Introduction to Data Science

Lab 1 – Exploring Data in Microsoft Excel Online

Overview

Rosie Reeves is an entrepreneurial middle-school student who sells homemade lemonade from a stand at the park near her house. To promote her lemonade-stand, she distributes leaflets in the park. Rosie records details of her sales and flyer (leaflet) distribution, along with weather measurements including the temperature and rainfall each day.

In this lab, you will explore and visualize the data Rosie recorded.

What You'll Need

To complete the labs, you will need the following:

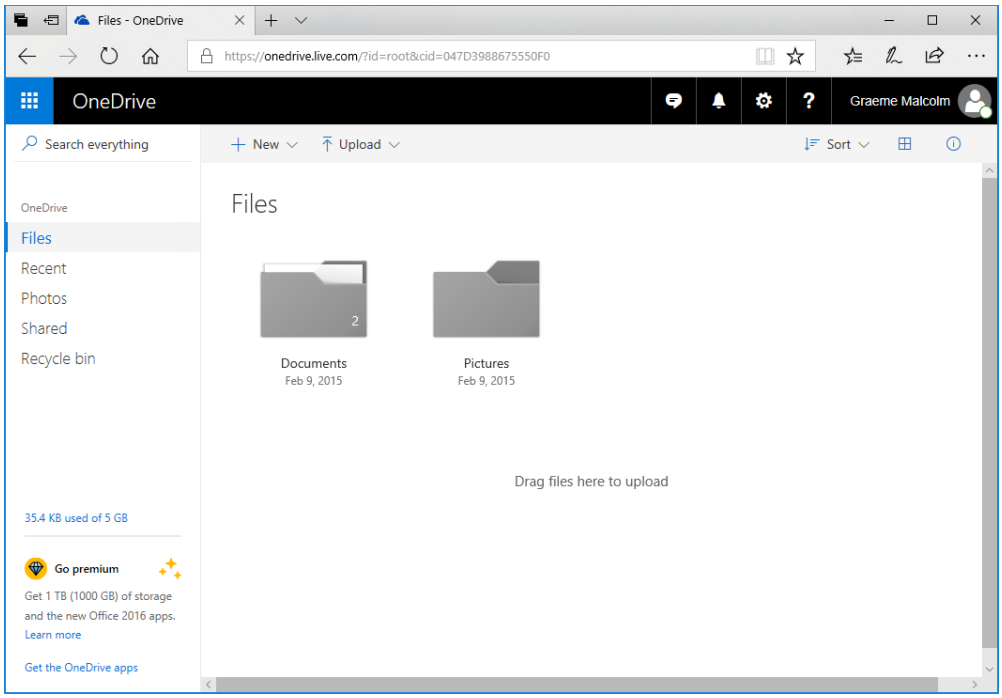
- A Windows, Linux, or Mac OS X computer with a web browser.
- A Microsoft account (for example a *hotmail.com*, *live.com*. or *outlook.com* account). If you do not already have a Microsoft account, sign up for one at https://signup.live.com.
- The lab files for this course. Download these from https://aka.ms/edx-dat101x-labfiles, and extract them to a folder on your computer.

Exercise 1: Viewing a Table of Data in Excel

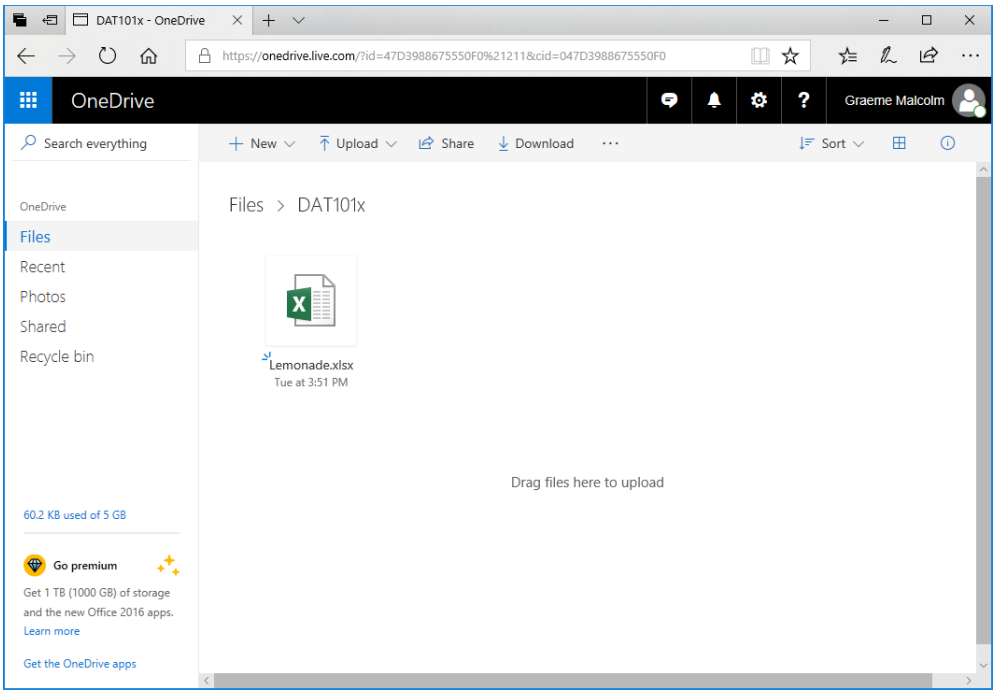
In this exercise, you will upload the Excel workbook containing Rosie's data to the OneDrive cloud storage account associated with your Microsoft account, and then explore the data in Microsoft Excel Online.

