

# 1

## Everything You Ever Needed to Know about Spreadsheets but Were Too Afraid to Ask

This book relies on you having a working knowledge of spreadsheets, and I'm going to assume that you already understand the basics. If you've never used a formula before in your life, then you've got a slight uphill battle here. I'd recommend going through a *For Dummies* book or some other intro-level tutorial for Excel before diving into this.

That said, even if you're a seasoned Excel veteran, there's some functionality that'll keep cropping up in this text that you may not have had to use before. It's not difficult stuff; just things I've noticed not everyone has used in Excel. You'll be covering a wide variety of little features in this chapter, and the example at this stage might feel a bit disjointed. But you can learn what you can here, and then, when you encounter it organically later in the book, you can slip back to this chapter as a reference.

As Samuel L. Jackson says in *Jurassic Park*, "Hold on to your butts!"

### EXCEL VERSION DIFFERENCES

As mentioned in the book's introduction, these chapters work with Excel 2007, 2010, 2013, 2011 for Mac, and LibreOffice. Sadly, in each version of Excel, Microsoft has moved stuff around for the heck of it.

For example, things on the Layout tab on 2011 are on the View tab in the other versions. Solver is the same in 2010 and 2013, but the performance is actually better in 2007 and 2011 even though 2007's Solver interface is grotesque.

The screen captures in this text will be from Excel 2011. If you have an older or newer version, sometimes your interactions will look a little different—mostly when it comes to where things are on the menu bar. I will do my best to call out these differences. If you can't find something, Excel's help feature and Google are your friends.

The good news is that whenever we're in the "spreadsheet part of the spreadsheet," everything works exactly the same.

As for LibreOffice, if you've chosen to use open source software for this book, then I'm assuming you're a do-it-yourself kind of person, and I won't be referencing the LibreOffice interface directly. Never you mind, though. It's a dead ringer for Excel.

## Some Sample Data

### NOTE

The Excel workbook used in this chapter, “Concessions.xlsx,” is available for download at the book’s website at [www.wiley.com/go/datasmart](http://www.wiley.com/go/datasmart).

Imagine you’ve been terribly unsuccessful in life, and now you’re an adult, still living at home, running the concession stand during the basketball games played at your old high school. (I swear this is only semi-autobiographical.)

You have a spreadsheet full of last night’s sales, and it looks like Figure 1-1.

	A	B	C	D
1	Item	Category	Price	Profit
2	Beer	Beverages	\$ 4.00	50%
3	Hamburger	Hot Food	\$ 3.00	67%
4	Popcorn	Hot Food	\$ 5.00	80%
5	Pizza	Hot Food	\$ 2.00	25%
6	Bottled Water	Beverages	\$ 3.00	83%
7	Hot Dog	Hot Food	\$ 1.50	67%
8	Chocolate Dipped Cone	Frozen Treat	\$ 3.00	50%
9	Soda	Beverages	\$ 2.50	80%
10	Chocolate Bar	Candy	\$ 2.00	75%
11	Hamburger	Hot Food	\$ 3.00	67%
12	Beer	Beverages	\$ 4.00	50%
13	Hot Dog	Hot Food	\$ 1.50	67%
14	Licorice Rope	Candy	\$ 2.00	50%
15	Chocolate Dipped Cone	Frozen Treat	\$ 3.00	50%

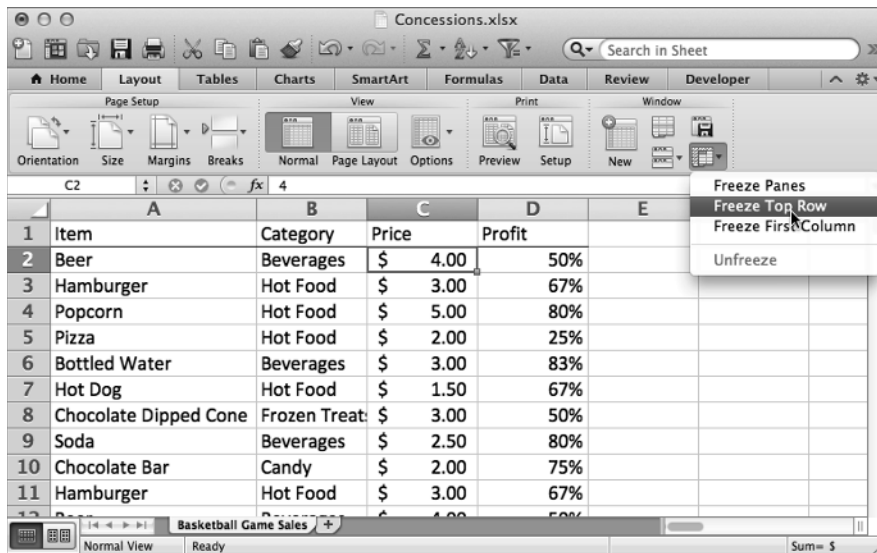
**Figure 1-1:** Concession stand sales

Figure 1-1 shows each sale, what the item was, what type of food or drink it was, the price, and the percentage of the sale going toward profit.

## Moving Quickly with the Control Button

If you want to peruse the records, you can scroll down the sheet with your scroll wheel, track pad, or down arrow. As you scroll, it’s helpful to keep the header row locked at the top of the sheet, so you can remember what each column means. To do that, choose

Freeze Panes or Freeze Top Row from the “View” tab on Windows (“Layout” tab on Mac 2011 as shown in Figure 1-2).



**Figure 1-2:** Freezing the top row

To move quickly to the bottom of the sheet to look at how many transactions you have, you can select a value in one of the populated columns and press **Ctrl+↓** (Command+↓ on a Mac). You’ll zip right to the last populated cell in that column. In this sheet, the final row is 200. Also, note that using **Ctrl/Command** to jump around the sheet from left to right works much the same.

If you want to take an average of the sales prices for the night, below the price column, column C, you can jot the following formula:

```
=AVERAGE(C2:C200)
```

The average is \$2.83, so you won’t be retiring wealthy anytime soon. Alternatively, you can select the last cell in the column, C200, hold **Shift+Ctrl+↑** to highlight the whole column, and then select the Average calculation from the status bar in the bottom right of the spreadsheet to see the simple summary statistic (see Figure 1-3). On Windows, you’ll need to right-click the status bar to select the average if it’s not there. On Mac, if your status bar is turned off, click the View menu and select “Status Bar” to turn it on.

