Introduction to Data Science

Lab 2 – Analyzing and Visualizing Data in Microsoft Excel Online

Overview

In the previous lab, you explored a dataset containing details of lemonade sales.

In this lab, you will analyze this data further, and create visualizations to help you gain insights from the data.

What You'll Need

To complete the labs, you will need the following:

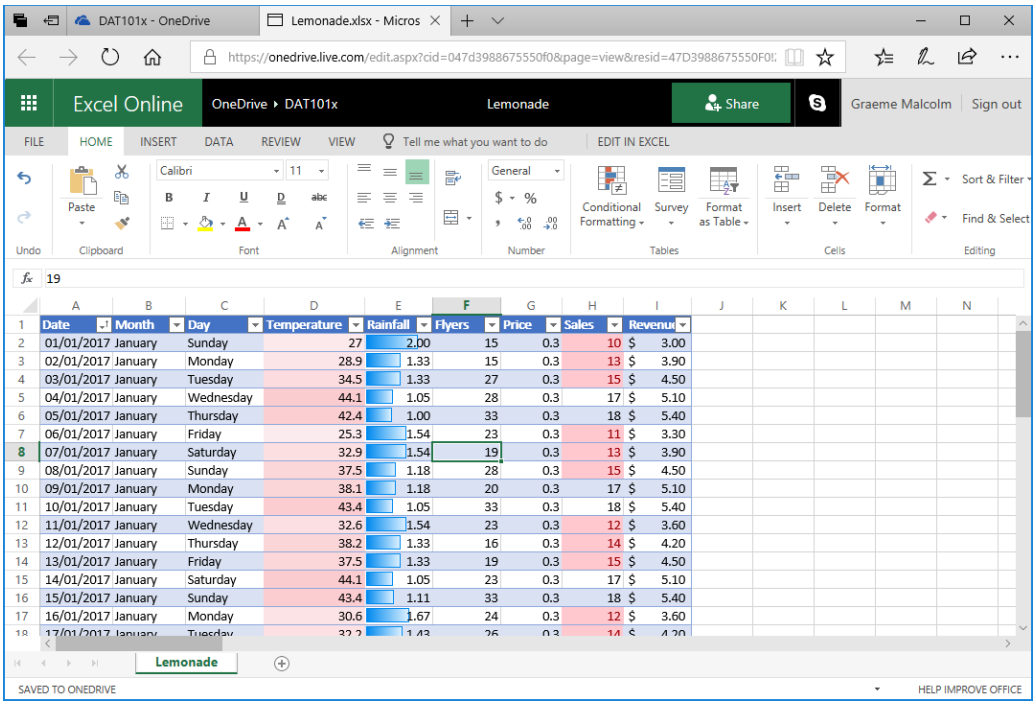
- A Windows, Linux, or Mac OS X computer with a web browser.
- A Microsoft account (for example a *hotmail.com*, *live.com*. or *outlook.com* account). If you do not already have a Microsoft account, sign up for one at https://signup.live.com.
- The **Lemonade.xlsx** workbook from the previous lab in your OneDrive folder.

Exercise 1: Analyzing Data with a PivotTable

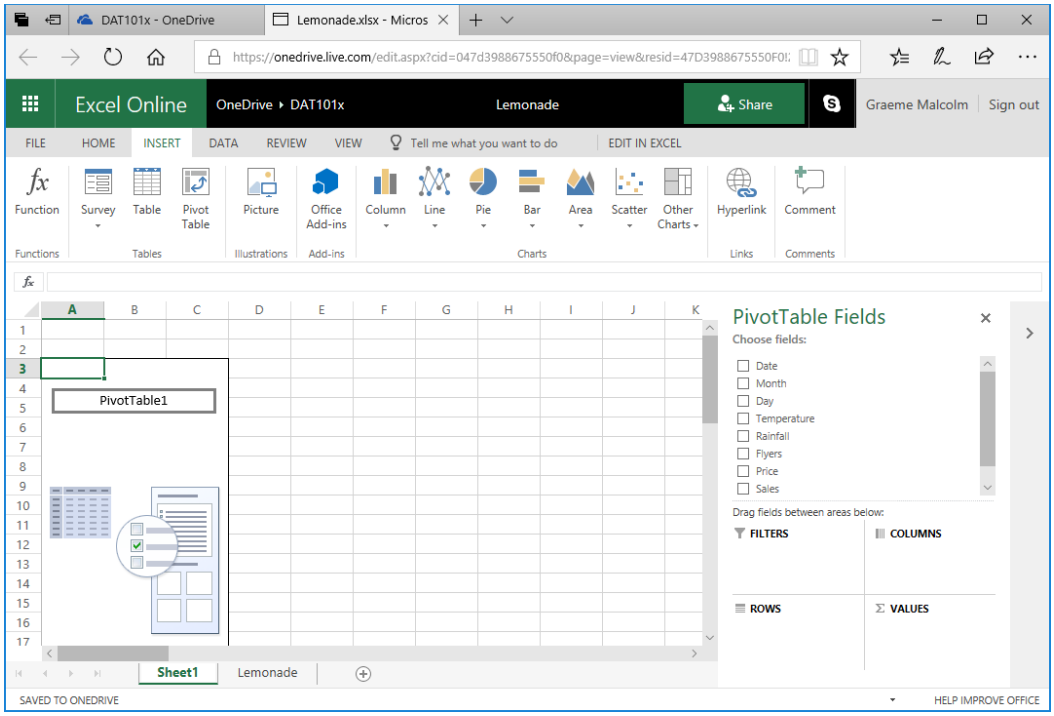
PivotTables are an excellent way to "slice and dice" data, summarizing numeric measures by one or more dimensions. In this exercise, you will use a PivotTable to view the lemonade data, aggregated in various ways.

Create a PivotTable

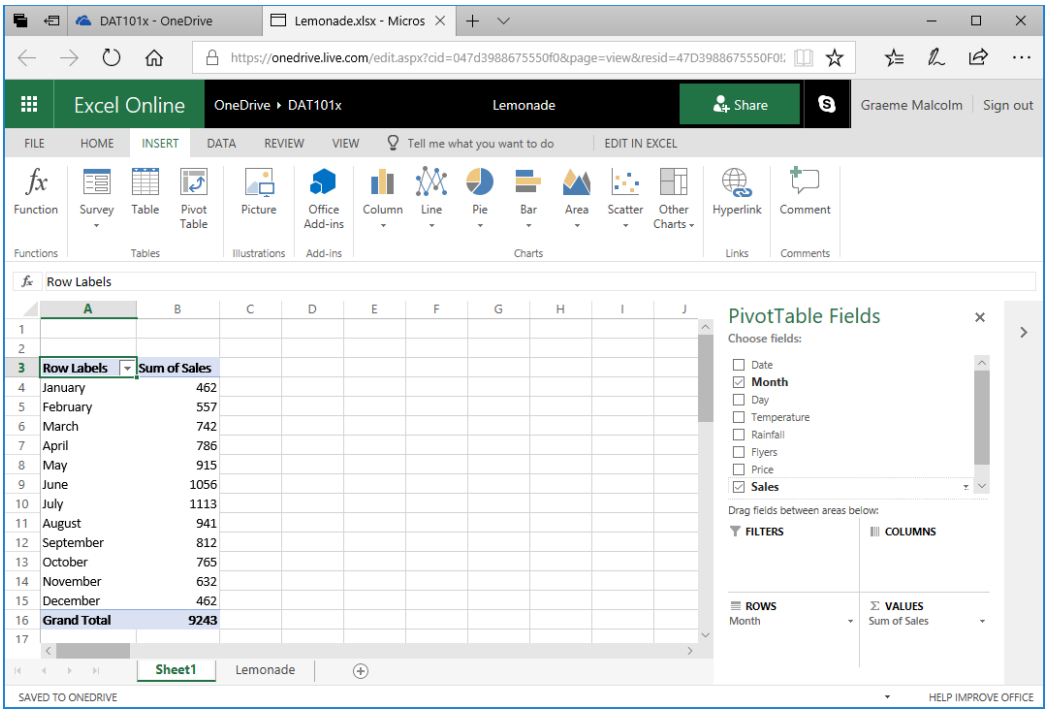
1. If you have not already done so, in your web browser, navigate to https://onedrive.live.com, and sign in using your Microsoft account credentials. Then open the **Lemonade.xlsx** workbook in the folder where you uploaded it in the previous lab. Your workbook should look like this:



