: Acrobat PDFMaker Office COM Addin

Manage: Excel Add-ins Go...

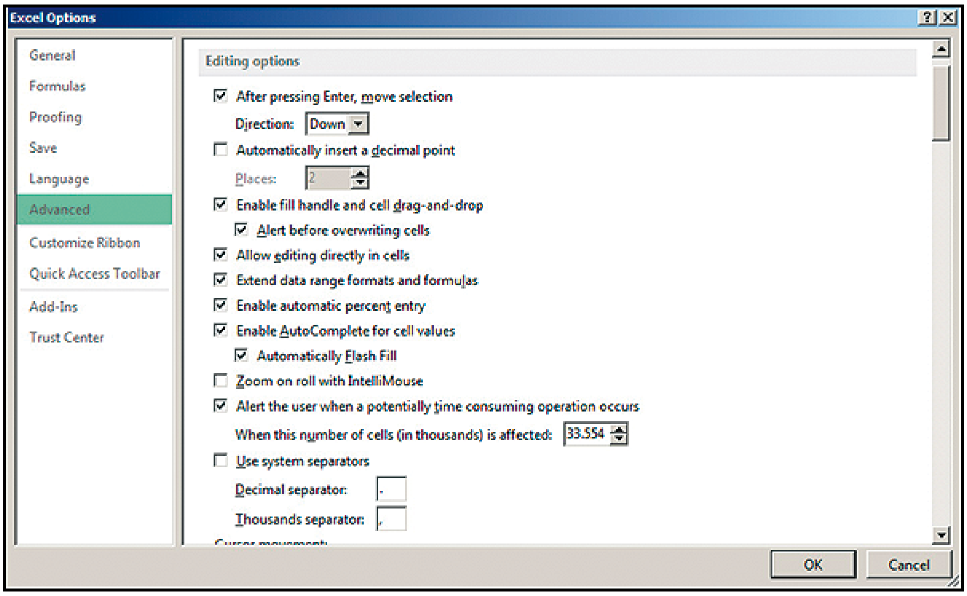
Figure 4 Activating the Data Analysis ToolPak

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 your Excel spreadsheet, click on **File,** select **Options** and then select the **Advanced tab**; the results of this are shown in Figure 5.



Editing options

☑ After pressing Enter, move selection

Direction: Down

□ Automatically insert a decimal point

Places: 2

☑ Enable fill handle and cell drag-and-drop

☑ Alert before overwriting cells

☑ Allow editing directly in cells

☑ Extend data range formats and formulas

☑ Enable automatic percent entry

☑ Enable AutoComplete for cell values

☑ Automatically Flash Fill

□ Zoom on roll with IntelliMouse

☑ Alert the user when a potentially time consuming operation occurs

When this number of cells (in thousands) is affected: 33.554

□ Use system separators

Decimal separator: .

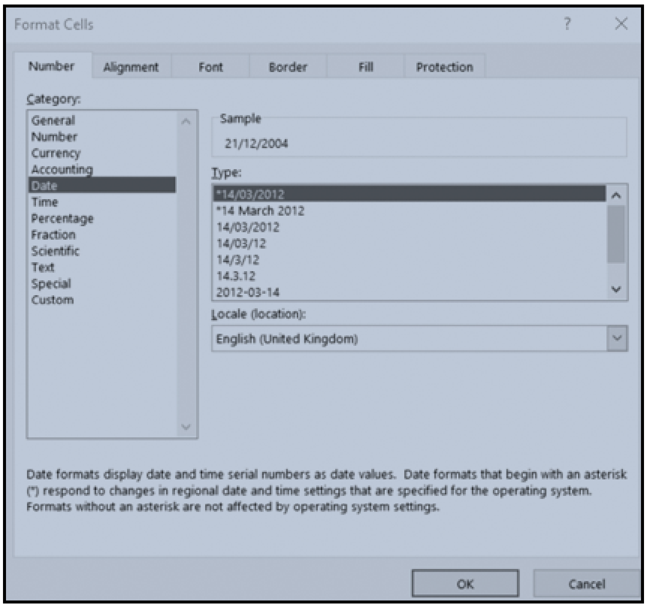
Thousands separator: ,

Figure 5 Advanced options in Excel

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

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.

Figure 6 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**.



Category:

General
Number
Currency
Accounting
Date
Time
Percentage
Fraction
Scientific
Text
Special
Custom

Sample

21/12/2004

Type:

*14/03/2012
*14 March 2012
14/03/2012
14/03/12
14/3/12
14.3.12
2012-03-14

Locale (location):

English (United Kingdom)

Date formats display date and time serial numbers as date values. Date formats that begin with an asterisk (*) respond to changes in regional date and time settings that are specified for the operating system. Formats without an asterisk are not affected by operating system settings.

Figure 6 Changing the date format

1.5 Using shortcut keys in Excel

One of the most beneficial things in Excel is being able to control the user interface without using the mouse. This will substantially speed up your Excel projects, especially when working under pressure in a professional environment.

Some of the most useful shortcuts are listed in the tables 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.
Ctrl + F	Display the ‘Find’ dialog box.
Ctrl + H	Display the ‘Find and Replace’ dialog box.
Ctrl + Tab	Set focus on next workbook if multiple workbooks are open.

Table 2 Data selection

Shift + Space	Select the entire row at the cursor position.
Ctrl + Space	Select the entire column 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.
Ctrl + Shift + Page Up	Select the current and previous sheet in a workbook. This is useful if you have similar worksheets and want to edit cells in all of them at the same time.
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.
Shift + F8	Add another range to the selected range of cells.
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.
F2	Edit current cell.
F4	Repeat last formatting action.
Alt + Enter	Start a new line in the same cell when entering text.
Ctrl + D	Copy above cell down.
Ctrl + ‘+’	Insert row/column.
Ctrl + ‘-’	Delete row/column.
Ctrl + 1	Show the ‘Format cells’ dialog.
Shift + F11	Insert new worksheet.

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.
F12	Display the ‘Save As’ dialog.
Ctrl + S	Save the current workbook. Extremely useful for the occasional power outage or computer crash.
Ctrl + F1	Minimise or show the ribbon.

1.6 Use of Excel spreadsheets

Excel spreadsheets are used to manage, organise and present data in a systematic way. They are particularly useful when there is a specific relationship between results in different cells. 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' exercises.

Excel also has many formatting options (borders, colour highlighting), to allow you to draw attention to aspects of your figures. One of the more advanced features is conditional formatting, in which Excel automatically assigns distinct colours to cells according to their value (e.g. red for negative, green for positive and appropriate shades in between).

Activity 1 Test your Excel knowledge

Allow approximately 15 minutes to complete this activity

Choose an answer to each of the questions.You can check your answer to each question as you go.

Which of the following functions counts all cells?

□ a) SUMIF

□ b) COUNTIF

□ c) AVERAGE

□ d) COUNT

What is the result of the following formula?

4 + 3*10

To enter the formula, click on an empty cell and type '=4+3*10'.

□ a) 24

□ b) 34

□ c) 30

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 in the keyboard to edit the formula or text?