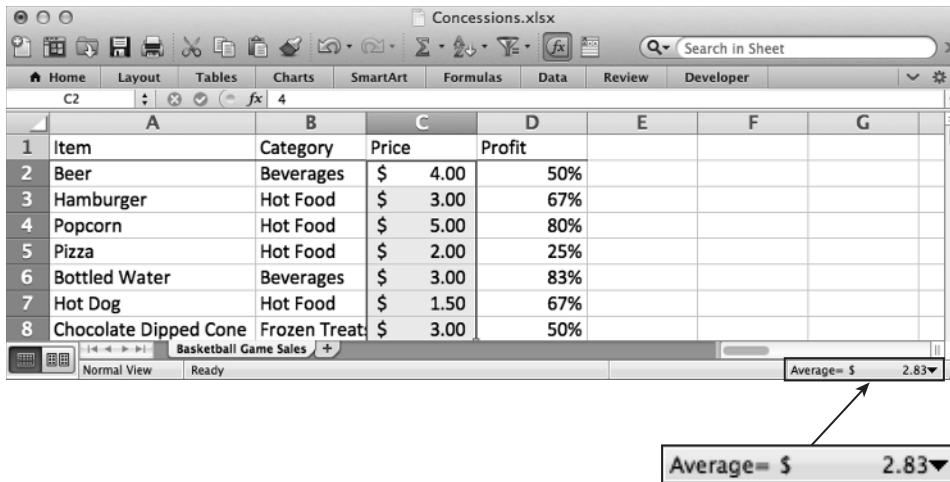


Figure 1-3: Average of the price column in the status bar

## Copying Formulas and Data Quickly

Perhaps you'd like to view your profits in actual dollars rather than as percentages. You can add a header to column E called "Actual Profit." In E2, you need only to multiply the price and profit columns together to obtain this:

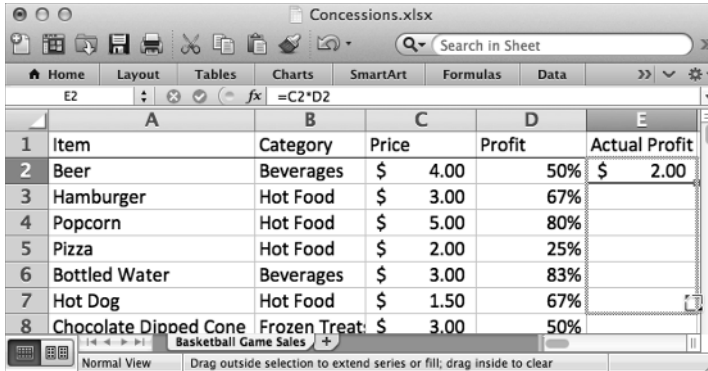
=C2\*D2

For beer, it's \$2. You don't have to rewrite this formula in every cell in the column. Instead, Excel lets you grab the right-bottom corner of the cell and drag the formula where you like. The referenced cells in columns C and D will update relative to where you copy the formula. If, as in the case of the concession data, the column to the left is fully populated, you can double-click the bottom-right corner of the formula to have Excel fill the whole column (see Figure 1-4). Try this double-click action for yourself, because I'll be using it all over the place in this book, and if you get the hang of it now, you'll save yourself a whole lot of heartache.

Now, what if you don't want the cells in the formula to change relative to the target when they're dragged or copied? Whatever you don't want changed, just add a \$ in front of it.

For example, if you changed the formula in E2 to:

=C\$2\*D\$2



The screenshot shows an Excel spreadsheet titled 'Concessions.xlsx'. The active cell is E2, and the formula bar displays '=C2\*D2'. The spreadsheet contains the following data:

	A	B	C	D	E
1	Item	Category	Price	Profit	Actual Profit
2	Beer	Beverages	\$ 4.00	50%	\$ 2.00
3	Hamburger	Hot Food	\$ 3.00	67%	
4	Popcorn	Hot Food	\$ 5.00	80%	
5	Pizza	Hot Food	\$ 2.00	25%	
6	Bottled Water	Beverages	\$ 3.00	83%	
7	Hot Dog	Hot Food	\$ 1.50	67%	
8	Chocolate Dipped Cone	Frozen Treat	\$ 3.00	50%	

**Figure 1-4:** Filling in a formula by dragging the corner

Then when you copy the formula down, nothing changes. The formula continues to reference row 2.

If you copy the formula to the right, however, C would become D, D would become E, and so on. If you don't want that behavior, you need to put a \$ in front of the column references as well. This is called an *absolute reference* as opposed to a *relative reference*.

## Formatting Cells

Excel offers static and dynamic options for formatting values. Take a look at column E, the Actual Profit column you just created. Select column E by clicking on the gray E column label. Then right-click the selection and choose Format Cells.

From within the Format Cells menu, you can tell Excel the type of number to be found in column E. In this case you want it to be Currency. And you can set the number of decimal places. Leave it at two decimals, as shown in Figure 1-5. Also available in Format Cells are options for changing font colors, text alignment, fill colors, borders, and so on.

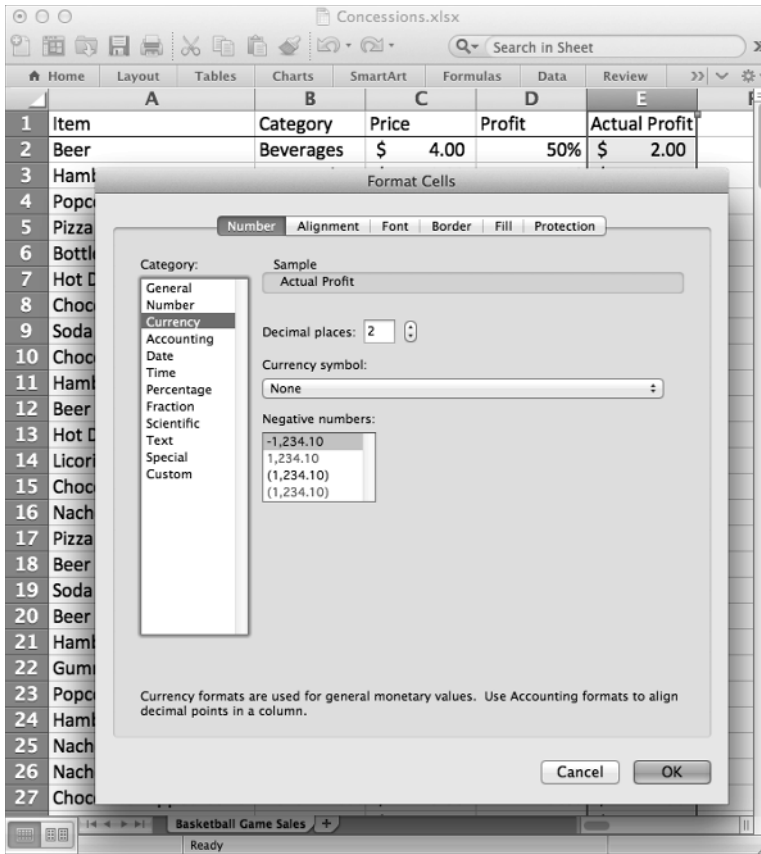


Figure 1-5: The Format Cells menu

But here's a conundrum. What if you want to format only the cells that have a certain value or range of values in them? And what if you want that formatting to change with the values?

That's called *conditional formatting*, and this book makes liberal use of it.

Cancel out of the Format Cells menu and navigate to the Home tab. In the Styles section (Mac calls it Format), you'll find the Conditional Formatting button (see Figure 1-6). Click the button to drop down a menu of options. The conditional formatting most used in this text is Color Scales. Pick a scale for column E and note how each cell in the column is colored based on its high or low value.