2. Select any cell in the table of data, and on the **Insert** tab of the ribbon, click **PivotTable**, and create a PivotTable from your table of data in a new worksheet. Excel adds a new worksheet with a PivotTable that looks like this:



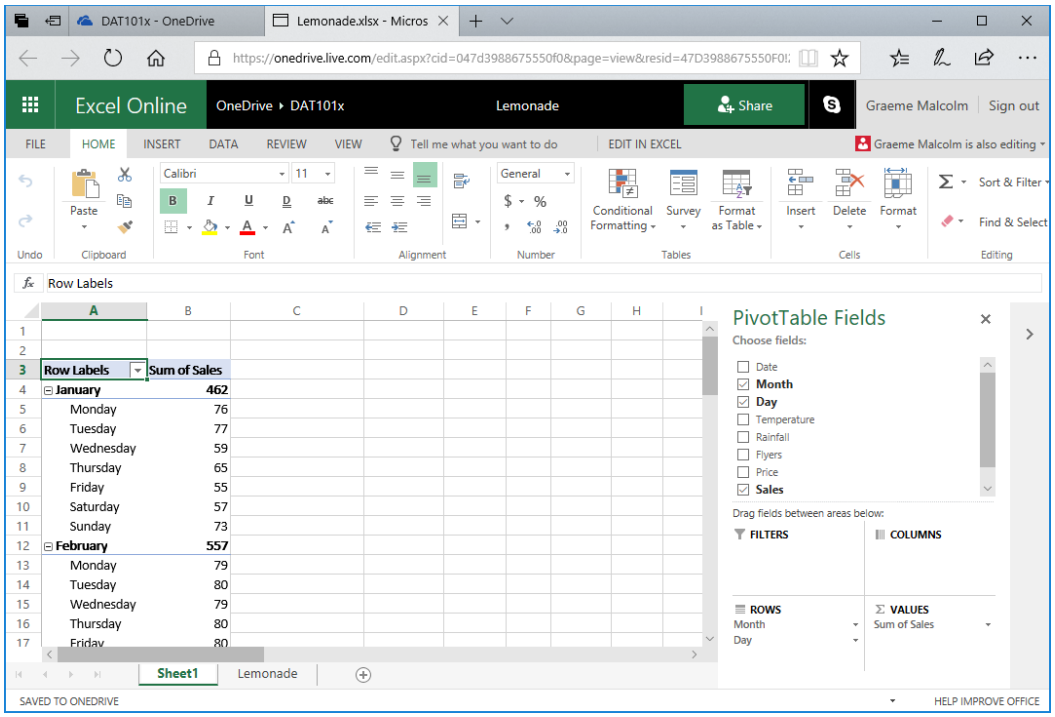
3. In the **PivotTable Fields** pane, select **Month**. Excel automatically adds **Month** to the **Rows** area of the PivotTable and displays the month names in chronological order.
4. In the **PivotTable Fields** pane, select **Sales**. Excel automatically adds **Sales** to the **Values** area of the PivotTable and displays the total number (sum) of lemonade sales for each month, like this:



You can now see the sales aggregated by month – so for example, there were 1,056 sales in June.

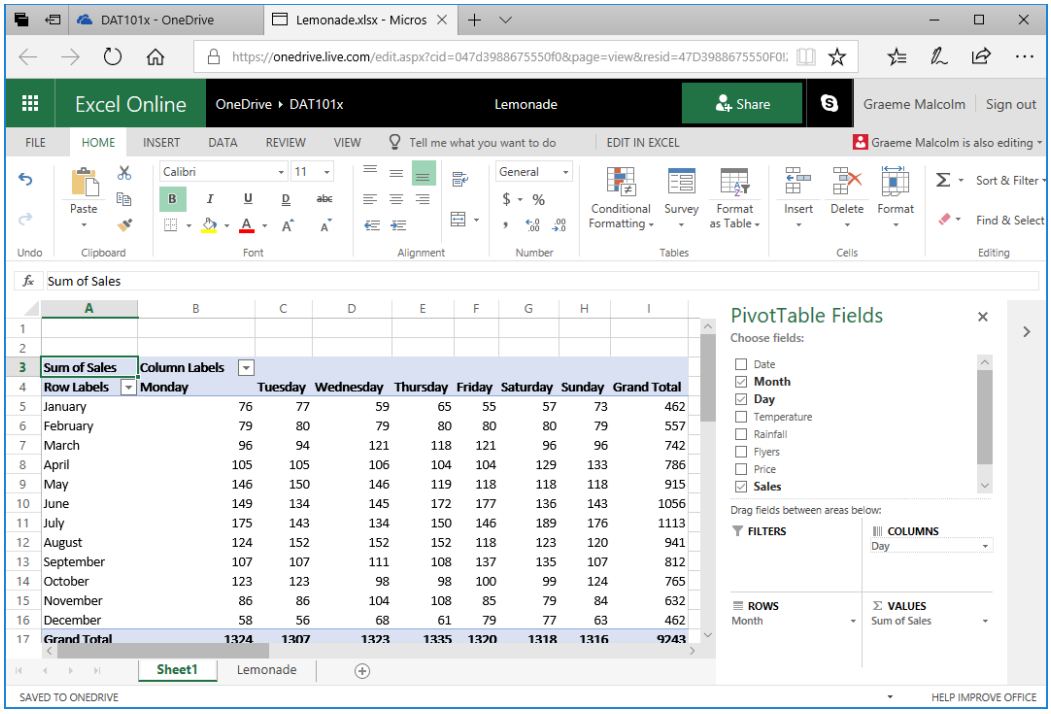
Add a Second Dimension

1. In the **PivotTable Fields** pane, select **Day**. Excel automatically adds **Day** to the **Rows** area of the PivotTable and displays the total number (sum) of lemonade sales for each weekday within each month, like this:



Now you can see monthly sales aggregated by weekday. For example, 57 of the sales in January were made on a Saturday. You can also expand/collapse months to “drill-up”/”drill-down” the levels of the hierarchy.