- □ a) F9
- □ b) F2
- □ c) Ctrl + F

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 insert a new worksheet in an Excel file?

- □ a) Ctrl + Z
- □ b) Ctrl + V
- □ c) Alt + Enter
- □ d) Shift + F11

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 dataset in a table and graphical form.

2 Univariate data visualisation

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

- tabular form
- graphical form.

While presenting and summarising 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 summarise the data and get an idea of its features. 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 you can use either a contingency table or scatter diagram to summarise the data and get an idea of its structure. You will learn about bivariate data visualisation later in the course.

The next section will briefly explain frequency tables.

2.1 Frequency tables

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

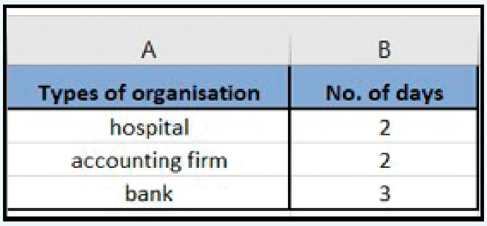
The frequency of a value is the number of times that value appears in a data set. A frequency distribution table displays the pattern of frequencies of a variable in a tabular form. It gives the information of how many times each value of a variable occurs in a data set. A frequency distribution table is an effective way to summarise and organise the collected raw data so that all its features are summarised in a table form. The first step that a researcher or analyst must do with collected raw data is to organise and present the data in such a way that makes it 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. You 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 a clear idea of how many days Anna works in each different organisation.

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
- grouped frequency distribution
- relative frequency distribution
- cumulative frequency distribution.

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. 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. production and sales). However, if you have raw data that is categorised, it is defined as **grouped data**. For example, if this manager knows that 50 employees work in production and 50 employees work in sales, 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 in Table 5, the lower-class limits are 7, 9, 11, 13, 15, and the upper limits are 8, 10, 12, 14, 16. The terms class and class interval are often used interchangeably, although the class interval is a symbol for the class.

Class boundaries: a class boundary is the number that is used to separate the two different classes. It is the midpoint between the upper limit of a class and the lower limit of the next class. Each class has both an upper and a lower limit boundary. The lower boundary of a class is calculated by subtracting half of the value of the interval from the lower-class limit, while the upper boundary of a class is calculated by adding half of the value of the interval to the upper-class limit.

Referring to the example from JC Electrics, its class boundaries are given in Table 5.

Table 5 Class intervals and boundaries for JC Electrics

Class intervals	Class boundaries
7–8	6.5–8.5
9–10	8.5–10.5
11–12	10.5–12.5

13–14	12.5–14.5
15–16	14.5–16.5

Referring to Table 5, you can say that the lower limit of the first-class interval is 6.5, as all values between 6.5 and 7.5 are recorded as 7. Meanwhile, the upper-class limit of 8 is 8.5, as all values between 7.5 and 8.5 are recorded as 8. The real class limit of a class is called a class boundary. A class boundary is obtained by adding two successive class limits and dividing the sum by 2. The value so obtained is taken as the upper-class boundary for the previous class, and lower-class boundary for the next class.

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.

The size or the width of a class interval: the size, or width, of a class interval is the difference between the lower- and upper-class boundaries and is also referred to as the class width, class size, or class length. If all class intervals of a frequency distribution have equal widths, this common width is denoted by c.

Range: this is the difference between the maximum value and the minimum value of the data set. For example, in the JC Electrics data set the maximum number of Electric Motors sold has a value of 25, while the minimum is 14. Hence, to calculate the range, you must calculate 25–14=9.

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; they are usually used to calculate the accurate frequency of individual data values.

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 your 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 4 people having 1 plant. Then you do the same for the rest of the values, so: 3 people have 2 plants, 6 people have 3 plants, 2 people have 4 plants, and 3 people have 5 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

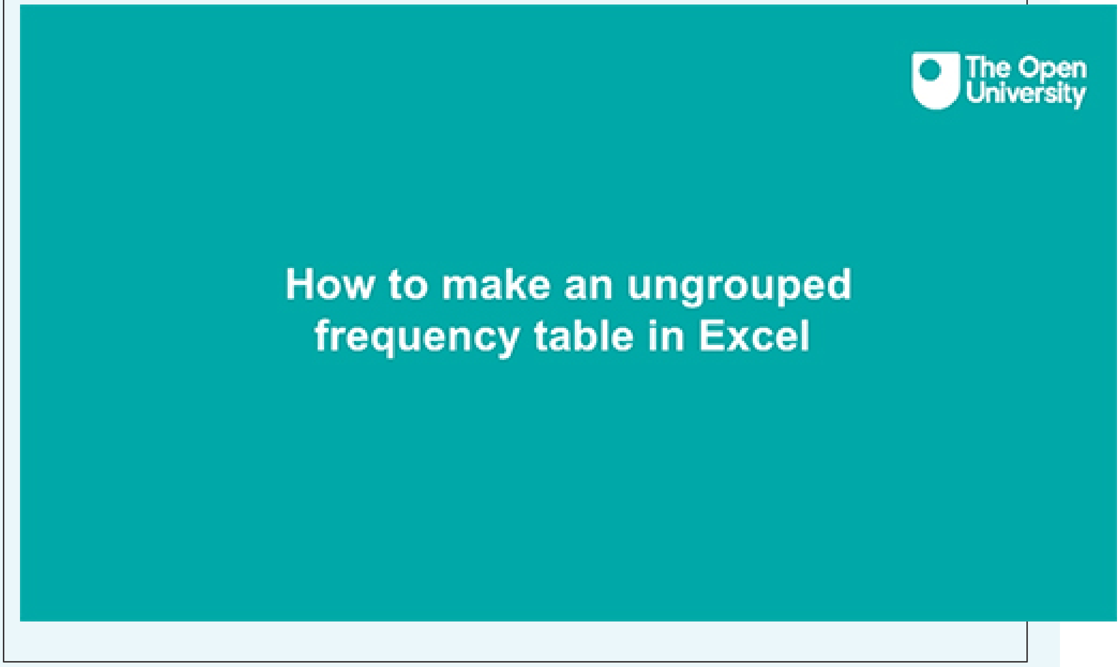
Allow approximately 30 minutes to complete this activity

In this activity, you will learn how to make an ungrouped frequency table in Excel. Once you have produced the ungrouped frequency table in Excel, you may need to compare it with the final output by clicking ‘Reveal discussion’. This will help you to see whether you have produced the accurate ungrouped frequency table or not.

Watch Video 1, which gives on how to create a frequency table, or follow the instructions below.