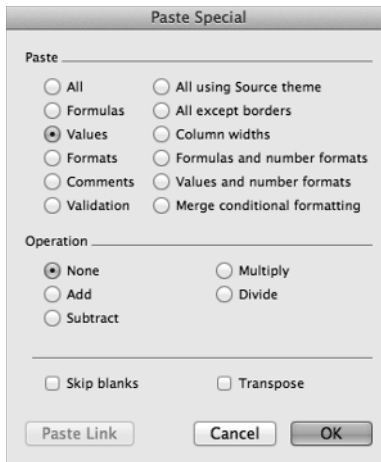
Figure 1-6: Applying conditional formatting to the profit

To remove conditional formatting, use the Clear Rules options under the Conditional Formatting menu.

## Paste Special Values

It's often in your best interest not to have a formula lying around like you see in Column E in Figure 1-4. If you were using the `RAND()` formula to generate a random value, for example, it changes each time the spreadsheet auto-recalculates, which while awesome, can also be extremely annoying. The solution is to copy and paste these cells back to the sheet as flat values.

To convert formulas to values only, simply copy a column filled with formulas (grab column E) and paste it back using the Paste Special option (found on the Home tab under the Paste option on Windows and under the Edit menu on Mac). In the Paste Special window, choose to paste as values (see Figure 1-7). Note also that Paste Special allows you to *transpose* the data from vertical to horizontal and vice versa when pasting. You'll be using that a fair bit in the chapters to come.



**Figure 1-7:** The Paste Special window in Excel 2011

## Inserting Charts

In the concession stand sales workbook, there's also a tab called Calories with a tiny table that shows the calorie count of each item the concession stand sells. You can chart data like this in Excel easily. On the Insert tab (Charts on a Mac), there is a charts section that provides different visualization options such as bar charts, line graphs, and pie charts.

### NOTE

In this book, we're going to use mostly column charts, line graphs, and scatter plots. Never be caught using a pie chart. And especially never use the 3D pie charts Excel offers, or my ghost will personally haunt you when I die. They're ugly, they don't communicate data well, and the 3D effect has less aesthetic value than the seashell paintings hanging on the wall of my dentist's office.

Highlighting columns A:B on the Calories workbook, you can select a Clustered Column chart to visualize the data. Play around with the graph. Sections can be right-clicked to bring up formatting menus. For example, right-clicking the bars, you can select "Format



Data Series...” under which you can change the fill color on the bars from the default Excel blue to any number of pleasing shades—black, for instance.

There’s no reason for the default legend, so you should select it and press delete to remove it. You might also want to select various text sections on the graph and increase the size of their font (font size is under the Home tab in Excel). This gives the graph shown in Figure 1-8.

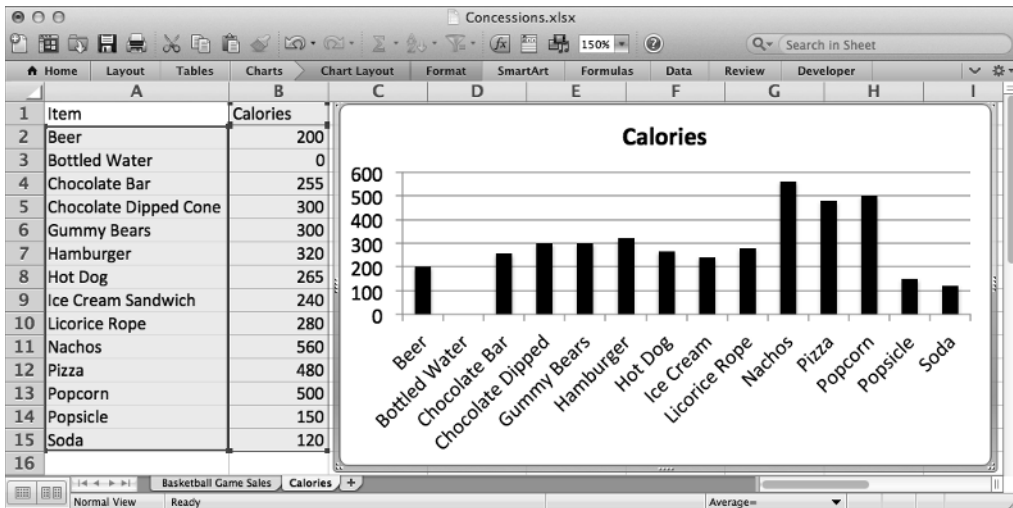
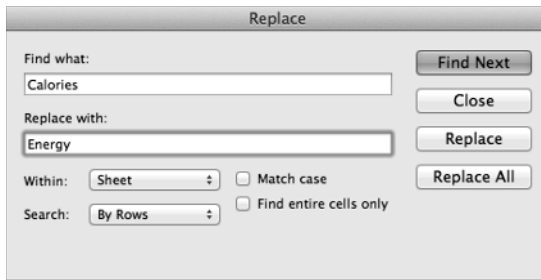


Figure 1-8: Inserting a calories column chart

## Locating the Find and Replace Menus

You’re going to use find and replace a fair bit in this book. On Windows you can either press Ctrl+F to open up the Find window (Ctrl+H for replace) or navigate to the Home tab and use the Find button in the Editing section. On Mac, there’s a search field on the top right of the sheet (press the down arrow for the Replace menu), or you can just press Cmd+F to bring up the Find and Replace menu.

Just to test it out, open up the replace menu on the Calories sheet. You can replace every instance of the word “Calories” with the word “Energy” (see Figure 1-9) by popping the words in the Find and Replace window and pressing Replace All.



**Figure 1-9:** Running a Find and Replace

## Formulas for Locating and Pulling Values

If I didn't assume you at least knew some formulas in Excel (SUM, MAX, MIN, PERCENTILE, and so on), we'd be here all day. And I want to get started. But there are some formulas used a lot in this book that you've probably not used unless you've dug deep into the wonderful world of spreadsheets. These formulas deal with *finding a value in a range and returning its location* or on the flip side *finding a location in a range and returning its value*.

I want to cover a few of those on the Calories tab.

Sometimes you want to know the place in line of some element in a column or row. Is it first, second, third? The MATCH formula handles that quite nicely. Below your calorie data, label A18 as **Match**. You can implement the formula one cell over in B18 to find where in the item list above the word "Hamburger" appears. To use the formula, you supply it a value to look for, a range to search in, and a 0 to force it to give you back the position of the keyword itself:

```
=MATCH("Hamburger", A2:A15, 0)
```

This yields a 6, because "Hamburger" is the sixth item in the list (see Figure 1-10).

Next up is the INDEX formula. Label A19 as **Index**.

This formula takes in a range of values and a row and column number and returns the value in the range at that location. For example, you can feed the INDEX formula our calorie table A1:B15, and to pull back the calorie count for bottled water, feed in 3 rows down and 2 columns over:

```
=INDEX(A1:B15, 3, 2)
```

This yields a calorie count of 0 as expected (see Figure 1-10).

Another formula you'll see a lot in this text is `OFFSET`. Go ahead and label A20 as `Offset`, and you can play with the formula in B20.

With this formula, you provide a range that acts like a cursor which is moved around with row and column offsets (similar to `INDEX` for the single valued case except it's 0-based). For example, you can provide `OFFSET` with a reference to the top left of the sheet, A1, and then pull back the value 3 cells below by providing a row offset of 3 and a column offset of 0:

```
=OFFSET(A1,3,0)
```

This returns the name of the third item on the list, "Chocolate Bar." See Figure 1-10.