Upload the Workbook to OneDrive

1. In your web browser, navigate to https://onedrive.live.com, and sign in using your Microsoft account credentials. You should see the files and folders in your OneDrive, like this:

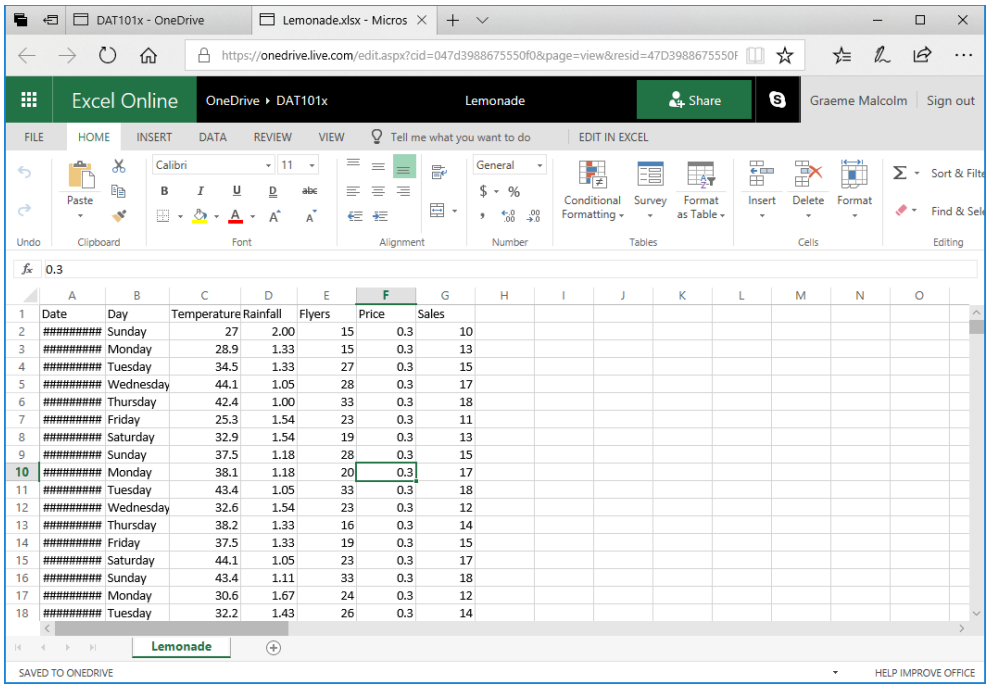


2. On the + **New** menu, click **Folder** to create a new folder. You can name this anything you like, for example **DAT101x**.
3. Click your new folder to open it, and then drag the **Lemonade.xlsx** Excel workbook file from the folder where you extracted the lab files for this course into the browser window to upload it to your folder. It should appear in your folder like this:

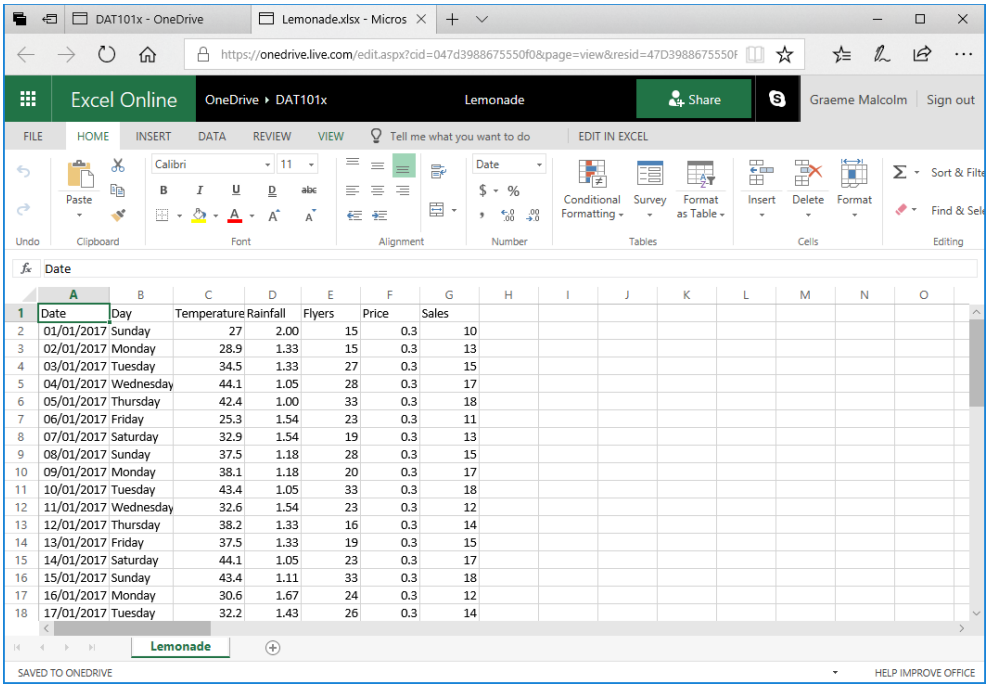


Open the Workbook in Excel Online

1. Click the **Lemonade.xlsx** file in your OneDrive folder to open it in Excel Online. When opened, it should look like this:

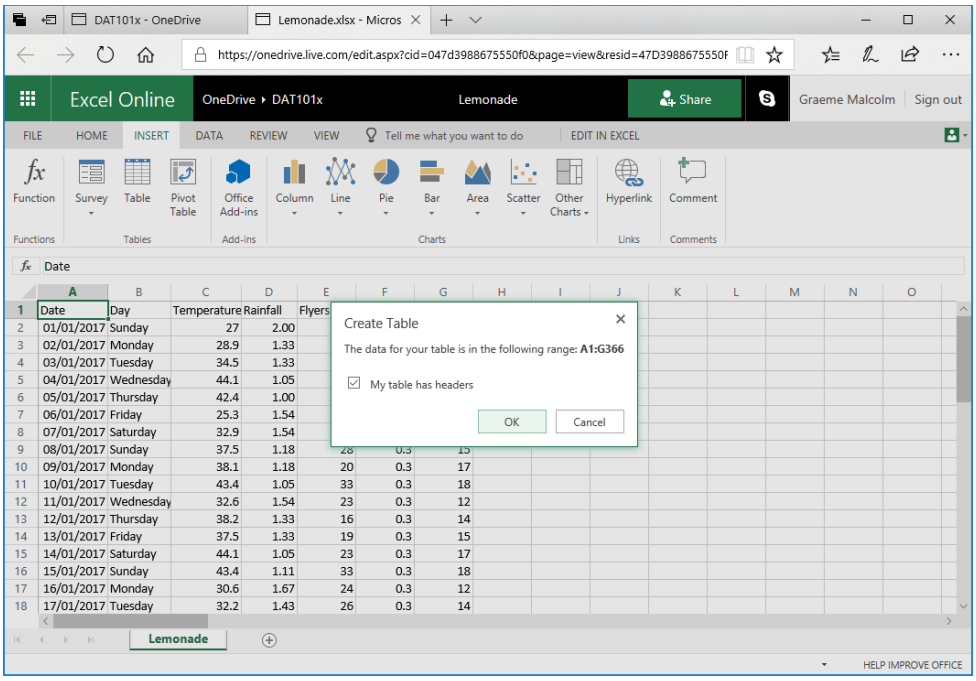


2. The dates column **A** may be too wide to be displayed, so the cells may contain **#######** as shown above. To see the dates, double-click the line between the **A** and **B** column headers. The dates will then be shown in the format for the locale associated with your Microsoft account. For example, in the following image, the dates are shown in UK format (*dd*/*MM*/*yyyy*):