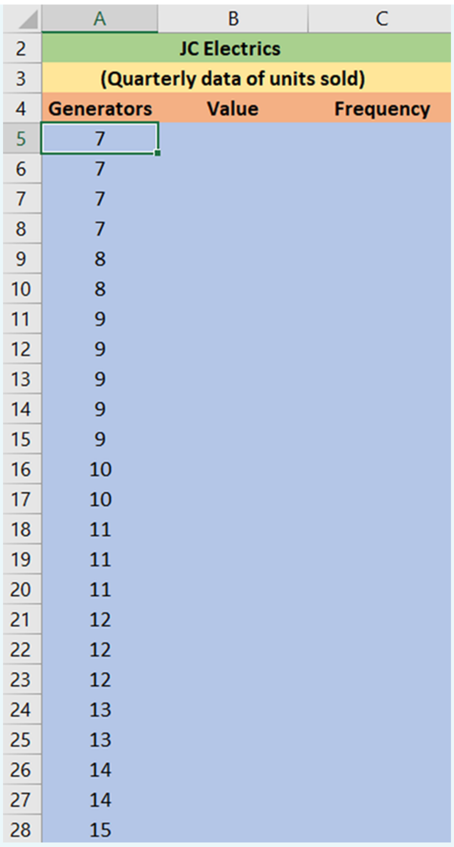
Video content is not available in this format.

Video 1



How to make an ungrouped frequency table in Excel

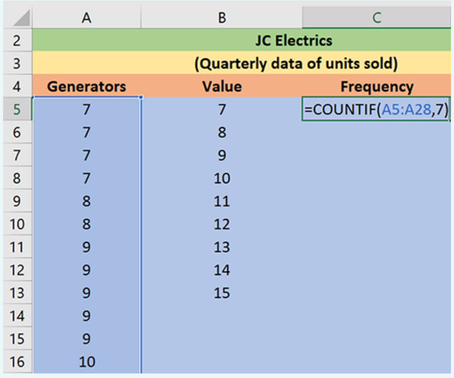
- Open the Excel file JC Electrics. This file contains the quarterly data of number of generators sold. Make sure that the data is arranged in columns.
- Copy the data containing the number of generators sold to **Column A** of a new worksheet.
- Label **Column B** ‘Value’ and label **Column C** ‘Frequency’.
- 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 $15 - 7 = 8$.
- Arrange the values in Column A in ascending 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**. Figure 8 shows how your information should be displayed.



	A	B	C
2	JC Electrics		
3	(Quarterly data of units sold)		
4	Generators	Value	Frequency
5	7		
6	7		
7	7		
8	7		
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		

Figure 8 Arranging the data from ascending to descending order

- To count the number of quarters in which 7 units were sold, you need to calculate the frequency in Column C. Type **=COUNTIF (Range, value)**. For example, **=COUNTIF (A5:A28,7)**

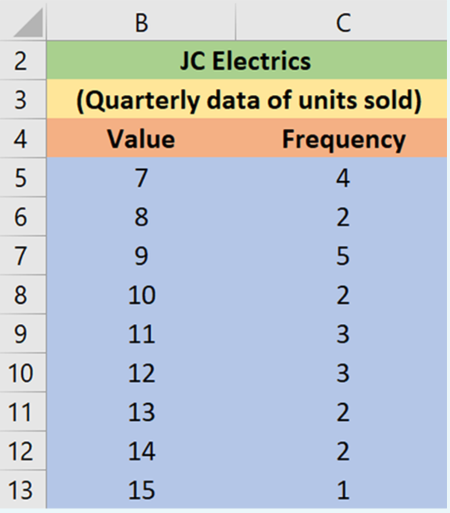


	A	B	C
2	JC Electrics		
3	(Quarterly data of units sold)		
4	Generators	Value	Frequency
5	7	7	=COUNTIF(A5:A28,7)
6	7	8	
7	7	9	
8	7	10	
9	8	11	
10	8	12	
11	9	13	
12	9	14	
13	9	15	
14	9		
15	9		
16	10		

Figure 9 Calculating the frequency of ungroup data

- You should now save your file as you will return to this ungrouped frequency table in a later activity.

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

Figure 10 above shows the completed frequency table. The same data is shown in Table 6 below.

Table 6 Ungrouped frequency table

Value	Frequency
7	4
8	2
9	5
10	2
11	3
12	3
13	2
14	2
15	1

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 amounts 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 much of the original detail of the data.

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

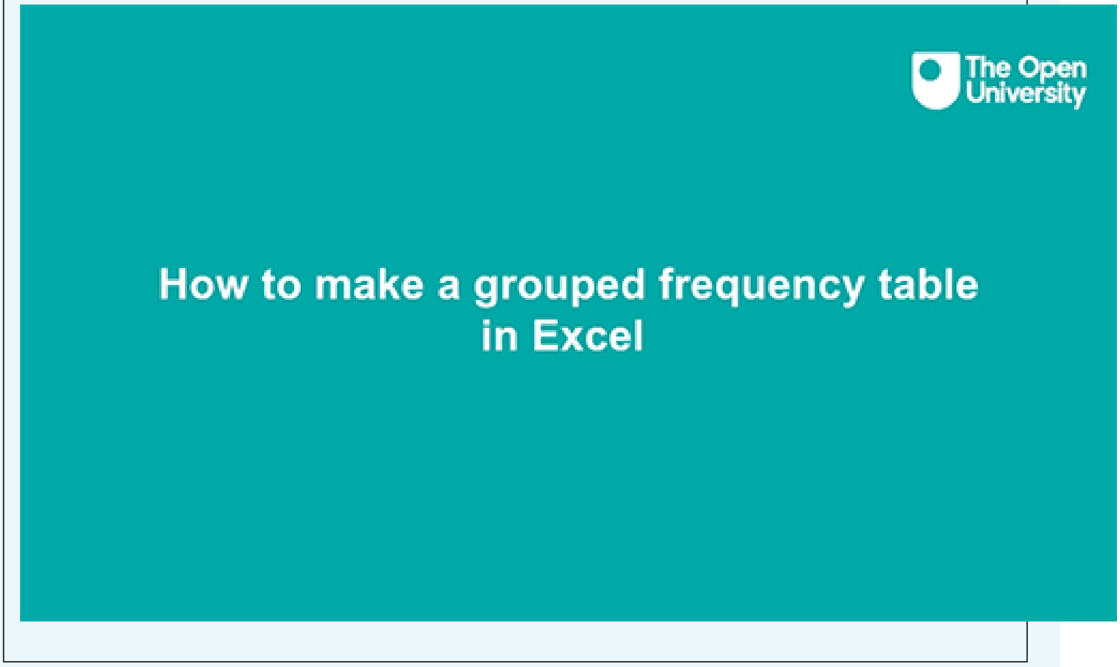
Activity 3 How to make a grouped frequency distribution table in Excel