The last formula I want to look at in this section is `SMALL` (it has a counterpart called `LARGE` that works the same way). If you have a list of values and you want to return, say, the third smallest, `SMALL` does that for you. To see this, label A21 as `Small` and in B21 feed in the list of calorie counts and an index of 3:

```
=SMALL(B2:B15,3)
```

This hands back a value of 150 which is the third smallest after 0 (bottled water) and 120 (soda). See Figure 1-10.

Now, there's one more formula used for looking up values that's kind of like `MATCH` on steroids and that's `VLOOKUP` (and its horizontal counterpart `HLOOKUP`). That's got its own section next because it's a beast.

	A	B	C
12	Pizza	480	
13	Popcorn	500	
14	Popsicle	150	
15	Soda	120	
16			
17			
18	Match	6	
19	Index	0	
20	Offset	Chocolate Bar	
21	Small	150	

Figure 1-10: Formulas you should learn

## Using VLOOKUP to Merge Data

Go ahead and flip back to the Basketball Game Sales tab. You can still reference a cell here from the previous tab, *Calories*, by simply placing the tab name and “!” in front of a referenced cell. For example, `Calories!B2` is a reference to the calories in beer regardless of what sheet you’re working in.

Now, what if you wanted to toss the calorie data into a column back on the sales sheet so that next to each item sold the appropriate calorie count was listed? You’d somehow have to look up the calorie count of each item sold and place it into a column next to the transaction. Well, it turns out there’s a formula for that called `VLOOKUP`.

Go ahead and label Column F in the spreadsheet *Calories* for this purpose. Cell F2 will include the calorie count for the first beer transaction from the *Calories* table. Using the `VLOOKUP` formula, you supply the item name from cell A2, a reference to the table `Calories!$A$1:$B$15`, and the relative column offset you want your return value to be read out of, which is to say the second column:

```
=VLOOKUP(A2,Calories!$A$1:$B$15,2,FALSE)
```

The `FALSE` at the end of the `VLOOKUP` formula means that you will not accept approximate matches for “Beer.” If the formula can’t find “Beer” on the calories table, it returns an error.

When you enter the formula, you can see that 200 calories is read in from the table on the *Calories* tab. Since you’ve put the \$ in front of the table references in the formula, you can copy this formula down the column by double-clicking the bottom-right corner of the cell. *Voila!* As shown in Figure 1-11, you have calorie counts for every transaction.

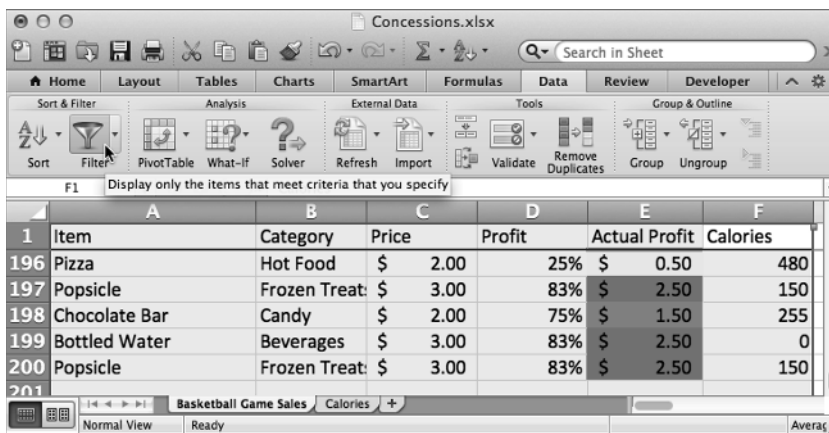
	A	B	C	D	E	F
1	Item	Category	Price	Profit	Actual Profit	Calories
2	Beer	Beverages	\$ 4.00	50%	\$ 2.00	200
3	Hamburger	Hot Food	\$ 3.00	67%	\$ 2.00	320
4	Popcorn	Hot Food	\$ 5.00	80%	\$ 4.00	500
5	Pizza	Hot Food	\$ 2.00	25%	\$ 0.50	480
6	Bottled Water	Beverages	\$ 3.00	83%	\$ 2.50	0

**Figure 1-11:** Using `VLOOKUP` to grab calorie counts

## Filtering and Sorting

Now that you have calories in there, say you now want to view only those transactions from the Frozen Treats category. What you want to do then is filter the sheet. To do so, first you select the data in range A1:F200. You can put the cursor in A1 and press Shift+Ctrl+↓ then →. An even easier method is to click the top of column A and hold the click as you mouse over to column F to highlight all six columns.

Then to place auto-filtering on these six columns, you press the Filter button in the Data section of the ribbon. It looks like a gray funnel as shown in Figure 1-12.



**Figure 1-12:** Place auto-filter on a selected range

Once auto-filter is activated, you can click the drop-down menu that appears in cell B1 and choose to show only certain categories (in this case, only the Frozen Treats transactions will be displayed). See Figure 1-13.

Once you've filtered, highlighting columns of data allows the summary bar in Excel to give you rolled-up information just on the cells that remain. For example, having filtered just the Frozen Treats, we can highlight the values in column E and use the summary bar to get a quick total of profit just from that category. See Figure 1-14.