2. In the **PivotTable Fields** pane, drag **Day** from the **Rows** area to the **Columns** area. Excel now shows total sales for each month on rows, broken down by weekday in columns; like this:

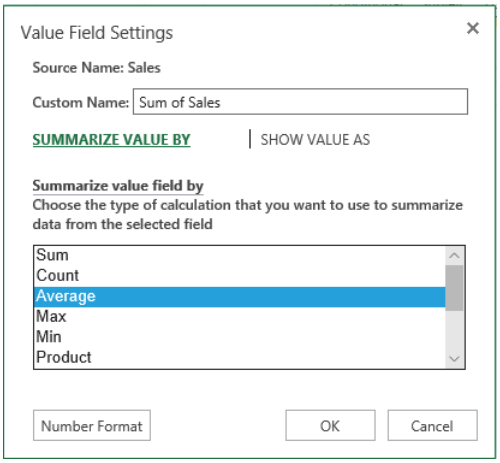


Sum of Sales	Column Labels							
Row Labels	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday	Grand Total
January	76	77	59	65	55	57	73	462
February	79	80	79	80	80	80	79	557
March	96	94	121	118	121	96	96	742
April	105	105	106	104	104	129	133	786
May	146	150	146	119	118	118	118	915
June	149	134	145	172	177	136	143	1056
July	175	143	134	150	146	189	176	1113
August	124	152	152	152	118	123	120	941
September	107	107	111	108	137	135	107	812
October	123	123	98	98	100	99	124	765
November	86	86	104	108	85	79	84	632
December	58	56	68	61	79	77	63	462
Grand Total	1324	1307	1323	1335	1320	1318	1316	9243

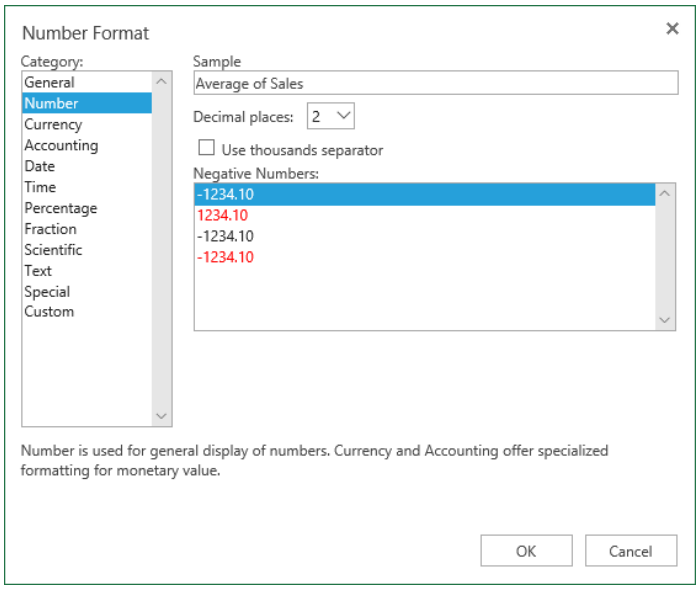
You can still see monthly sales broken down by weekday, but you can also see (in the bottom row) the totals for each week day across the entire year. For example, a total of 1,324 sales were made on a Monday.

Change the Aggregation

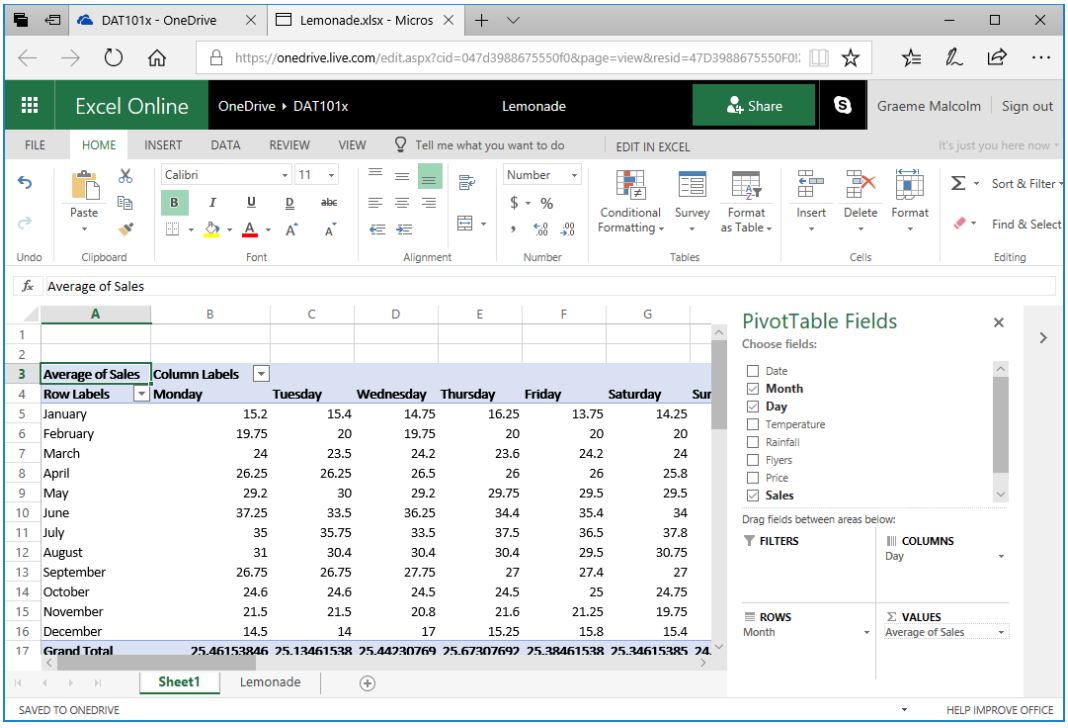
1. In the **PivotTable Fields** pane, in the **Values** area, click the drop-down arrow next to **Sum of Sales**, and then click **Value Field Settings**.
2. In the **Value Field Settings** dialog box, select **Average** as shown here, and then click **Number Format**:



3. In the **Number Format** dialog box, select the **Number** category and ensure that **Decimal places** is set to **2** as shown here. Then click **OK**.



The table of data now shows the average number of sales for each month and weekday, as shown here.



Average of Sales	Column Labels						
Row Labels	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sur
January	15.2	15.4	14.75	16.25	13.75	14.25	
February	19.75	20	19.75	20	20	20	
March	24	23.5	24.2	23.6	24.2	24	
April	26.25	26.25	26.5	26	26	25.8	
May	29.2	30	29.2	29.75	29.5	29.5	
June	37.25	33.5	36.25	34.4	35.4	34	
July	35	35.75	33.5	37.5	36.5	37.8	
August	31	30.4	30.4	30.4	29.5	30.75	
September	26.75	26.75	27.75	27	27.4	27	
October	24.6	24.6	24.5	24.5	25	24.75	
November	21.5	21.5	20.8	21.6	21.25	19.75	
December	14.5	14	17	15.25	15.8	15.4	
Grand Total	25.46153846	25.13461538	25.44230769	25.67307692	25.38461538	25.34615385	24

You can now see the average number of sales for each weekday by month. For example, the average number of sales on a Wednesday in February is 19.75.

Challenge: PivotTable Analysis

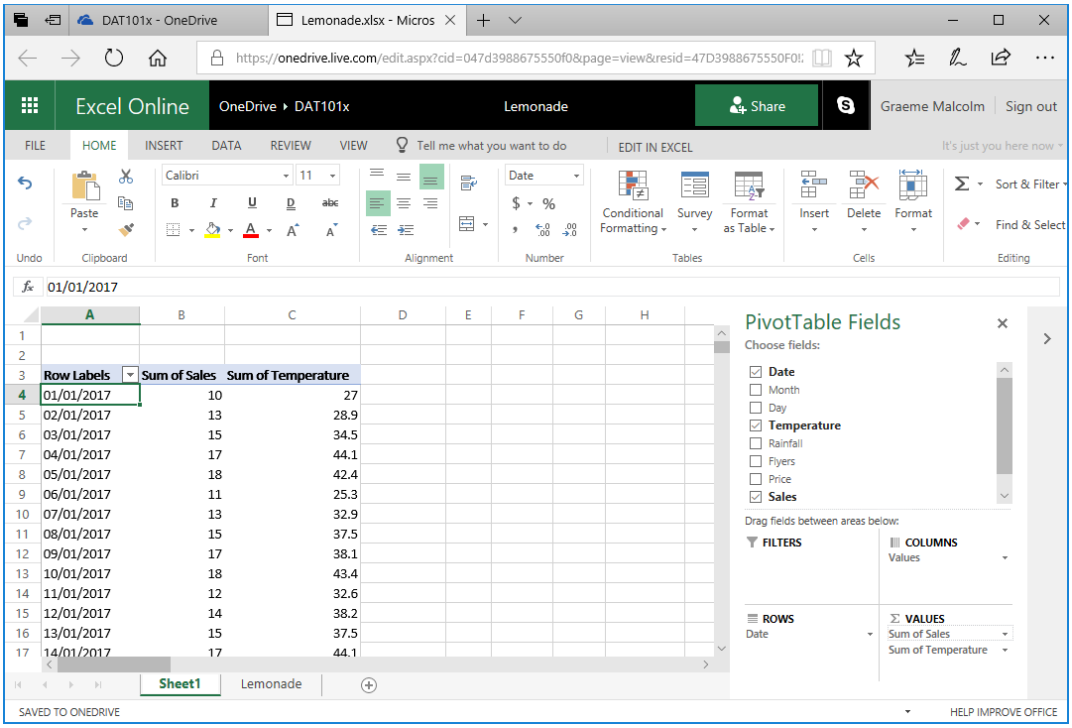
1. Modify the fields in the PivotTable to find the following information:
   - The total sum of revenue for August.
   - The temperature on the hottest Saturday in July.
   - The lowest number of flyers distributed in a day during November.