Allow approximately 30 minutes to complete this activity

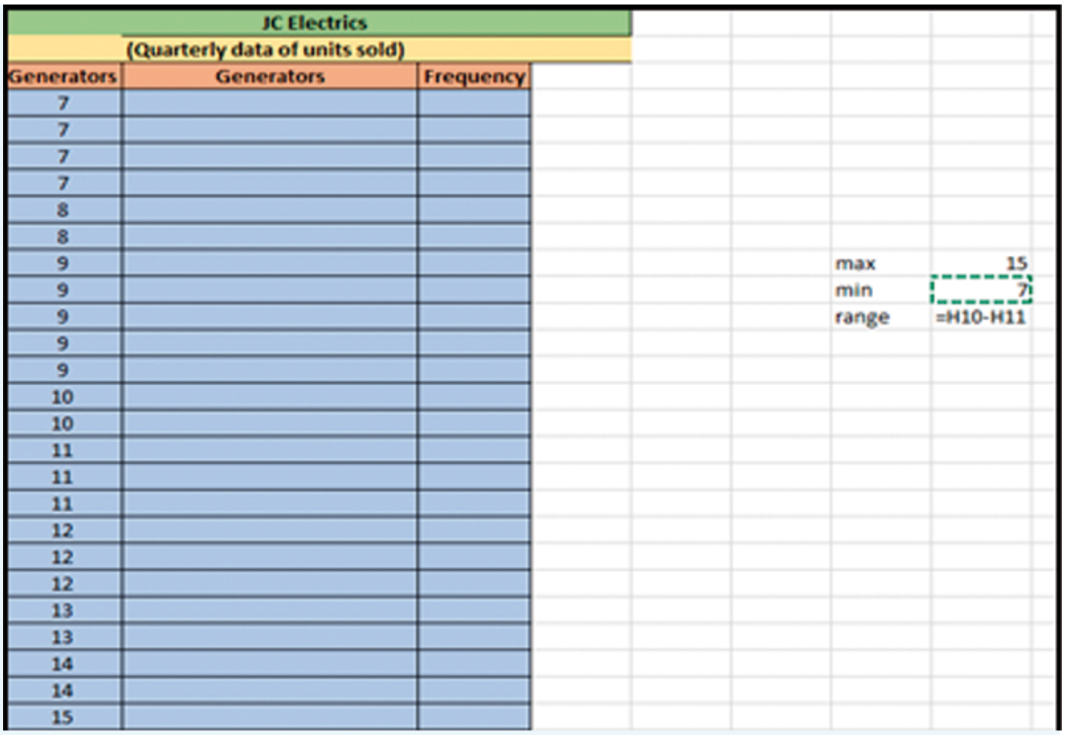
In this activity, you need to produce a grouped frequency table in Excel either by watching the screencast in Video 2 or by following the instructions given below. Once you have produced the grouped frequency table in Excel, you can check your answer by clicking 'Reveal discussion'.

Video content is not available in this format.

Video 2



- Open the Excel file JC Electrics. This file contains quarterly data of the number of generators sold. Make sure that the data is arranged in columns.
- Find the range which is the difference between the maximum and minimum value in the data set. You can do this either by entering the formula **=MAX (A2:A25)-MIN (A2:A25)**, or by simply using the results you have calculated in Column H as, **=H10-H11** (see Figure 11).



JC Electrics

(Quarterly data of units sold)

Generators	Generators	Frequency
7		
7		
7		
7		
8		
8		
9		
9		
9		
9		
9		
10		
10		
11		
11		
11		
12		
12		
12		
13		
13		
14		
14		
15		

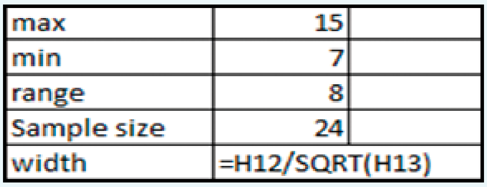
max	15
min	7
range	=H10-H11

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:

$$\text{Width} = \frac{\text{Range}}{\sqrt{\text{Sample size}}}$$

- 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 1.6, so will be taken as **2** (see Figure 12).

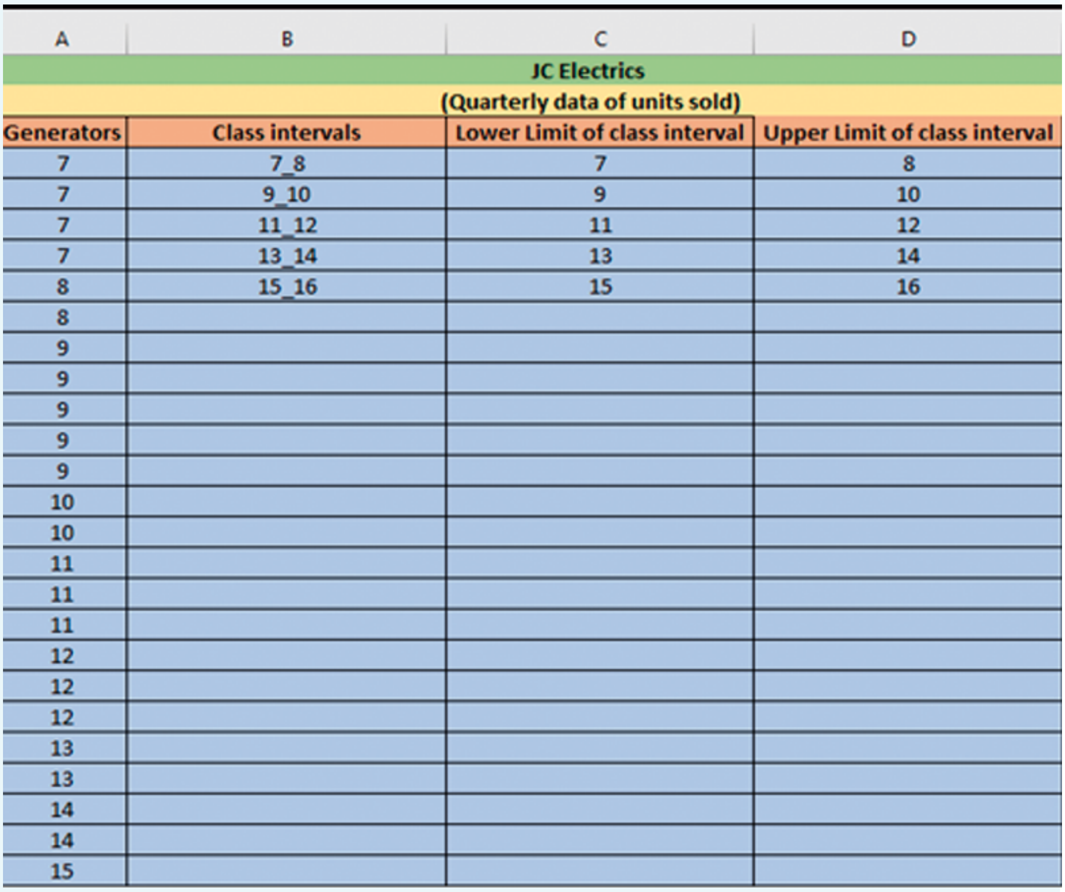


max	15	
min	7	
range	8	
Sample size	24	
width	=H12/SQRT(H13)	

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. Add the class interval width to find the upper limit of the first interval and 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.