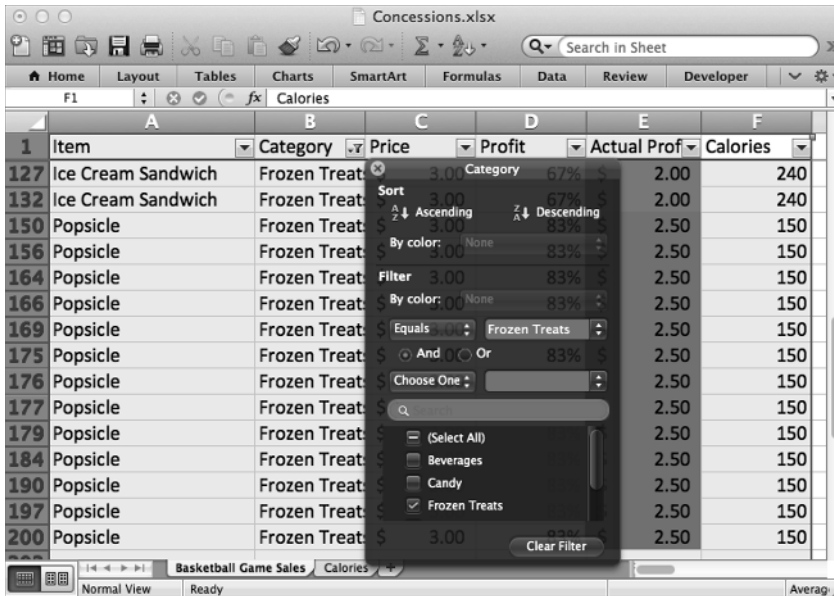


Figure 1-13: Filtering on category

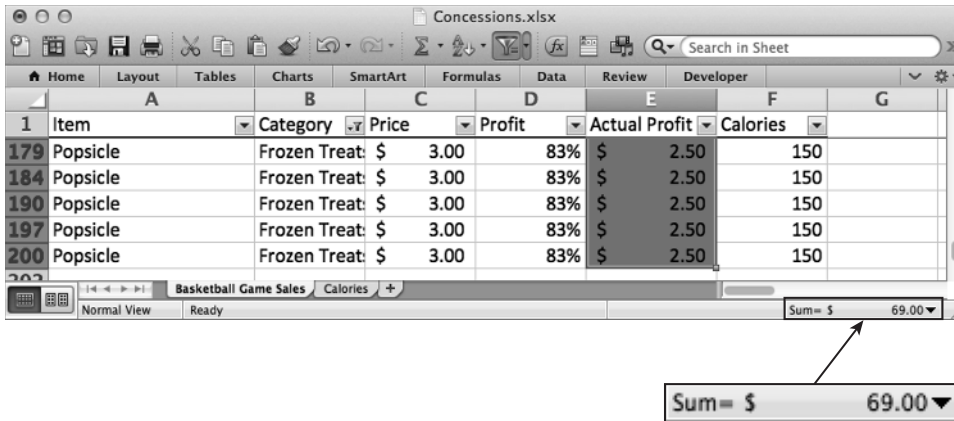


Figure 1-14: Summarizing a filtered column

Auto-filter allows you to sort as well. For example, if you want to sort by profit, just click the auto-filter menu on the Profit cell (D1) and select Sort Ascending (or “Smallest to Largest” in some versions). See Figure 1-15.

	A	B	C	D	E	F
1	Item	Category	Price	Profit	Actual Profit	Calories
8	Chocolate Dipped Cone	Frozen Treat	\$ 3.00	50%	1.50	300
15	Chocolate Dipped Cone	Frozen Treat	\$ 3.00	50%	3.00	300
27	Chocolate Dipped Cone	Frozen Treat	\$ 3.00	50%	3.00	300
31	Chocolate Dipped Cone	Frozen Treat	\$ 3.00	50%	3.00	300
34	Chocolate Dipped Cone	Frozen Treat	\$ 3.00	50%	3.00	300
41	Chocolate Dipped Cone	Frozen Treat	\$ 3.00	50%	3.00	300
48	Chocolate Dipped Cone	Frozen Treat	\$ 3.00	50%	3.00	300
56	Chocolate Dipped Cone	Frozen Treat	\$ 3.00	50%	3.00	300
67	Chocolate Dipped Cone	Frozen Treat	\$ 3.00	50%	3.00	300
71	Chocolate Dipped Cone	Frozen Treat	\$ 3.00	50%	3.00	300
81	Chocolate Dipped Cone	Frozen Treat	\$ 3.00	50%	3.00	300
84	Ice Cream Sandwich	Frozen Treat	\$ 3.00	67%	2.00	240
89	Ice Cream Sandwich	Frozen Treat	\$ 3.00	67%	2.00	240
91	Ice Cream Sandwich	Frozen Treat	\$ 3.00	67%	2.00	240
98	Ice Cream Sandwich	Frozen Treat	\$ 3.00	67%	2.00	240
103	Ice Cream Sandwich	Frozen Treat	\$ 3.00	67%	2.00	240
111	Ice Cream Sandwich	Frozen Treat	\$ 3.00	67%	2.00	240

**Figure 1-15:** Sorting in ascending order by profit

To remove all the filtering you’ve applied, either you can go back into the Category filter menu and check the other boxes, or you can un-toggle the filter button on the ribbon that you pressed in the first place. You’ll see that although you have all of your data back, the Frozen Treats are still in the order you sorted them in.

Excel also offers the Sort interface for doing more complex sorts than might be possible with auto-filter. To use the feature, you highlight the data to be sorted (grab A:F again) and select Sort from the Sort & Filter section of the Data tab in Excel. This will bring up the sort menu. On Mac, to get this window, you must press the down arrow in the sort button and select Custom Sort...

In the sort menu, shown in Figure 1-16, you can note whether your data has column headers or not, and if it does have headers like this example does, then you can select, by name, the columns to be sorted.

Now, the most awesome part of this sorting interface is that under the “Options...” button, you can select to sort left to right instead of column data. That’s something you cannot do with auto-filter. In top to bottom of this book you’ll need to randomly sort data by both columns and rows in two quick steps, and this interface is going to be your friend. For now, just cancel out of it as the data is already ordered the way you want it.

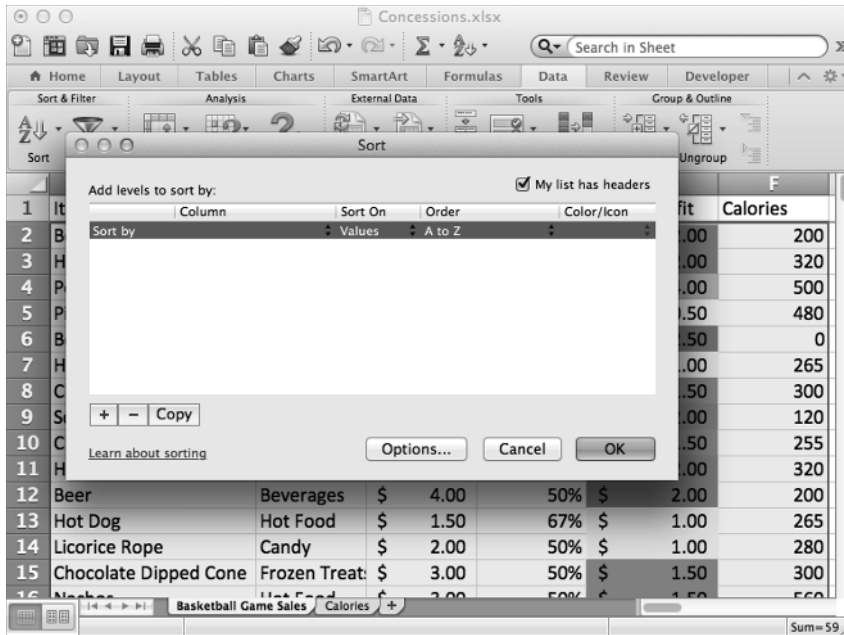


Figure 1-16: Using the Sort menu

## Using PivotTables

What if you wanted to know the total counts of each item type you sold? Or you wanted to know revenue totals by item?

These questions are akin to “aggregate” or “group by” queries that you’d run in a traditional SQL database. But this data isn’t in a database. It’s in a spreadsheet. That’s where PivotTables come to the rescue.

Just as when you filtered your data, you start by selecting the data you want to manipulate—in this case, the purchase data in the range A1:F:200. From the Insert tab (Data tab on Mac), you can press the PivotTable button and select for Excel to create a new sheet with a PivotTable. While some versions of Excel allow you to insert a PivotTable into an existing sheet, it’s standard practice to select the new sheet option unless you have a really good reason not to.

In this new sheet, the PivotTable Builder will be aligned to the right of the table (it floats on a Mac). The builder allows you to take the columns from the original selected data and use them as report filters, column and row labels for grouping, or values. A report filter is similar in function to a filter from the previous section—it allows you to select only a subset of the data, such as Frozen Treats. The Column Labels and Row Labels fill in the meat of the PivotTable report with distinct values from the selected columns.



On Windows, the initial PivotTable built will be completely empty, while on Mac it is often prepopulated with distinct values from the first selected column down the rows of the table and distinct values from the second column across the columns. If you're on a Mac, go ahead and uncheck all the boxes in the builder, so that you can work along from an empty table.