Exercise 2: Visualizing Data with Charts

It can often be easier to identify trends and relationships in data by creating data visualizations such as charts.

View the Sales Trend for the Year

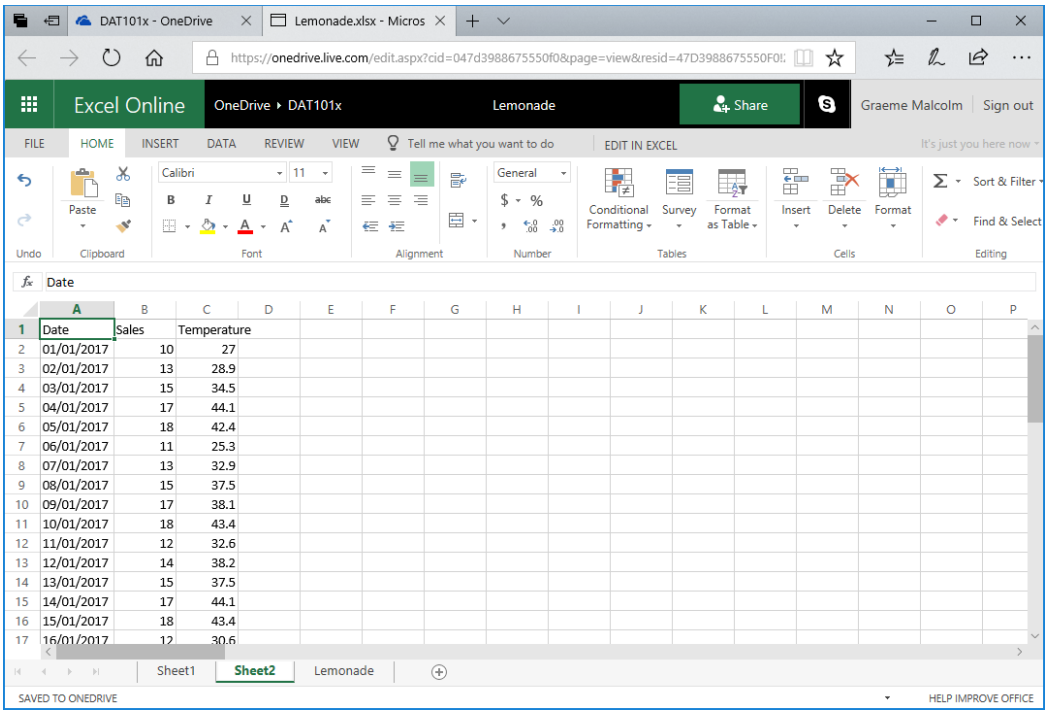
1. Modify the PivotTable you created in the previous exercise so that it shows **Date** in the **Rows** area and the sum of **Sales** and sum of **Temperature** (in that order) in the **Values** area, like this:



Row Labels	Sum of Sales	Sum of Temperature
01/01/2017	10	27
02/01/2017	13	28.9
03/01/2017	15	34.5
04/01/2017	17	44.1
05/01/2017	18	42.4
06/01/2017	11	25.3
07/01/2017	13	32.9
08/01/2017	15	37.5
09/01/2017	17	38.1
10/01/2017	18	43.4
11/01/2017	12	32.6
12/01/2017	14	38.2
13/01/2017	15	37.5
14/01/2017	17	44.1

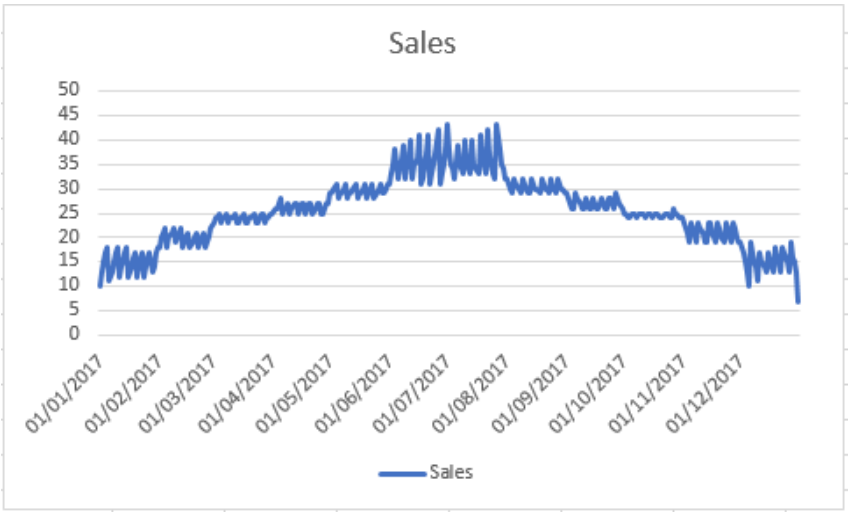
Note that the date may be formatted differently for your location.

2. Select the cells containing the daily sales data, but not the **Row Labels**, **Sum of Sales**, and **Sum of Temperature** header cells or the **Grand Total** footer cells; and then on the **Home** tab of the ribbon, click the **Copy** button (📋) to copy the selected cells to the clipboard.
3. Under the worksheet, click the **New Sheet** button (**+**) to add a new worksheet to the workbook.
4. In the new sheet, select cell A2, and then on the Home tab click the **Paste** button (📋) to paste the copied cells into the new worksheet. You may need to widen the A column to see the dates.
5. In cells A1 to C1, add the columns headers **Date, Sales**, and **Temperature**. Your new worksheet should look like this:



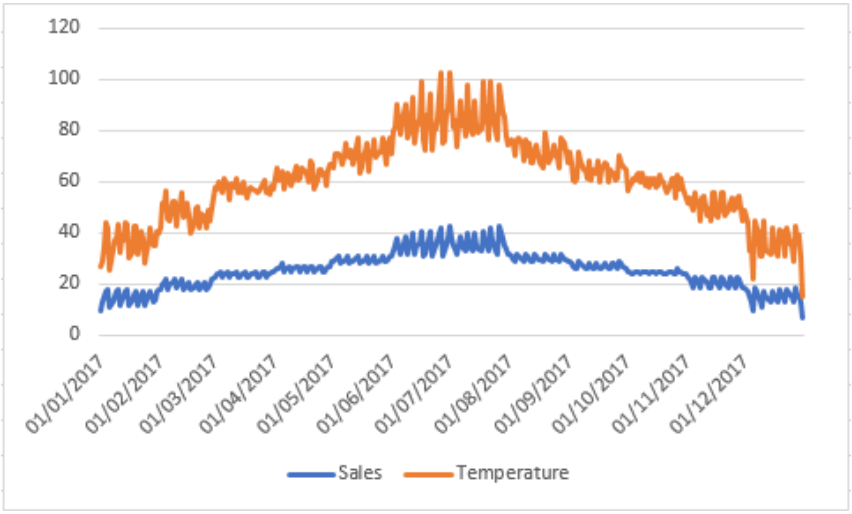
	A	B	C
1	Date	Sales	Temperature
2	01/01/2017	10	27
3	02/01/2017	13	28.9
4	03/01/2017	15	34.5
5	04/01/2017	17	44.1
6	05/01/2017	18	42.4
7	06/01/2017	11	25.3
8	07/01/2017	13	32.9
9	08/01/2017	15	37.5
10	09/01/2017	17	38.1
11	10/01/2017	18	43.4
12	11/01/2017	12	32.6
13	12/01/2017	14	38.2
14	13/01/2017	15	37.5
15	14/01/2017	17	44.1
16	15/01/2017	18	43.4
17	16/01/2017	12	30.6