JC Electrics

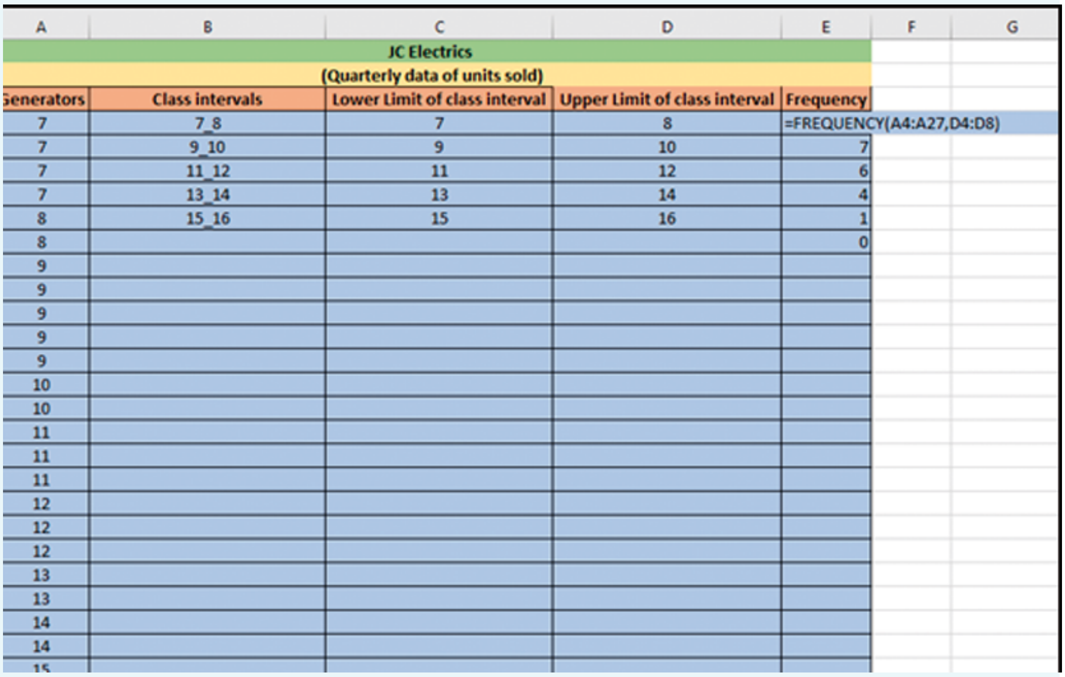
(Quarterly data of units sold)

A	B	C	D
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 Maing the class intervals

- The next step is to calculate the frequency. Select the range **E2:E6** and enter FREQUENCY function as shown in the Figure 14 in the discussion.
- Press **CTRL + SHIFT + ENTER** to submit the FREQUENCY formula above as an array formula. If it is entered correctly, you would see a formula wrapped in curly braces {}.
- You should now save your file as you will return to this grouped frequency table in a later activity.

Discussion



A	B	C	D	E	F	G
JC Electrics						
(Quarterly data of units sold)						
Generators	Class intervals	Lower Limit of class interval	Upper Limit of class interval	Frequency		
7	7_8	7	8	=FREQUENCY(A4:A27,D4:D8)		
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				0		
9						
9						
9						
9						
9						
10						
10						
11						
11						
11						
12						
12						
12						
13						
13						
14						
14						
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 that 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 section, 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 you the proportion of each value or class interval of a variable. 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

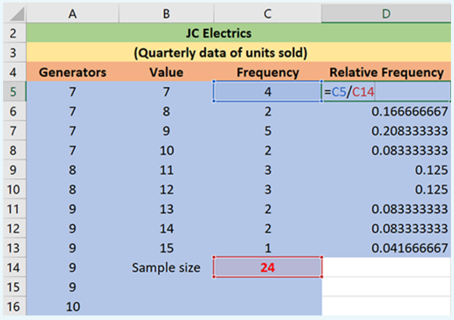
Allow approximately 30 minutes to complete this activity

In this activity, you will build a relative frequency table using the ungrouped frequency distribution table from Activity 2. 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 ‘Value’ and Column C is labelled ‘Frequency’. Add a fourth column to the table for the relative frequencies.

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.

Discussion



	A	B	C	D
2	JC Electrics			
3	(Quarterly data of units sold)			
4	Generators	Value	Frequency	Relative Frequency
5	7	7	4	=C5/C14
6	7	8	2	0.166666667
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			

Figure 15 Calculating the relative frequency of ungrouped data

Figure 15 represents the quarterly data of number of units of generators sold. Column A (‘Generators’) contains values between 7 to 15. Column B is labelled ‘Value’. Column C is labelled ‘Frequency’.

To calculate the relative frequency in Column D, you need to divide each frequency by sample size. The sample size of 24 can be found by summing the ‘Frequency’ column.

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. It is the sum of the frequencies less than or equal to each value or class interval of a variable. 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

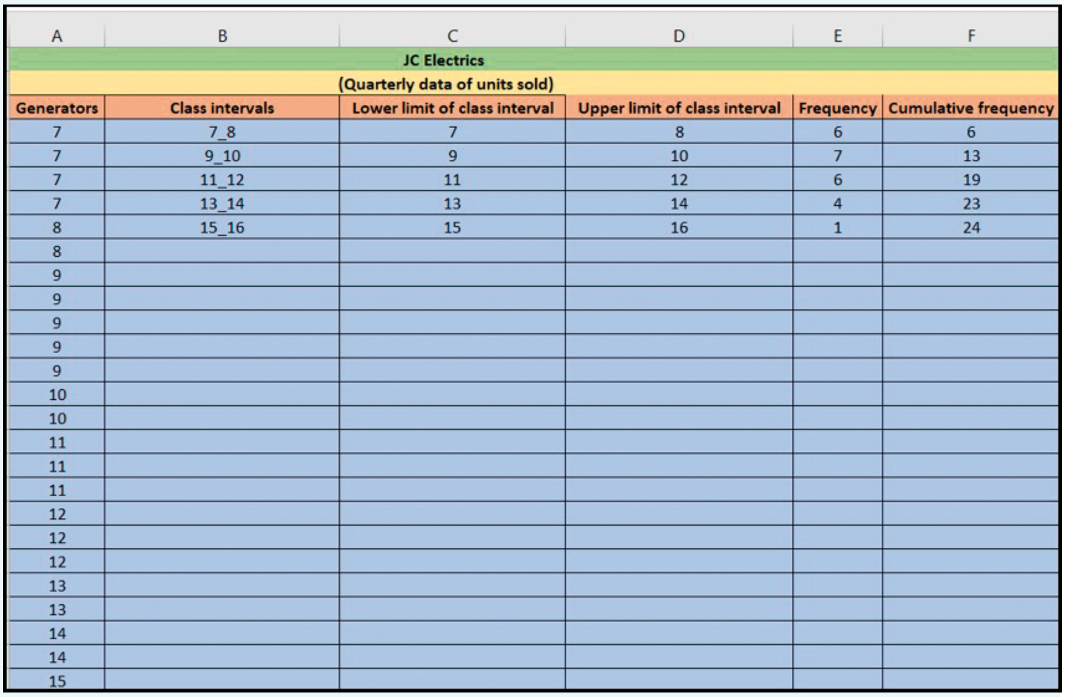
Allow approximately 30 minutes to complete this activity

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.

Borrow the grouped frequency distribution table from Activity 3. This table consists of five 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.

Add another column, Column F, to the table for 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.