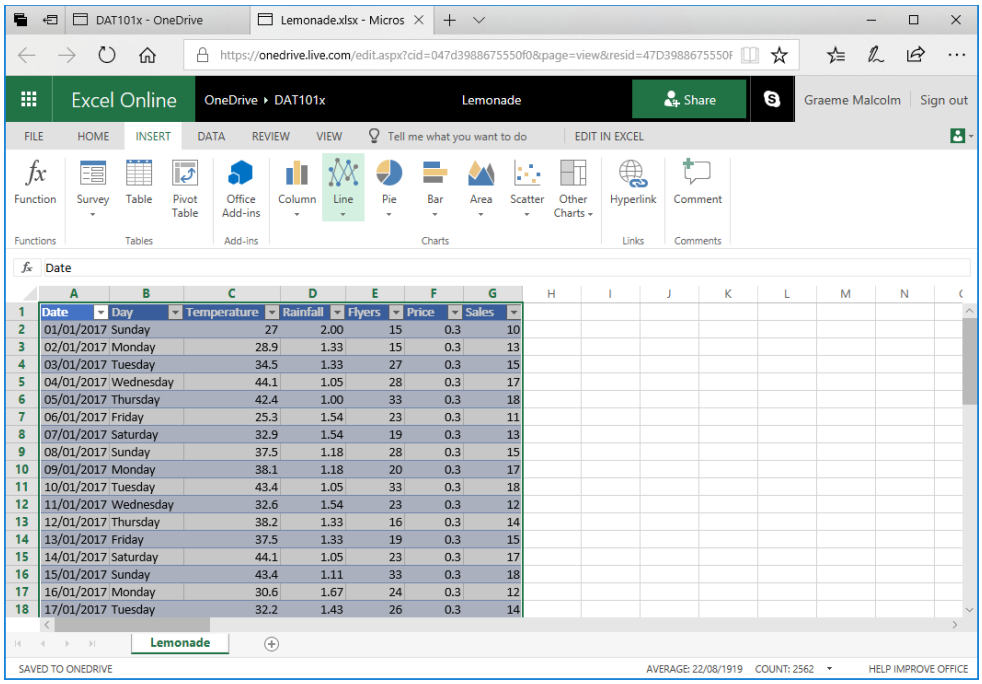


Filter and Sort the Data

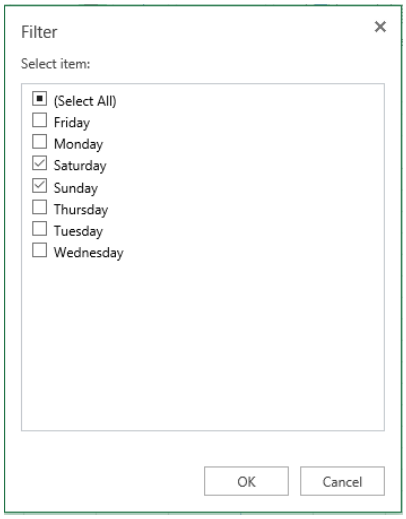
1. Select cell **A1**, and then on the **Insert** tab of the ribbon above the worksheet, click **Table**. Verify that Excel has automatically detected the data in the range **A1:G366**, and that the **My table has headers** checkbox is selected, and then click **OK**; as shown here:



Excel automatically formats the data as a table and adds drop-down buttons to the header row as shown here:

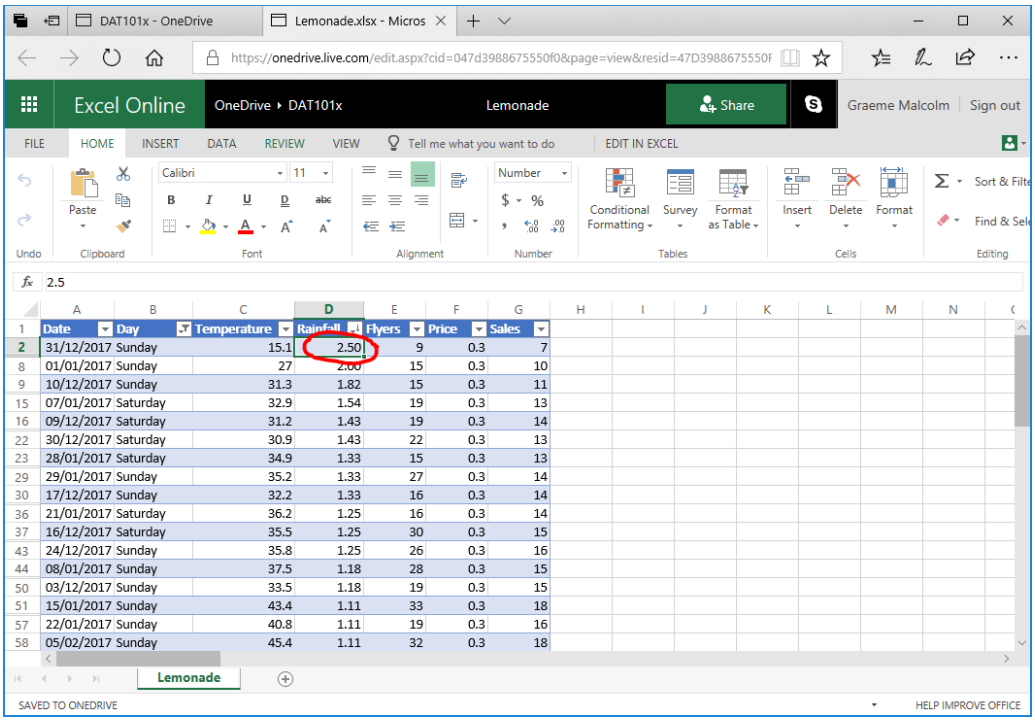


2. Click any cell to deselect the table, and then click the drop-down button for the **Day** column, and click **Filter…**
3. In the **Filter** dialog box, clear the **(Select All)** checkbox, and then select only the **Saturday** and **Sunday** checkboxes as shown here before clicking **OK**:



The table of data is filtered to show only the records for weekend days (Saturday and Sunday).

4. Click the drop-down arrow for the **Rainfall** column and click **Sort Descending**. The table of data is sorted in descending order of rainfall, so the first row contains the data for the weekend day with the most rain. This was a Sunday on which there was 2.50 cm of rain as shown here:



5. Click the drop-down arrow for the **Day** column again and then click **Clear Filter from 'Day'**. The table now shows all the data.
6. Click the drop-down arrow for **Date** and click **Sort Ascending** to re-order the data into chronological order.