Now, say you wanted to know total revenue by item. To get at that, you'd drag the Item tile in the PivotTable Builder into the Rows section and the Price tile into the Values section. This means that you'll be operating on revenue grouped by item name.

Initially, however, the PivotTable is set up to merely count the number of price records that are within a group. For example, there are 20 Beer rows. See Figure 1-17.

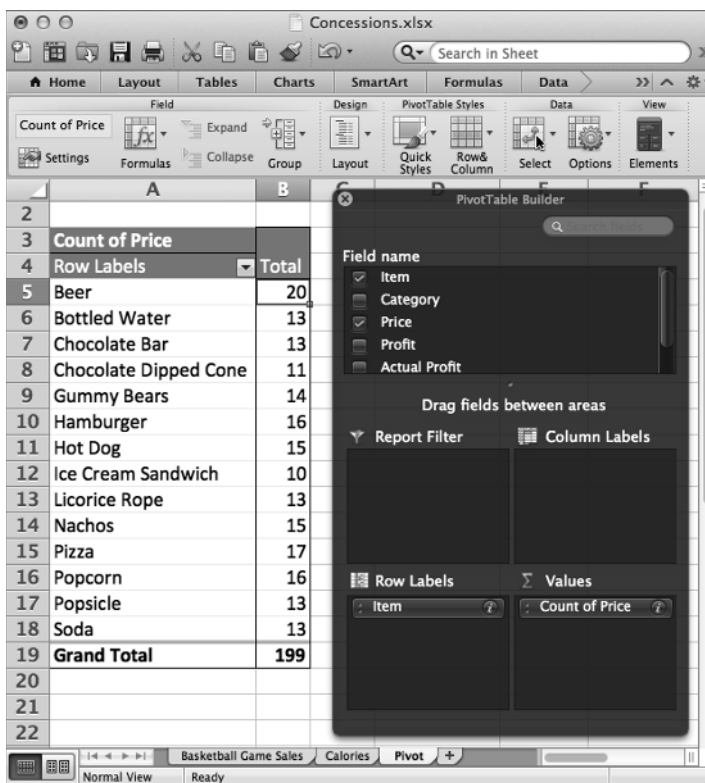
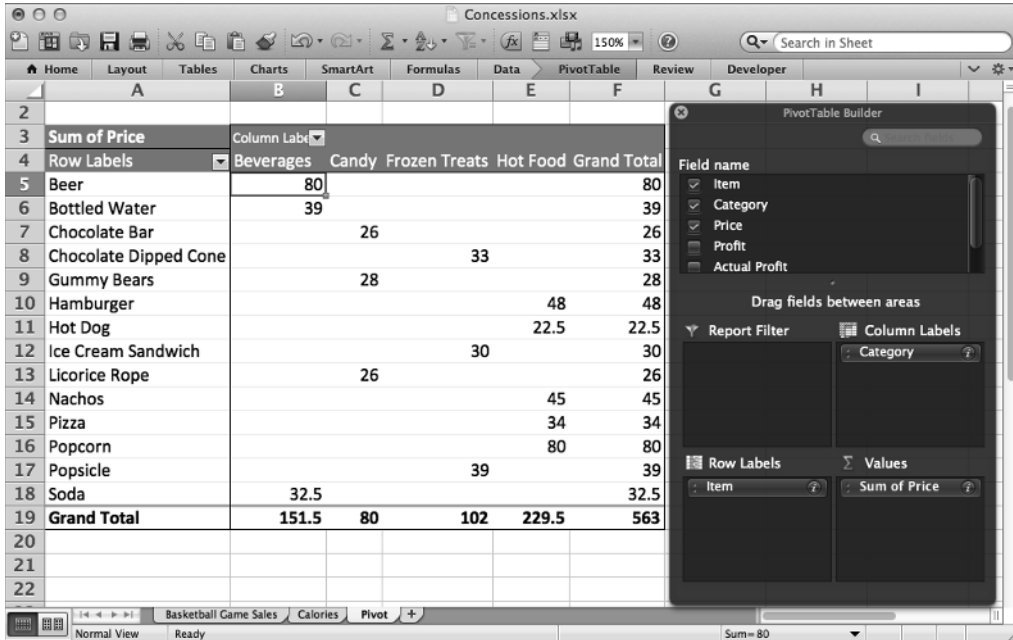


Figure 1-17: The PivotTable builder and a count of sales by item

You need to change the count to a sum in order to examine revenue. To do so, on Windows, drop the menu down on the Price tile in the Values section of the builder and select “Value Field Settings...” On Mac, press the little “i” button. From there, “sum” can be selected from the various summary options.

What if you wanted to break out these sums by category? To do so, you drag the Category tile into the Columns section of the builder. This gives the table shown in Figure 1-18. Note that the PivotTable in the figure automatically totals up rows and columns for you.

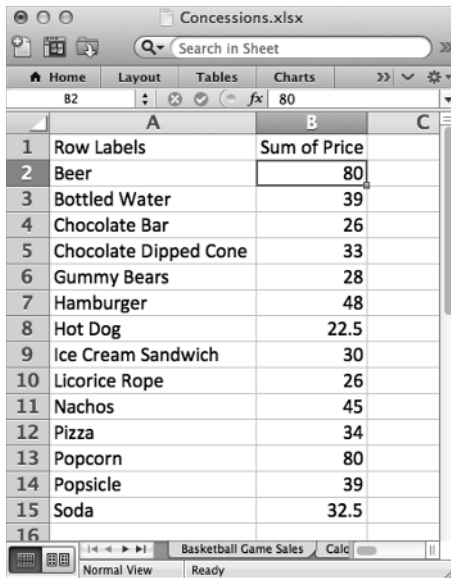


**Figure 1-18:** Revenue by item and category

And if you want to ever get rid of something from the table, just uncheck it or grab the tile from the section it's in and drag it out of the sheet as if you were tossing it away. Go ahead and drop the Category tile.

Once you get a report you want in a PivotTable, you can always select the values and paste them to another sheet to work on further. In this example, you can copy the table (A5:B18 on Mac) and Paste Special its values into a new tab called Revenue By Item (see Figure 1-19).

Feel free to swap in various row and column labels until you get the hang of what's going on. For instance, try to get a total calorie count sold by category using a PivotTable.



The screenshot shows an Excel spreadsheet titled 'Concessions.xlsx'. The active sheet is 'Basketball Game Sales'. The table displays the following data:

	A	B	C
1	Row Labels	Sum of Price	
2	Beer	80	
3	Bottled Water	39	
4	Chocolate Bar	26	
5	Chocolate Dipped Cone	33	
6	Gummy Bears	28	
7	Hamburger	48	
8	Hot Dog	22.5	
9	Ice Cream Sandwich	30	
10	Licorice Rope	26	
11	Nachos	45	
12	Pizza	34	
13	Popcorn	80	
14	Popsicle	39	
15	Soda	32.5	
16			

Figure 1-19: Revenue by Item tab created by pasting values from a PivotTable

## Using Array Formulas

In the concession transaction workbook, there is a tab called Fee Schedule. As it turns out, Coach O'Shaughnessy would let you run the snack stand only if you kicked some of the profit back to him (perhaps to subsidize his tube sock-buying habit). The Fee Schedule tab shows the percent cut he takes on each item sold.