Discussion



A	B	C	D	E	F
JC Electrics					
(Quarterly data of units sold)					
Generators	Class intervals	Lower limit of class interval	Upper limit of class interval	Frequency	Cumulative frequency
7	7_8	7	8	6	6
7	9_10	9	10	7	13
7	11_12	11	12	6	19
7	13_14	13	14	4	23
8	15_16	15	16	1	24
8					
9					
9					
9					
9					
9					
10					
10					
11					
11					
11					
12					
12					
12					
13					
13					
14					
14					
15					

Figure 16 Calculating the cumulative frequency

Figure 16 presents the data of the number of generators sold. Column A (‘Generators’) shows the values between 7 to 15 in ascending order. Column B shows the class intervals. Column C shows the values of the lower limit of each class intervals. Column D shows the values of upper limit of each class intervals. Column E shows the frequency.

To calculate the cumulative frequency in Column F, add each frequency to the frequencies in the previous rows. If you do it correctly, the value in the last row will be equal to the sample size.

In the next section, you will learn how to visualise these tables by drawing histograms in Excel.

2.3 Histograms: a graphical visualisation of frequency tables

A histogram is a popular visualisation tool to summarise the distribution of continuous data. In a histogram, the variable is divided into intervals called ‘bins’. You then count the number of observations in each bin and plot the resulting table in a bar chart. The horizontal *x*-axis displays the ‘bins’ and the 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 bar chart below. Note that the values on the x-axis show the upper limit of the interval. In a proper histogram, there are no spaces or gaps between the bars.

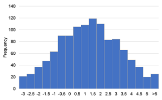


Figure 17 A typical histogram in Excel

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

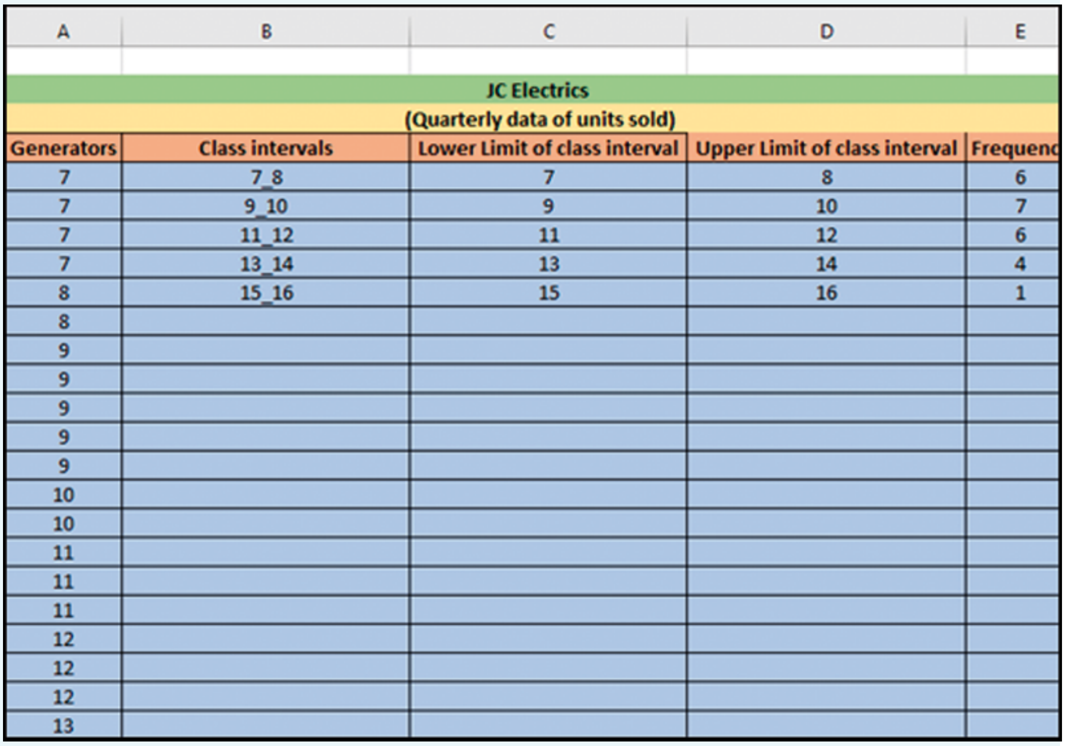
Activity 6 Using Excel to draw a histogram

Allow approximately 35 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, which contains the quarterly data of number of generators sold. The third column C contains information about the number of generators sold in each quarter of the year.
- Find the minimum and maximum value in the data set. You can obtain them through the min (range) and max (range) functions in Excel. Type **=MAX(A5:A28)** into cell L10 and **=MIN(A5:A28)** into cell L11. This will give the minimum and maximum values of the data set, which are 7 and 15.
- Next, you need to specify a range of intervals (often called ‘bins’) for which to count the number of observations that fall into each bin. The maximum value is 15 and the minimum value is 7, so you can make the class intervals 7–8, 9–10, 11–12, 13–14, 14–15, 15–16 etc. This means that the first class has the lower

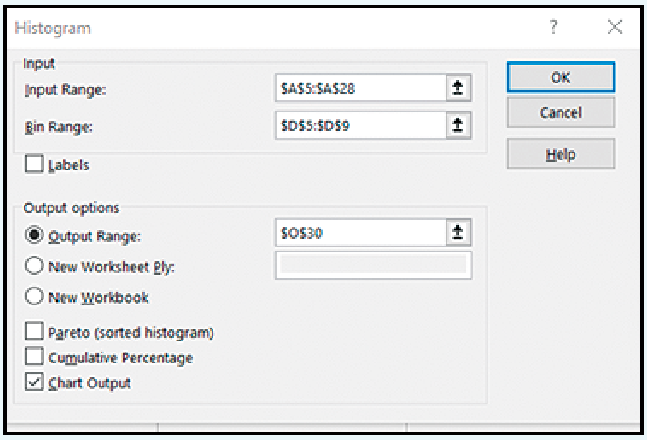
value 7 and the maximum value 8 and so on. See Columns C and D in the worksheet in Figure 18.



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

- There are many ways to calculate the width of the bin in Excel. One of the easiest ways to calculate it is as the width of the bin or class intervals (sample size / range), which is 3 (i.e. 24/8=3). In this example, the bin width is 2.
- 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
- Tick the box ‘Chart Output’ and specify the output location as H5, as shown in Figure 19 below.



Histogram

Input

Input Range: $A$5:$A$28

Bin Range: $D$5:$D$9

□ Labels

Output options

● Output Range: $O$30

○ New Worksheet Ply:

○ New Workbook

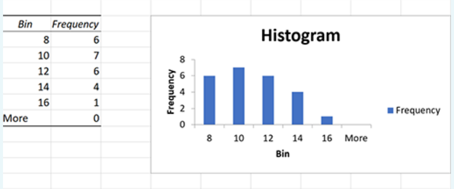
□ Pareto (sorted histogram)

□ Cumulative Percentage

☑ Chart Output

Figure 19 Histogram dialog box in Excel

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



Bin	Frequency
8	6
10	7
12	6
14	4
16	1
More	0

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

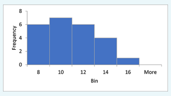


Figure 21 Histogram showing units sold of generators

Working through the steps given above should return the completed histogram shown in Figure 21.