Challenge: Find the Weekday with the Lowest Temperature

1. Using the filter and sort capabilities in Excel Online, filter the data so that only weekdays (Monday to Friday) are shown, and sort the data so that the first row contains data for the weekday with the lowest temperature.

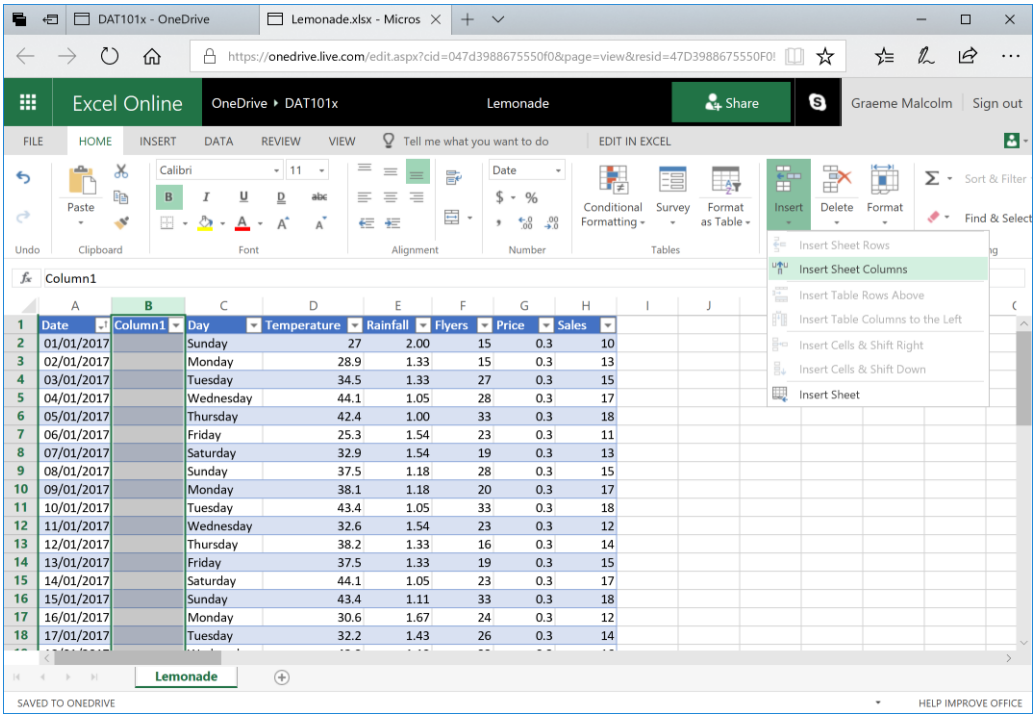
2. Make a note of the day and the temperature, and then clear the filter and re-sort the data back into chronological order.

Exercise 2: Using Formulae to Explore Data in Excel

In this exercise, you will use formulae to create derived columns that extend the data recorded by Rosie.

Add Derived Columns

1. Click the **B** column header to select the entire **B** column. Then on the **Home** tab of the ribbon, in the **Insert** drop-down menu, click **Insert Sheet Columns** This inserts a new **Column1** column between the **Date** and **Day** columns as shown here:

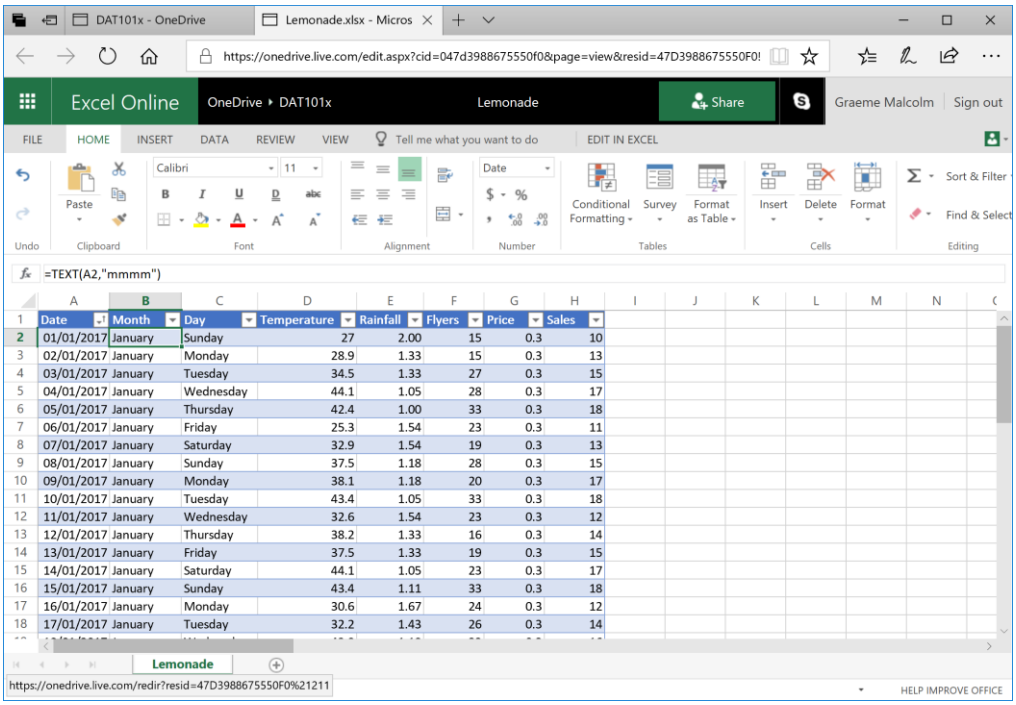


	A	B	C	D	E	F	G	H
1	Date	Column1	Day	Temperature	Rainfall	Flyers	Price	Sales
2	01/01/2017		Sunday	27	2.00	15	0.3	10
3	02/01/2017		Monday	28.9	1.33	15	0.3	13
4	03/01/2017		Tuesday	34.5	1.33	27	0.3	15
5	04/01/2017		Wednesday	44.1	1.05	28	0.3	17
6	05/01/2017		Thursday	42.4	1.00	33	0.3	18
7	06/01/2017		Friday	25.3	1.54	23	0.3	11
8	07/01/2017		Saturday	32.9	1.54	19	0.3	13
9	08/01/2017		Sunday	37.5	1.18	28	0.3	15
10	09/01/2017		Monday	38.1	1.18	20	0.3	17
11	10/01/2017		Tuesday	43.4	1.05	33	0.3	18
12	11/01/2017		Wednesday	32.6	1.54	23	0.3	12
13	12/01/2017		Thursday	38.2	1.33	16	0.3	14
14	13/01/2017		Friday	37.5	1.33	19	0.3	15
15	14/01/2017		Saturday	44.1	1.05	23	0.3	17
16	15/01/2017		Sunday	43.4	1.11	33	0.3	18
17	16/01/2017		Monday	30.6	1.67	24	0.3	12
18	17/01/2017		Tuesday	32.2	1.43	26	0.3	14