So how much money do you owe him for last night's game? To answer that question, you need to multiply the total revenue of each item from the PivotTable by the cut for the coach and sum them all up.

There's a great formula for this operation that will do all the multiplication and summation in a single step. Rather creatively named, it's called `SUMPRODUCT`. In cell E1 on the Revenue By Item sheet, add a label called **Total Cut for Coach**. In C2, determine the `SUMPRODUCT` of the revenue and the fees by adding this formula:

```
=SUMPRODUCT(B2:B15, 'Fee Schedule'!B2:O2)
```

Uh oh. There's an error; the cell just reads #value. What's going wrong?

Even though you've selected two ranges of equal size and put them in `SUMPRODUCT`, the formula can't see that the ranges are equal because one range is vertical and one's horizontal.

Fortunately, Excel has a function for flipping arrays in the right direction. It's called `TRANSPOSE`. You need to write the formula like this:

```
=SUMPRODUCT(B2:B15,TRANSPOSE('Fee Schedule'!B2:O2))
```

Nope! Still getting an error.

The reason you're still getting an error is that every formula in Excel, by default, returns a single value. Even `TRANSPOSE` returns the first value in the transposed array. If you want the *whole array* returned, you have to turn `TRANSPOSE` into an "array formula," which means exactly what you might think. Array formulas hand you back arrays, not single values.

You don't have to change the way you type your `SUMPRODUCT` to make this happen. All you need to do is when you're done typing the formula, instead of pressing Enter, press Ctrl+Shift+Enter. On the Mac, you use Command+Return.

Victory! As shown in Figure 1-20, the calculation now reads \$57.60. But I suggest rounding that down to \$50, because how many socks does Coach really need?

	A	B	C	D	E
1	Row Labels	Sum of Price		Total Cut for Coach:	
2	Beer	80		57.6	
3	Bottled Water	39			
4	Chocolate Bar	26			

**Figure 1-20:** Taking a `SUMPRODUCT` with an array formula

## Solving Stuff with Solver

Many of the techniques you'll study in this book can be boiled down to *optimization models*. An optimization problem is one where you have to make the best decision (choose the best investments, minimize your company's costs, find the class schedule with the

fewest morning classes, or so on). In optimization models then, the words “minimize” and “maximize” come up a lot when articulating an objective.

In data science, many of the practices, whether that’s artificial intelligence, data mining, or forecasting, are actually just some data prep plus a model-fitting step that’s actually an optimization model. So it’d make sense to teach optimization first. But learning all there is to know about optimization is tough to do straight off the bat. So you’ll do an in-depth optimization study in Chapter 4 *after* you do some more fun machine learning problems in Chapters 2 and 3. To fill in the gaps though, it’s best if you get a little practice with optimization now. Just a taste.

In Excel, optimization problems are solved using an Add-In that ships with Excel called Solver.

- On Windows, Solver may be added in by going to File (in Excel 2007 it’s the top left Windows button) ⇨ Options ⇨ Add-ins, and under the Manage drop-down choosing Excel Add-ins and pressing the Go button. Check the Solver Add-In box and press OK.
- On Mac, Solver is added by going to Tools then Add-ins and selecting Solver.xlam from the menu.

A Solver button will appear in the Analysis section of the Data tab in every version.

All right! Now that Solver is installed, here’s an optimization problem: You are told you need 2,400 calories a day. What’s the fewest number of items you can buy from the snack stand to achieve that? Obviously, you could buy 10 ice cream sandwiches at 240 calories a piece, but is there a way to do it for fewer items than that?

Solver can tell you!

To start, make a copy of the Calories sheet, name the sheet **Calories-Solver**, and clear out everything but the calories table on the copy. If you don’t know how to make a copy of a sheet in Excel, you simply right-click the tab you’d like to copy and select the Move or Copy menu. This gives you the new sheet shown in Figure 1-21.

To get Solver to work, you need to provide it with a range of cells it can set with decisions. In this case, Solver needs to decide how many of each item to buy. So in Column C next to the calorie counts, label the column **How many?** (or whatever you feel like), and you can allow Solver to store its decisions in this column.

Excel considers blank cells to be 0s so you needn’t fill in these cells with anything to start. Solver will do that for you.

	A	B
1	Item	Calories
2	Beer	200
3	Bottled Water	0
4	Chocolate Bar	255
5	Chocolate Dipped Cone	300
6	Gummy Bears	300
7	Hamburger	320
8	Hot Dog	265
9	Ice Cream Sandwich	240
10	Licorice Rope	280
11	Nachos	560
12	Pizza	480
13	Popcorn	500
14	Popsicle	150
15	Soda	120

**Figure 1-21:** The copied Calories-Solver sheet

In cell C16, sum up the number of items to be bought above as:

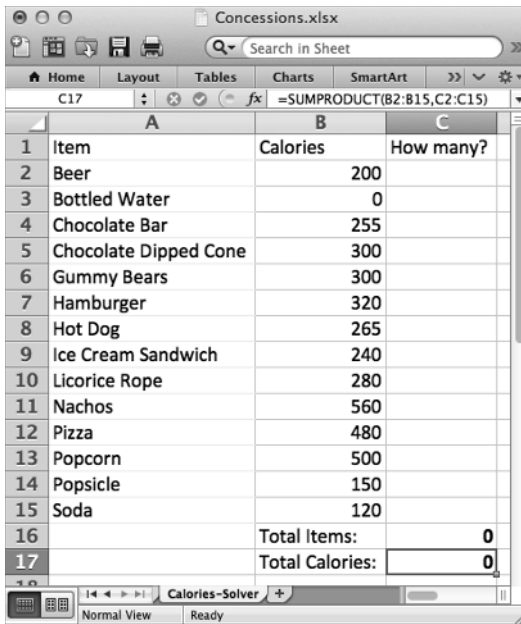
```
=SUM(C2:C15)
```

And below that you can sum up the total calorie count of these items (which you'll want eventually to equal 2,400) using the `SUMPRODUCT` formula:

```
=SUMPRODUCT(B2:B15,C2:C15)
```

This gives the initial sheet shown in Figure 1-22.

Now you're ready to build the model, so bring up the Solver window by pressing the Solver button on the Data tab.



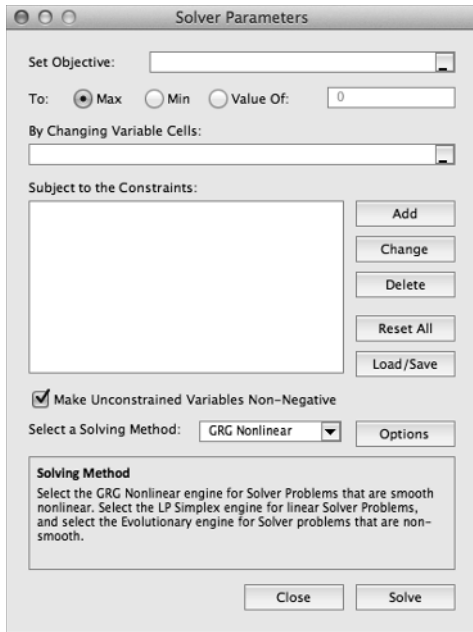
	A	B	C
1	Item	Calories	How many?
2	Beer	200	
3	Bottled Water	0	
4	Chocolate Bar	255	
5	Chocolate Dipped Cone	300	
6	Gummy Bears	300	
7	Hamburger	320	
8	Hot Dog	265	
9	Ice Cream Sandwich	240	
10	Licorice Rope	280	
11	Nachos	560	
12	Pizza	480	
13	Popcorn	500	
14	Popsicle	150	
15	Soda	120	
16		Total Items:	0
17		Total Calories:	0

**Figure 1-22:** Getting calorie and item counts set up

## NOTE

The Solver window, shown in Figure 1-23 in Excel 2011, looks pretty similar in Excel 2010, 2011, and 2013. In Excel 2007, the layout is slightly different, but the only substantive difference is that there is no algorithm selection box. Rather, there's an "Assume Linear Model" checkbox under the Options menu. We'll learn all about these elements later.