6. Select the **Date** and **Sales** data, including the headers (but not the temperature data). Then on the **Insert** tab of the ribbon, in the **Line** drop-down list, click the first line chart format. Excel inserts a line chart like this:



Note that the line chart shows daily fluctuations in sales, but the general trend seems to indicate that sales are higher during the summer months and lower at the beginning and end of the year.

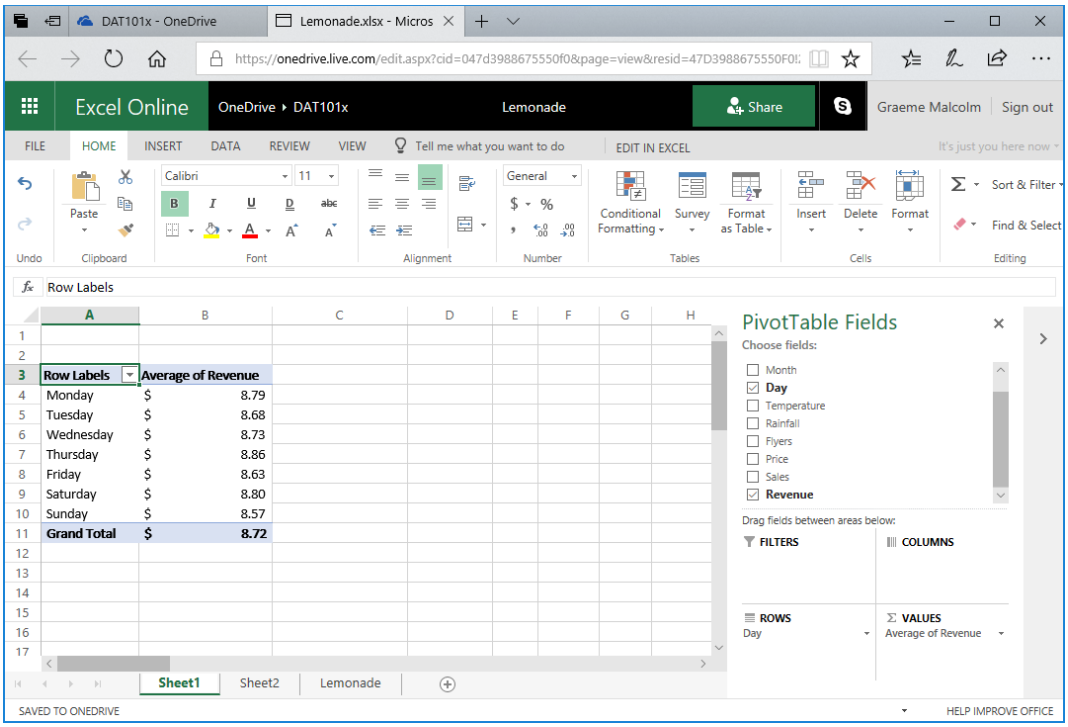
7. Delete the chart, and then select all the data and headers, including **Temperature** and insert a new line chart. This inserts a chart like this:



This time, the chart includes separate series for **Sales** and **Temperature**. Both series show a similar pattern; it seems sales and temperature both increase over the summer months.

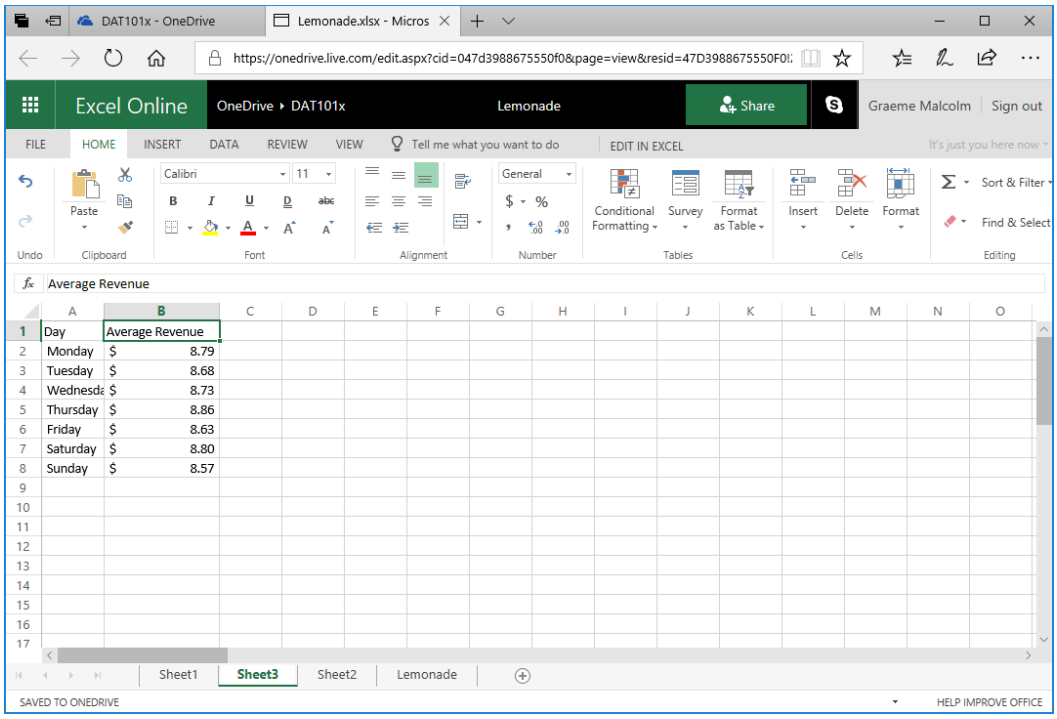
View Revenue by Weekday

1. Return to the worksheet containing the PivotTable, and modify it to show **Day** on rows with the *average* of **Revenue** in the **Accounting** number format, like this:

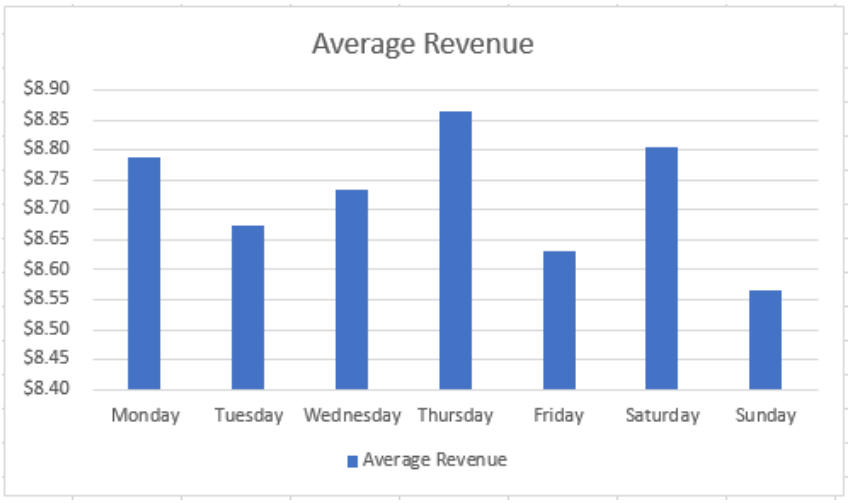


Row Labels	Average of Revenue
Monday	$ 8.79
Tuesday	$ 8.68
Wednesday	$ 8.73
Thursday	$ 8.86
Friday	$ 8.63
Saturday	$ 8.80
Sunday	$ 8.57
Grand Total	**$ 8.72**