2. In cell **B1**, rename **Column1** to **Month**. Then in cell **B2**, enter the following formula:

```
=TEXT(A2, "mmmm")
```

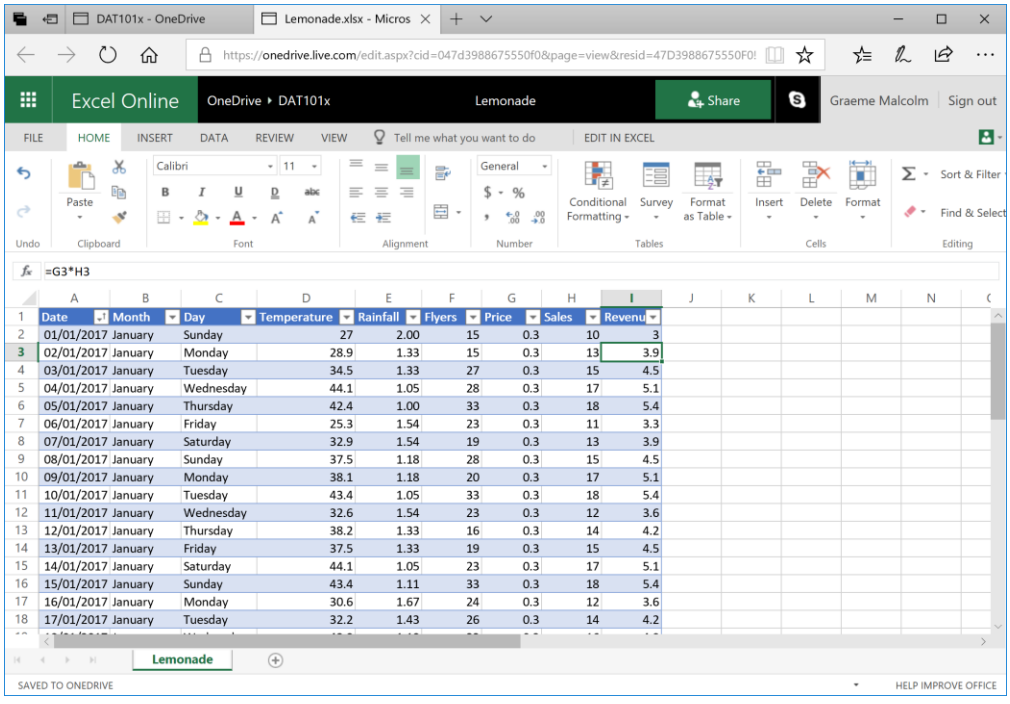
After you enter the formula, it should be copied automatically to all the other **Month** cells in the table, and the name of the month for each record should be displayed as shown here:



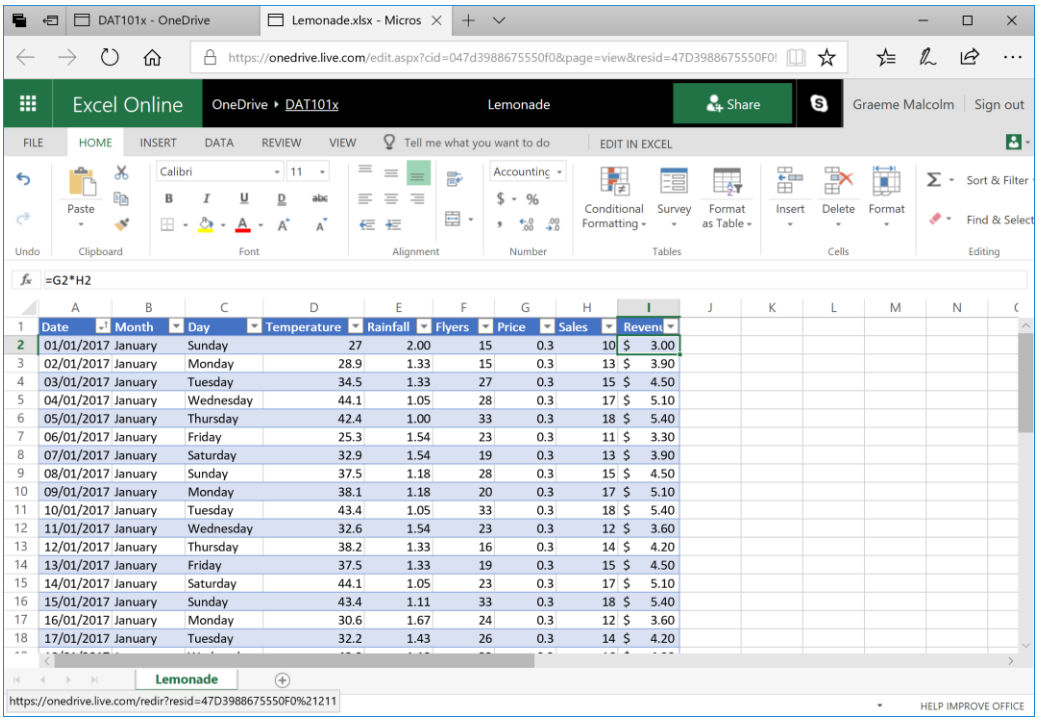
3. In cell **I1**, enter the text **Revenue** to add a new **Revenue** column to the table. Then in cell **I2**, enter the following formula:

```
= G2*H2
```

The formula is again automatically copied to the remaining rows in the table, and the revenue (calculated as **Price** multiplied by **Sales**) is displayed as shown here:



4. Click the **I** column header to select the entire column, and then on the **Home** tab of the ribbon, in the **Number** section, in the *Accounting Number Format* (**$**) drop-down list, select **$ English (United States)**. This formats the revenue data as US dollars:

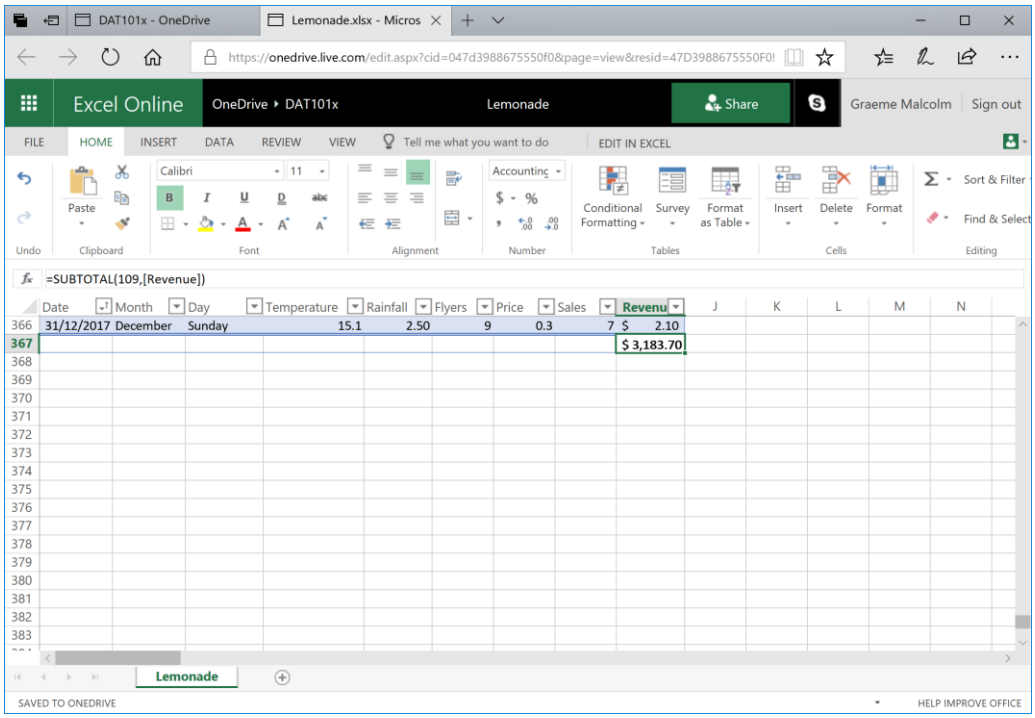


	A	B	C	D	E	F	G	H	I
1	Date	Month	Day	Temperature	Rainfall	Flyers	Price	Sales	Revenu
2	01/01/2017	January	Sunday	27	2.00	15	0.3	10	$ 3.00
3	02/01/2017	January	Monday	28.9	1.33	15	0.3	13	$ 3.90
4	03/01/2017	January	Tuesday	34.5	1.33	27	0.3	15	$ 4.50
5	04/01/2017	January	Wednesday	44.1	1.05	28	0.3	17	$ 5.10
6	05/01/2017	January	Thursday	42.4	1.00	33	0.3	18	$ 5.40
7	06/01/2017	January	Friday	25.3	1.54	23	0.3	11	$ 3.30
8	07/01/2017	January	Saturday	32.9	1.54	19	0.3	13	$ 3.90
9	08/01/2017	January	Sunday	37.5	1.18	28	0.3	15	$ 4.50
10	09/01/2017	January	Monday	38.1	1.18	20	0.3	17	$ 5.10
11	10/01/2017	January	Tuesday	43.4	1.05	33	0.3	18	$ 5.40
12	11/01/2017	January	Wednesday	32.6	1.54	23	0.3	12	$ 3.60
13	12/01/2017	January	Thursday	38.2	1.33	16	0.3	14	$ 4.20
14	13/01/2017	January	Friday	37.5	1.33	19	0.3	15	$ 4.50
15	14/01/2017	January	Saturday	44.1	1.05	23	0.3	17	$ 5.10
16	15/01/2017	January	Sunday	43.4	1.11	33	0.3	18	$ 5.40
17	16/01/2017	January	Monday	30.6	1.67	24	0.3	12	$ 3.60
18	17/01/2017	January	Tuesday	32.2	1.43	26	0.3	14	$ 4.20

5. Scroll down to the bottom of the table of data, select cell **I367** (under the **Revenue** column). Then on the **Home** tab of the ribbon, in the **Editing** section, in the *AutoSum* (**Σ**) drop-down menu, click **Σ Sum**. This enters the following formula:

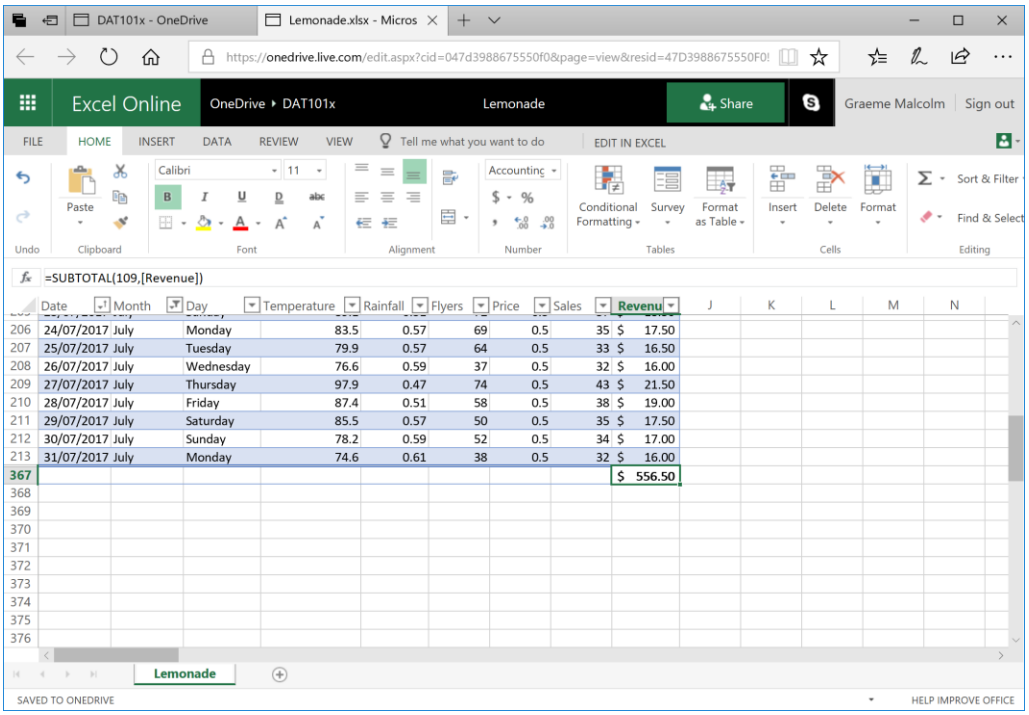
```
=SUBTOTAL(109,[Revenue])
```

This formula references **Revenue** as a named column in the table and calculates the total of the values in that column. You could achieve the same result by entering `=SUM(I2:I366)` but by using the AutoSum function, the value is included in the definition of the table:



	Date	Month	Day	Temperature	Rainfall	Flyers	Price	Sales	Revenu
366	31/12/2017	December	Sunday	15.1	2.50	9	0.3	7	$ 2.10
367									$ 3,183.70

6. Filter the **Month** column to show only the records for July, and then look at the subtotal at the bottom of the **Revenue** column (you may need to scroll to find it). It now shows the total revenue for July:



	Date	Month	Day	Temperature	Rainfall	Flyers	Price	Sales	Revenue
206	24/07/2017	July	Monday	83.5	0.57	69	0.5	35	$ 17.50
207	25/07/2017	July	Tuesday	79.9	0.57	64	0.5	33	$ 16.50
208	26/07/2017	July	Wednesday	76.6	0.59	37	0.5	32	$ 16.00
209	27/07/2017	July	Thursday	97.9	0.47	74	0.5	43	$ 21.50
210	28/07/2017	July	Friday	87.4	0.51	58	0.5	38	$ 19.00
211	29/07/2017	July	Saturday	85.5	0.57	50	0.5	35	$ 17.50
212	30/07/2017	July	Sunday	78.2	0.59	52	0.5	34	$ 17.00
213	31/07/2017	July	Monday	74.6	0.61	38	0.5	32	$ 16.00
367									$ 556.50

7. Clear the filter on **Month** to show all the data.

Challenge: Find the Total Number of Flyers Distributed

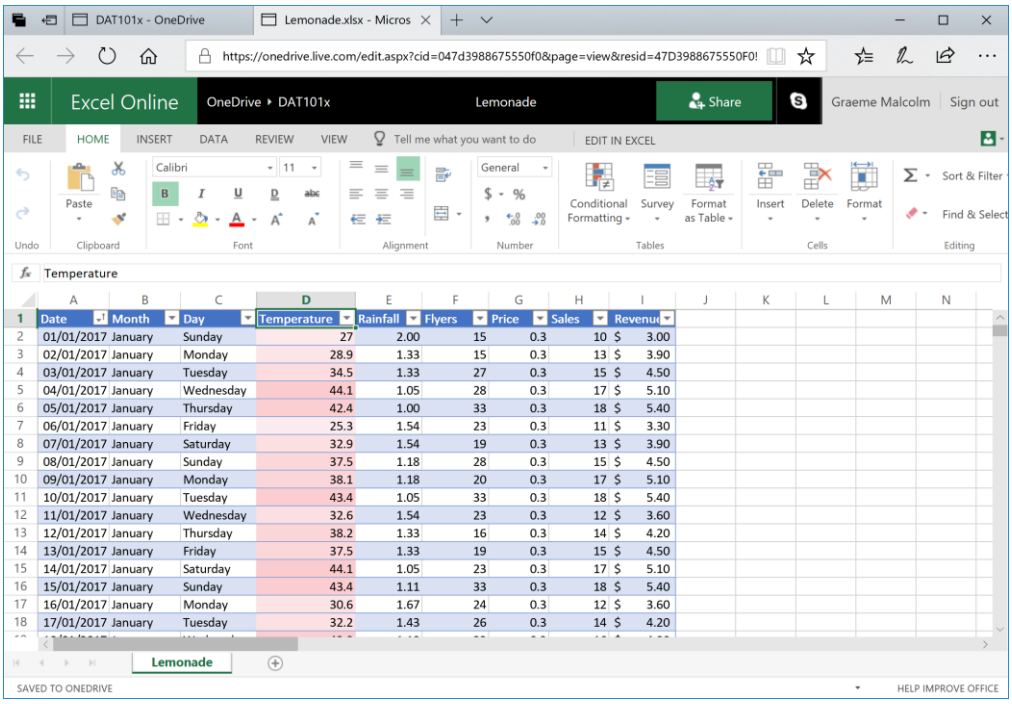
1. Add a cell under the **Flyers** column that contains the total number of flyers Rosie distributed. Format this column using the *Comma Style* (**,**) number format so that the total is formatted like 00,000.00.
2. Note the total amount for the year, and then filter the data to find the number of flyers distributed in the month of January. Don't forget to clear the filter when you're done!

Exercise 3: Using Conditional Formatting to Explore Data

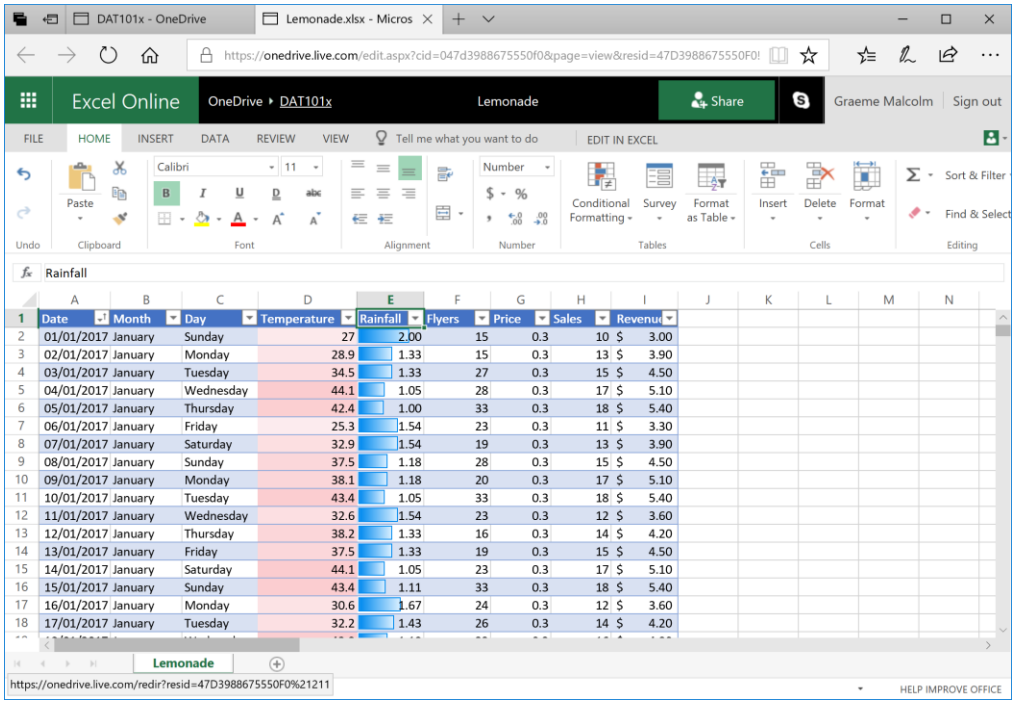
In this exercise, you will apply conditional formatting to data to highlight key values of interest.