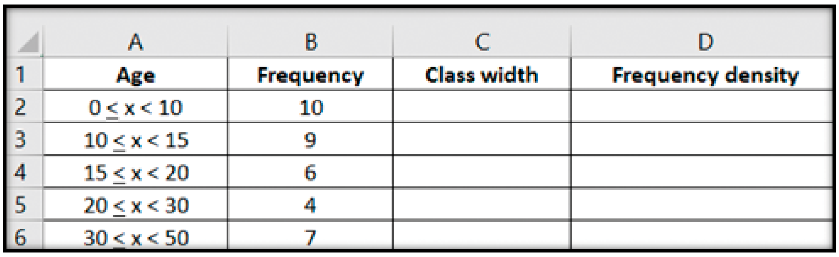
2.4 Frequency density

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 width may not be equal.

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

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

You can calculate frequency density when you have a set of grouped data that consists of unequal widths of class intervals. For example, see the following Excel worksheet in Figure 22, which shows information about the ages of a group of people playing football.

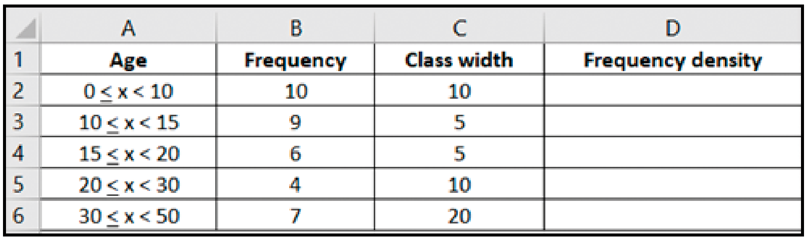


	A	B	C	D
1	Age	Frequency	Class width	Frequency density
2	0 ≤ x < 10	10		
3	10 ≤ x < 15	9		
4	15 ≤ x < 20	6		
5	20 ≤ x < 30	4		
6	30 ≤ x < 50	7		

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

To calculate the frequency densities:

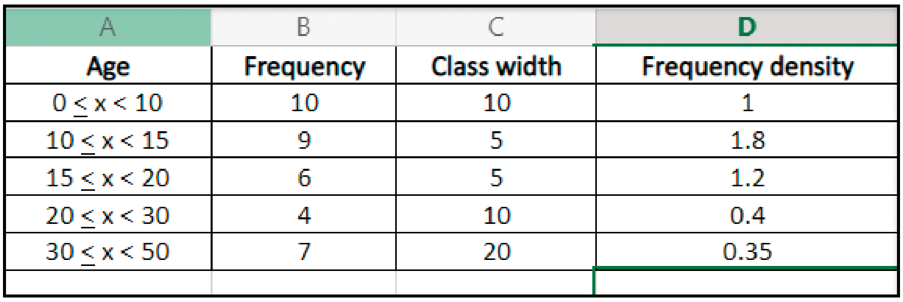
- In Column C, find the class width of the class intervals by finding the difference of upper and lower bounds/limits. (For example, $10 - 0 = 10$, $15 - 10 = 5$, and so on.)



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

Figure 23 Calculating the class width

- Then, in Column D, divide the frequency of each class interval by its width.



A	B	C	D
Age	Frequency	Class width	Frequency density
0 ≤ x < 10	10	10	1
10 ≤ x < 15	9	5	1.8
15 ≤ x < 20	6	5	1.2
20 ≤ x < 30	4	10	0.4
30 ≤ 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.

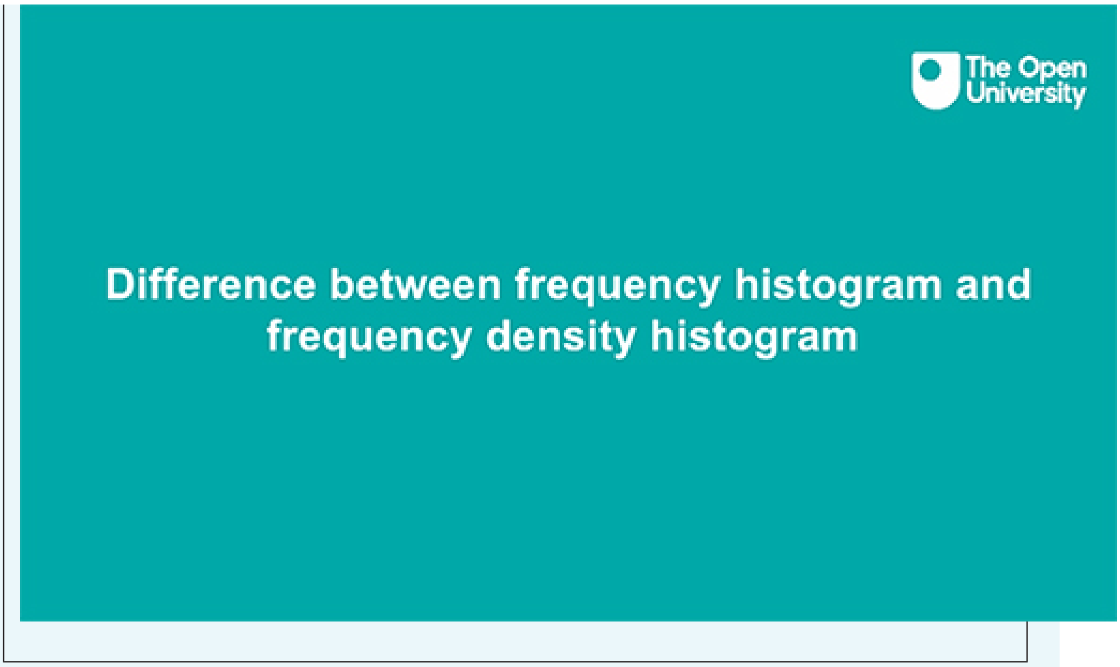
Activity 7 The difference between frequency histograms and frequency density histograms

Allow approximately 15 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 3



Difference between frequency histogram and frequency density histogram

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 well employees perform their work, that is, to measure the efficiency of the employees. In this example, the performance manager may examine two variables: the number of tasks they complete and the quality of the tasks.

Bivariate data is collected to explore the relationship between two variables and then use this relationship to inform future decisions. One of the main aims of the researcher is to find out whether changes in one variable may be caused by changes in another 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 whether 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 15 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 4

Bivariate data

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 you want to summarise the following sample of firms in Table 7 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 (Table 8), which summarises the information of the sample data. It counts the number of firms for each combination of sector and number of employees.

Table 8 Cross table for sample firms

Sector	<50 Employees	50 + Employees	Total employees
Technology	4	0	4
Food	2	4	6
Total	6	4	10

The cross table shows that there are two firms in the food manufacturing sector that have less than 50 employees. However, the cell 50+ shows that there are four firms in the food manufacturing sector that have more than 50 employees. The sum of the total food manufacturing firm is six which is the 60% of the grand total.

Contingency tables vary in size and type because the size of the contingency table depends on the sample size and number of observations.