2. Copy the day and average revenue values (but not the headers or total) to the clipboard, and then add a worksheet, paste the copied data in cell **A2**, and add **Day** and **Average Revenue** headers like this:

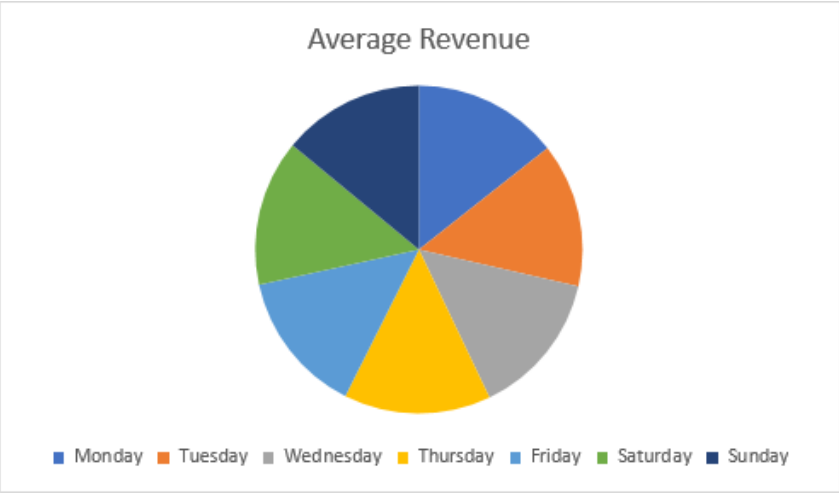


3. Select all the data, including the Day and Average Revenue headers, and on the Insert tab of the ribbon, in the **Column** drop-down list, select the first column chart format. A chart like this is created:



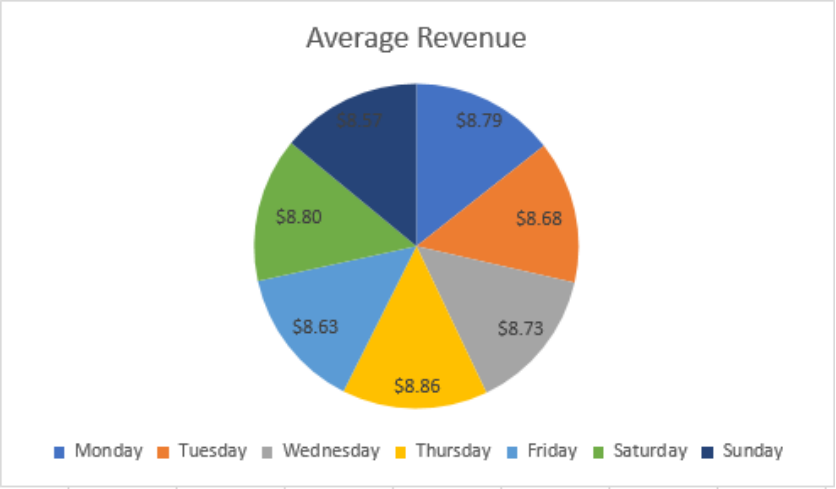
At first glance this chart appears to show some significant variation between average revenue of different days of the week; with revenue on Thursdays much higher than on Sundays. However, look more closely at the scale on the vertical (Y) axis – The difference is less than 30 cents.

4. Select the column chart, and on the **Chart** tab of the ribbon, in the **Pie** drop-down list select the 2D Pie chart format. The chart changes to a pie chart like this:



Note that the pie segments are more or less the same size for each day.

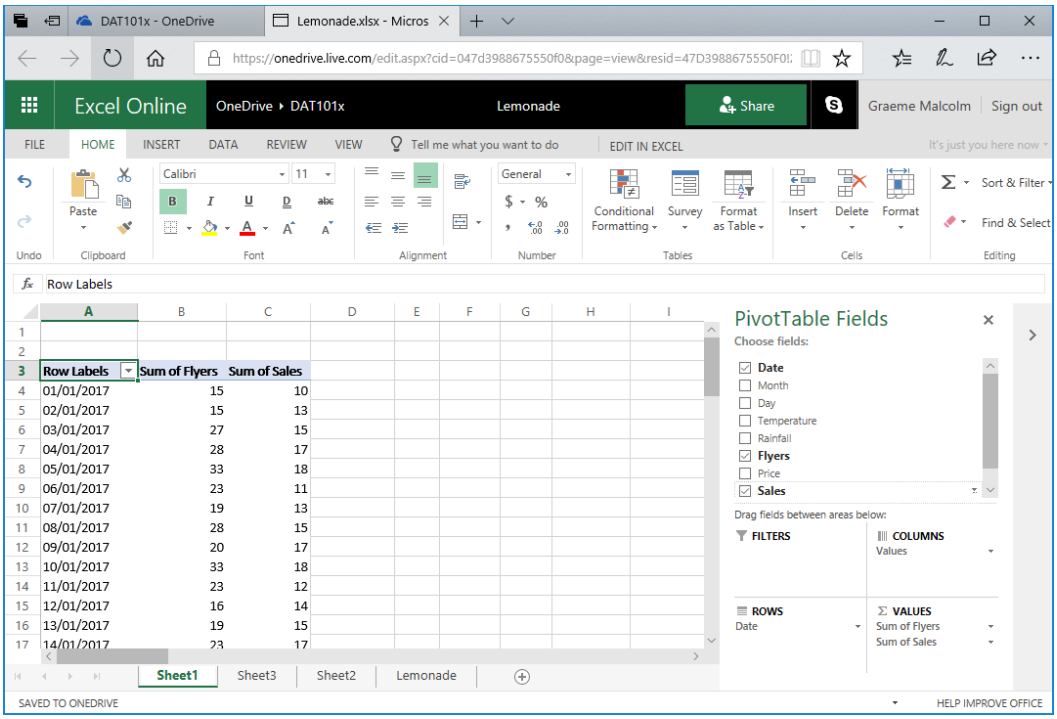
5. Select the pie chart and on the **Chart** tab, in the **Data Labels** drop-down list, select **Inside End**. This displays the actual data amounts in the chart, like this:



Now it's clearer that there's little apparent variation in average revenue for different days of the week.