Highlighting Extremes and Outliers

1. Select cell **D2** and then hold the **Shift** and **Ctrl** keys and press the **Down-Arrow** key to select all the values in the **Temperature** column (if you are using a Mac OSX computer, hold the **Shift** and ⌘ keys, and press the **Down-Arrow** key).
2. On the **Home** tab of the ribbon, in the **Conditional Formatting** drop-down list, point to **Color Scales**, and select the **Red-White Color Scale** (with red at the top and white at the bottom). The **Temperature** cells are reformatted so that the hottest days are colored an intense red, and the coolest days are much lighter in color intensity. Scrolling through the data now, it is easier to find days that are particularly hot or cool.

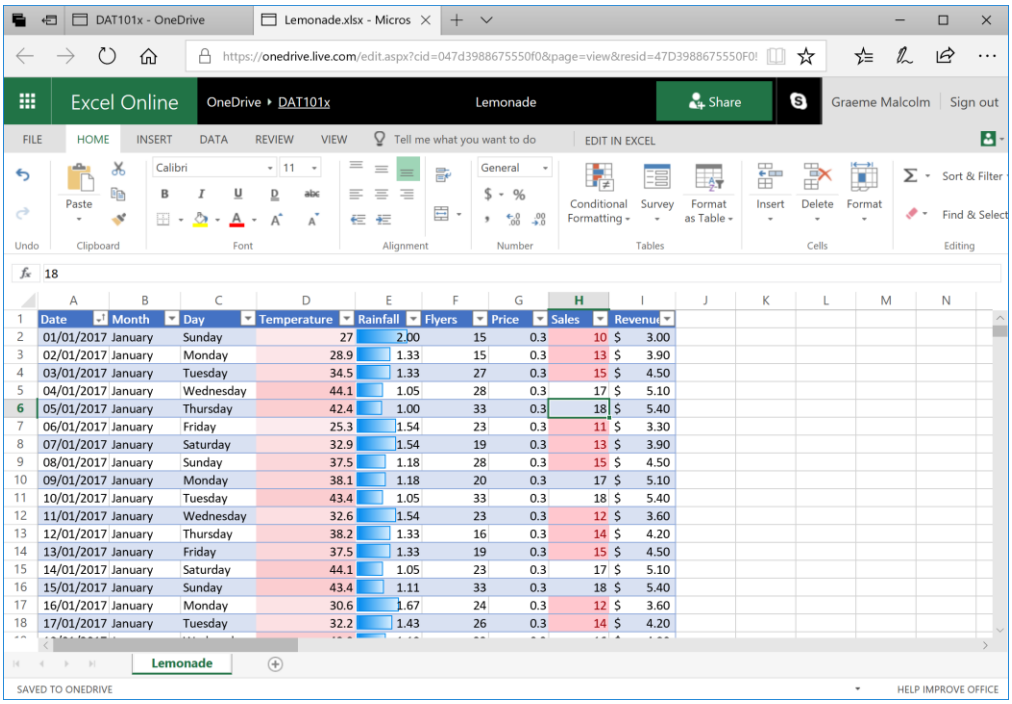


3. Select all the values in the **Rainfall** column, and then in the **Conditional Formatting** drop-down list, point to **Data Bars**, and select the **Light Blue Data Bar** gradient fill. The cells are formatted with a visual indication of the comparative level of rainfall for each day.



4. Select all the values in the **Sales** column, and then in the **Conditional Formatting** drop-down list, point to **Top/Bottom Rules**, and select **Top 10%**. Then in the **Top 10%** dialog box, select **Green Fill with Dark Green Text** and click **OK**. The cells containing sales values in the top 10% are highlighted in green (you may need to scroll to see them).
5. Reselect the values in the **Sales** column if you deselected them, and then in the **Conditional Formatting** drop-down list, point to **Top/Bottom Rules**, and select **Bottom 10%**. Then in the

Bottom 10% dialog box, select **Red Fill with Dark Red Text** and click **OK**. The cells containing sales values in the bottom 10% are highlighted in red (again, you may need to scroll to see them).



Date	Month	Day	Temperature	Rainfall	Flyers	Price	Sales	Revenue
01/01/2017	January	Sunday	27	2.00	15	0.3	10	$ 3.00
02/01/2017	January	Monday	28.9	1.33	15	0.3	13	$ 3.90
03/01/2017	January	Tuesday	34.5	1.33	27	0.3	15	$ 4.50
04/01/2017	January	Wednesday	44.1	1.05	28	0.3	17	$ 5.10
05/01/2017	January	Thursday	42.4	1.00	33	0.3	18	$ 5.40
06/01/2017	January	Friday	25.3	1.54	23	0.3	11	$ 3.30
07/01/2017	January	Saturday	32.9	1.54	19	0.3	13	$ 3.90
08/01/2017	January	Sunday	37.5	1.18	28	0.3	15	$ 4.50
09/01/2017	January	Monday	38.1	1.18	20	0.3	17	$ 5.10
10/01/2017	January	Tuesday	43.4	1.05	33	0.3	18	$ 5.40
11/01/2017	January	Wednesday	32.6	1.54	23	0.3	12	$ 3.60
12/01/2017	January	Thursday	38.2	1.33	16	0.3	14	$ 4.20
13/01/2017	January	Friday	37.5	1.33	19	0.3	15	$ 4.50
14/01/2017	January	Saturday	44.1	1.05	23	0.3	17	$ 5.10
15/01/2017	January	Sunday	43.4	1.11	33	0.3	18	$ 5.40
16/01/2017	January	Monday	30.6	1.67	24	0.3	12	$ 3.60
17/01/2017	January	Tuesday	32.2	1.43	26	0.3	14	$ 4.20

Challenge: Compare Temperature, Rainfall, and Sales

Now that you've highlighted the cells, you can more easily make visual comparisons between temperature, rainfall, and sales values.

Scroll through the data, and just by looking at the visual formatting you've added, try to see if you can spot any relationship between temperature, rainfall, and sales that might form the basis of a hypothesis you'll want to investigate more thoroughly.