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. If you would like to experiment with PivotTables, go to the ‘Insert’ ribbon in Excel and select ‘PivotTable’.

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). You can then plot the corresponding point on the diagram.

In the next activity, you will produce a scatter diagram in Excel either by following the video or the instructions provided.

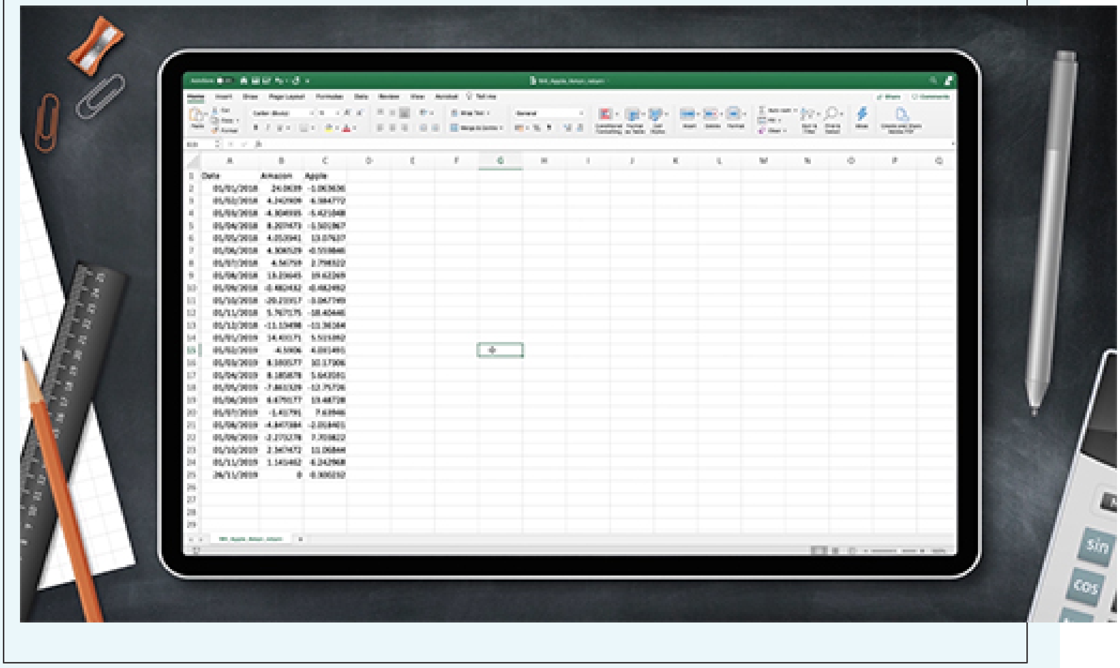
Activity 9 Drawing scatter diagrams

Allow approximately 35 minutes to complete this activity

The screencast in Video 5 gives you instructions on how to draw scatter plots in Excel.

Video content is not available in this format.

Video 5



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

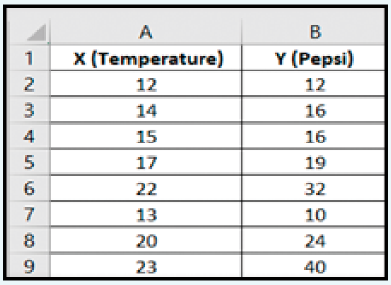
Table 9 Temperature and number of drinks sold

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

Here, temperature is the independent variable and number of drinks is the dependent variable, as the sale of Pepsi is affected by changes in the temperature. Hence, you will plot temperature on the x-axis and drinks 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.

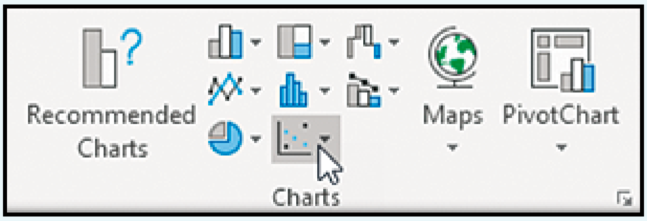
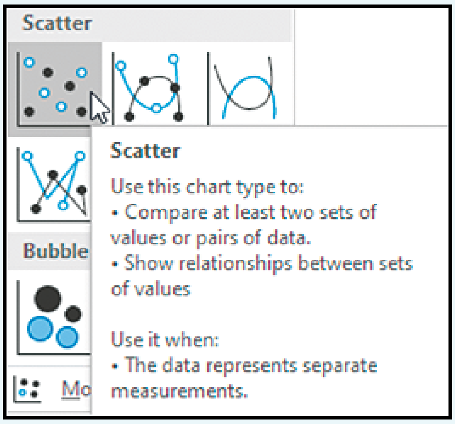


Figure 26 How to select the Scatter symbol

- This will open a drop-down menu showing various types of scatter plots. The standard type is the one with unconnected dots in the top left. Click the **Scatter** symbol to insert this chart.



Scatter

Use this chart type to:
• Compare at least two sets of values or pairs of data.
• Show relationships between sets of values

Use it when:
• The data represents separate measurements.

Figure 27 How to select the Scatter chart type

Discussion

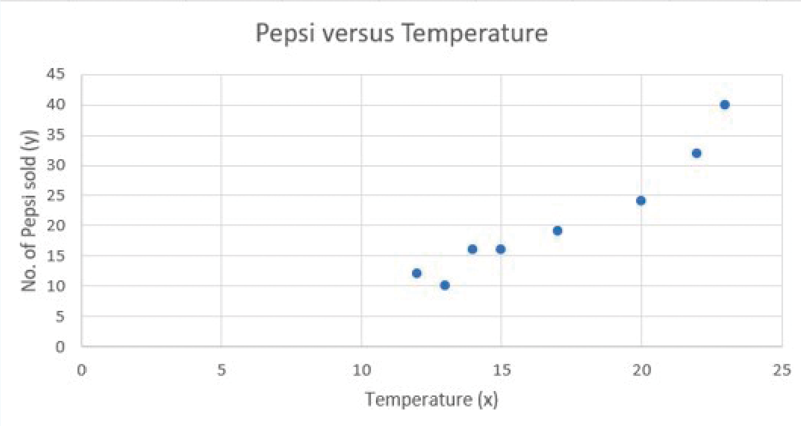


Figure 28 Temperature versus Pepsi sales

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.

Conclusion

In this course, 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: for example, in 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 course are only the first step toward starting the decision-making process using data.

The next step could be to study descriptive statistics, which gets you closer to a comprehensive analysis of the data. You could then become more confident when examining and summarising data and using Excel tools such as measures of location and measures of dispersion to numerically analyse 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 Henry Lahr.

The material acknowledged below is Proprietary and used under licence (not subject to Creative Commons Licence). Grateful acknowledgement is made to the following sources for permission to reproduce material in this free course:

Course image: ismagilov / iStock / Getty Images Plus

Figure 1: jadamprostore / iStock / Getty Images Plus

Figures 2–28: © Excel

Every effort has been made to contact copyright owners. If any have been inadvertently overlooked, the publishers will be pleased to make the necessary arrangements at the first opportunity.