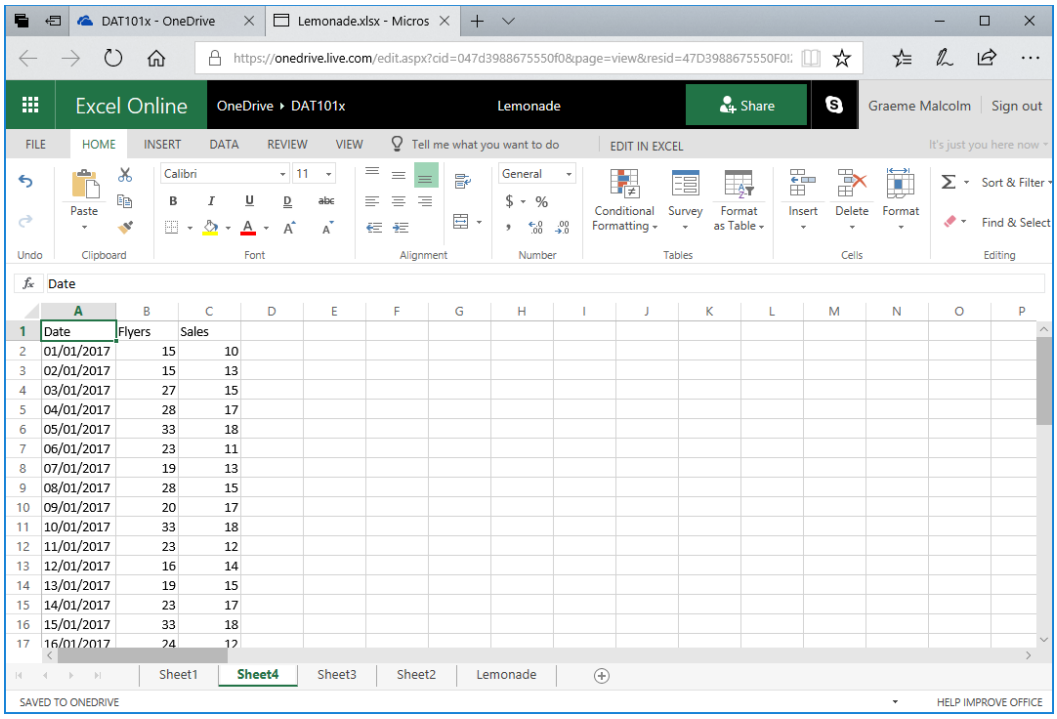
View Sales by Flyers

1. Return to the worksheet containing the PivotTable, and modify it to show **Date** on rows with the sum of **Flyers** and the sum of **Sales** as values in **General** number format, like this:



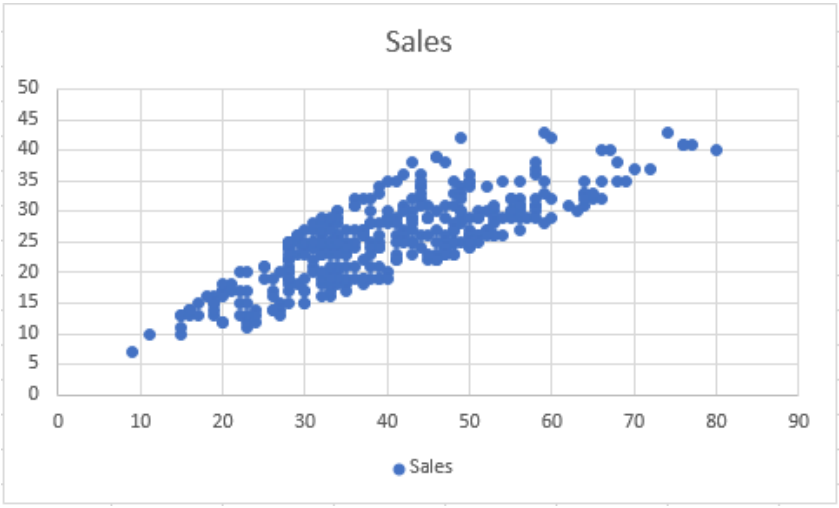
Row Labels	Sum of Flyers	Sum of Sales
01/01/2017	15	10
02/01/2017	15	13
03/01/2017	27	15
04/01/2017	28	17
05/01/2017	33	18
06/01/2017	23	11
07/01/2017	19	13
08/01/2017	28	15
09/01/2017	20	17
10/01/2017	33	18
11/01/2017	23	12
12/01/2017	16	14
13/01/2017	19	15
14/01/2017	23	17

2. Copy the date, flyers, and sales values (but not the headers or totals) to a new worksheet and add **Date**, **Flyers**, and **Sales** headers like this:



Date	Flyers	Sales
01/01/2017	15	10
02/01/2017	15	13
03/01/2017	27	15
04/01/2017	28	17
05/01/2017	33	18
06/01/2017	23	11
07/01/2017	19	13
08/01/2017	28	15
09/01/2017	20	17
10/01/2017	33	18
11/01/2017	23	12
12/01/2017	16	14
13/01/2017	19	15
14/01/2017	23	17
15/01/2017	33	18
16/01/2017	24	12

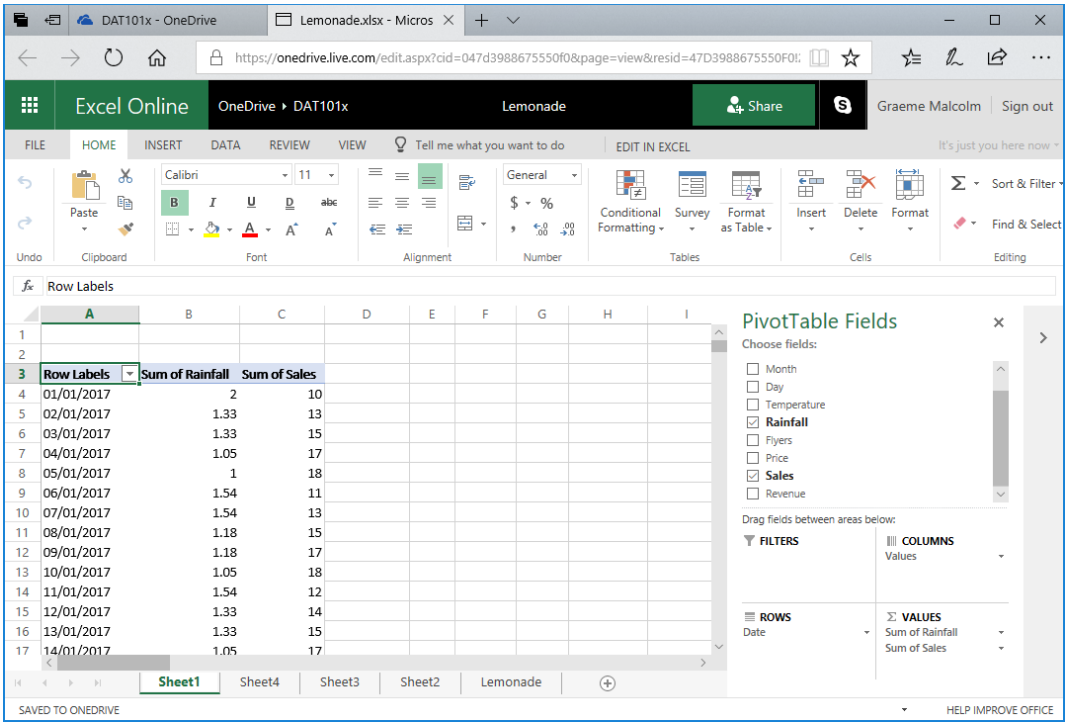
3. Select the **Flyers** and **Sales** data and headers (but not the dates). Then on the **Insert** tab, in the **Scatter** drop-down list, select the first scatter-plot format. This creates a scatter-plot chart like this:



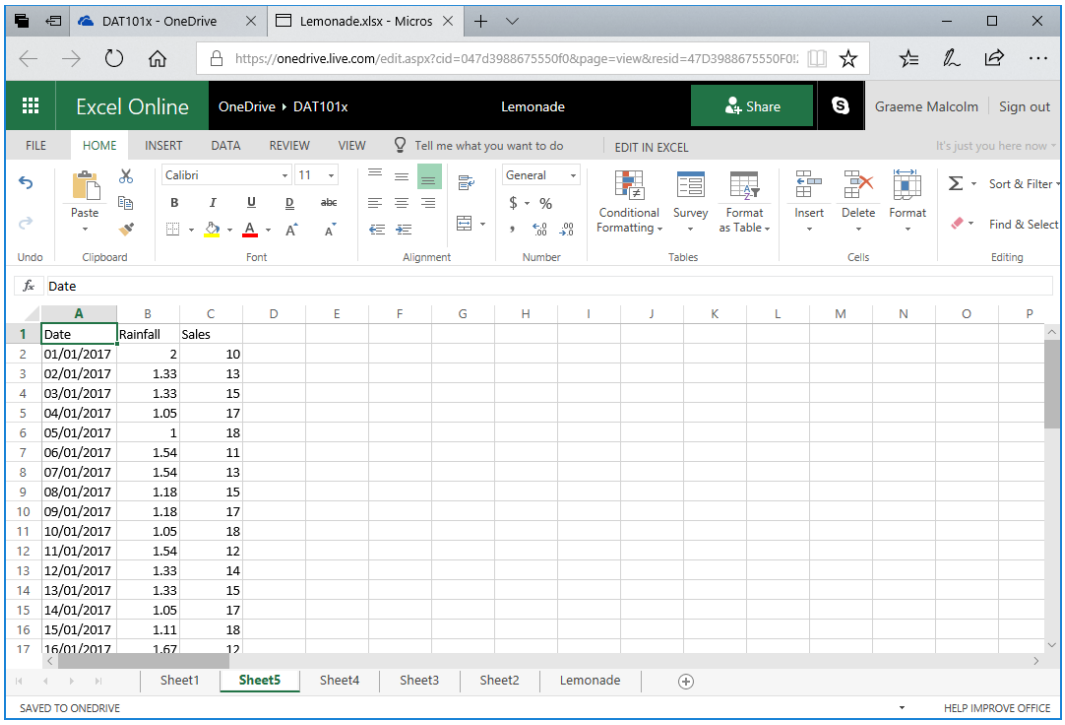
Note that the chart shows the number of flyers distributed each day on the horizontal (X) axis, and the number of sales each day on the vertical (Y) axis. The plot forms a roughly diagonal line (with some variance), indicating a general trend where the number of sales tends to increase in-line with the number of flyers distributed.

View Sales by Rainfall

1. Return to the worksheet containing the PivotTable, and modify it to show **Date** on rows with the sum of **Rainfall** and the sum of **Sales** as values in **General** number format, like this:

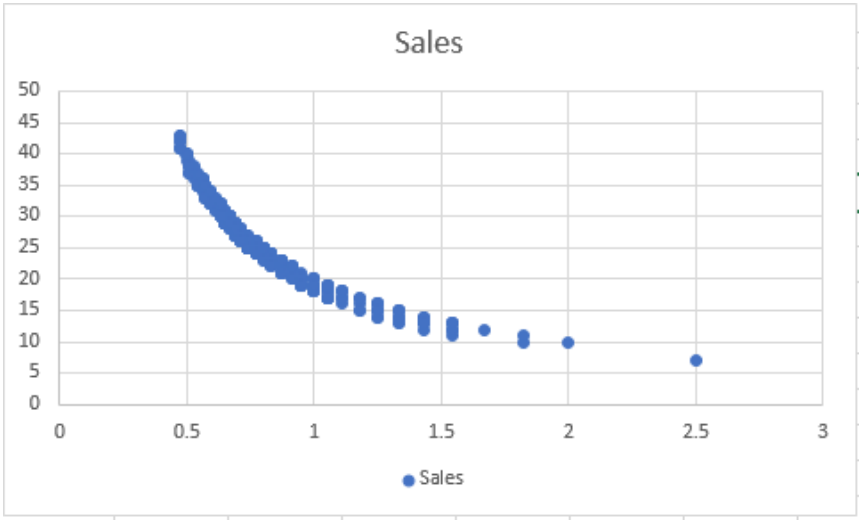


2. Copy the date, rainfall, and sales values (but not the headers or totals) to a new worksheet and add **Date**, **Rainfall**, and **Sales** headers like this:



	A	B	C
1	Date	Rainfall	Sales
2	01/01/2017	2	10
3	02/01/2017	1.33	13
4	03/01/2017	1.33	15
5	04/01/2017	1.05	17
6	05/01/2017	1	18
7	06/01/2017	1.54	11
8	07/01/2017	1.54	13
9	08/01/2017	1.18	15
10	09/01/2017	1.18	17
11	10/01/2017	1.05	18
12	11/01/2017	1.54	12
13	12/01/2017	1.33	14
14	13/01/2017	1.33	15
15	14/01/2017	1.05	17
16	15/01/2017	1.11	18
17	16/01/2017	1.67	12

3. Select the **Rainfall** and **Sales** data and headers (but not the dates). Then on the **Insert** tab, in the **Scatter** drop-down list, select the first scatter-plot format. This creates a scatter-plot chart like this:



This plot seems to indicate some kind of relationship between rainfall and sales, with sales falling as rainfall increases. However, the line formed by the plots is curved. This often means there is a non-linear, possibly logarithmic relationship.

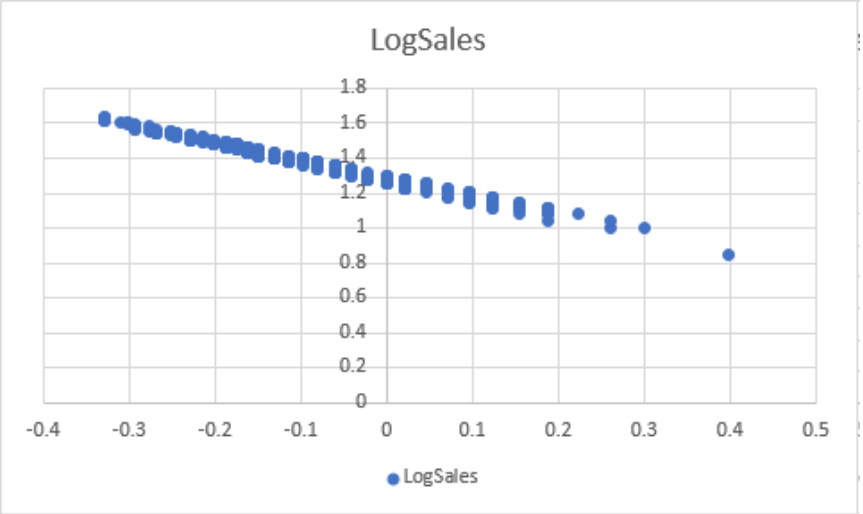
4. Move the chart so you can see the empty D and E columns after the daily rainfall and sales data.
5. In D1, add the column header **LogRainfall**, and in cell D2 enter the following formula to calculate the base 10 log of the rainfall value:

```
=log(B2)
```

6. Copy the formula to the other cells in the **LogRainfall** column. The easiest way to do this is to select the cell containing the formula and double-click on the small square "handle" at the bottom right of the selected cell.
7. In E1, add the column header **LogSales**, and in cell E2 enter the following formula to calculate the base 10 log of the rainfall value:

   ```
   =log(C2)
   ```

8. Copy the formula to the other cells in the **LogSales** column.
9. Select the **LogRainfall** and **LogSales** data and headers. Then on the **Insert** tab, in the **Scatter** drop-down list, select the first scatter-plot format. This creates a scatter-plot chart like this:



   Note that this plot shows a linear relationship between the log of rainfall and the log of sales. This is potentially useful as we explore relationships in the data, as it is easier to calculate a linear equation that relates rainfall to sales than to define a logarithmic equation to do the same.

Challenge: Visualizing Data

1. Create a column chart showing the sum of flyers distributed on each day of the week and note the days on which the highest and lowest number of flyers were distributed.
2. Create a scatter plot showing daily temperature and rainfall and examine the apparent relationship between these fields.