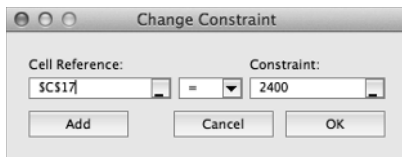
The main elements you plug into Solver to solve a problem, as shown in Figure 1-23, are an objective cell, an optimization direction (minimization or maximization), some decision variables that can be changed by Solver, and some constraints.



**Figure 1-23:** The uninitialized Solver window

In your case, the objective is to minimize the total items in cell C16. The cells that can be altered are the item selections in C2:C15. And the constraints are that C17, the total calories, needs to be equal to 2,400. Also, we'll need to add a constraint that our decisions be counting numbers, so we'll need to check the non-negative box (under the options menu in Excel 2007) and add an integer constraint to the decisions. After all, you can't buy 1.7 sodas. These integer constraints will be covered in depth in Chapter 4.

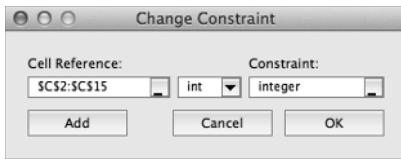
To add in the total calorie constraint, press the Add button and set C17 equal to 2,400 as shown in Figure 1-24.



**Figure 1-24:** Adding the calorie constraint

Similarly, add a constraint setting C2:C15 to be integers as shown in Figure 1-25.





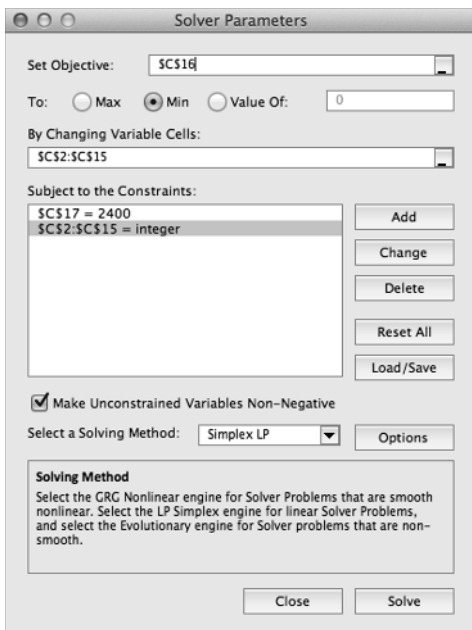
**Figure 1-25:** Adding an integer constraint

Press OK.

In Excel 2010, 2011, and 2013, make sure the solving method is set to Simplex LP. Simplex LP is appropriate for this problem, because this problem is *linear* (the “L” in LP stands for linear as you’ll see in Chapter 4). By linear, I mean that the problem involves nothing but linear combinations of the decisions in C2 through C15 (sums, products with constants such as calorie counts, etc.).

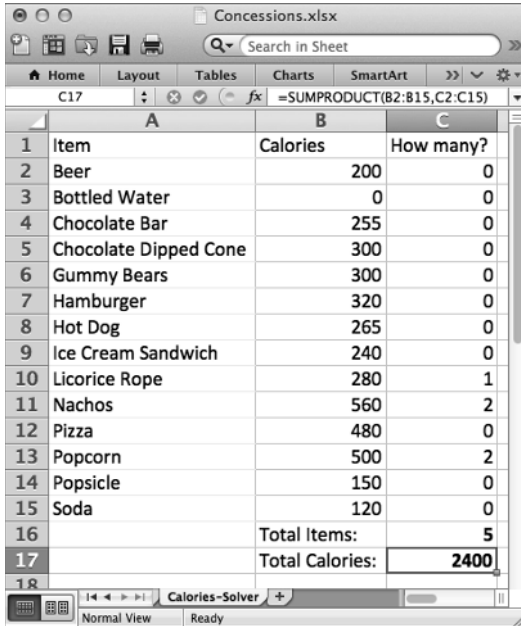
If we had non-linear calculations in the model (perhaps a square root of a decision, a logarithm, or an exponential function), then we could use one of the other algorithms Excel provides in Solver. Chapter 4 covers this in great detail.

In Excel 2007, you would denote the problem as linear by clicking the Assume Linear Model under the Options screen. Your final setup should appear as in Figure 1-26.



**Figure 1-26:** Final Solver setup for minimizing items needed for 2,400 calories

All right! Go ahead and press the Solve button. Excel should find a solution almost immediately. And that solution, as shown in Figure 1-27, is 5. Now, your Excel might pick a different 5 items than mine in the screenshot, but the minimum is 5 nonetheless.



	A	B	C
1	Item	Calories	How many?
2	Beer	200	0
3	Bottled Water	0	0
4	Chocolate Bar	255	0
5	Chocolate Dipped Cone	300	0
6	Gummy Bears	300	0
7	Hamburger	320	0
8	Hot Dog	265	0
9	Ice Cream Sandwich	240	0
10	Licorice Rope	280	1
11	Nachos	560	2
12	Pizza	480	0
13	Popcorn	500	2
14	Popsicle	150	0
15	Soda	120	0
16		Total Items:	5
17		Total Calories:	2400

**Figure 1-27:** The optimized item selection

## OpenSolver: I Wish We Didn't Need This, but We Do

This book was originally designed to work completely with Excel's built-in Solver. However, as it turns out, functionality was *removed* from Solver in later versions for mysterious and unadvertised reasons.

What that means is that while this whole book works using vanilla Solver in Excel 2007 and Excel 2011 for Mac, in Excel 2010 and Excel 2013, the built-in Solver will occasionally complain that a linear optimization model is too large (I'll give you a heads-up in this book whenever a model gets that complex).

Luckily, there's an excellent free tool called OpenSolver that's available for the Windows versions of Excel that addresses this deficiency. With OpenSolver, you can still build your model in the regular Solver interface, but OpenSolver provides a button that you press to use its Simplex LP algorithm implementation, which is blazingly fast.

To set up OpenSolver, navigate to <http://OpenSolver.org> and download the zip file. Uncompress the file into a folder, and whenever you want to solve a beefy model, just set it up in a spreadsheet like normal and double-click the OpenSolver.xlam file, which will give you an OpenSolver section on the Data tab in Excel. Press the Solve button to solve an existing model. As shown in Figure 1-28, I've applied OpenSolver in Excel 2013 to the model from the previous section, and it buys five slices of pizza.

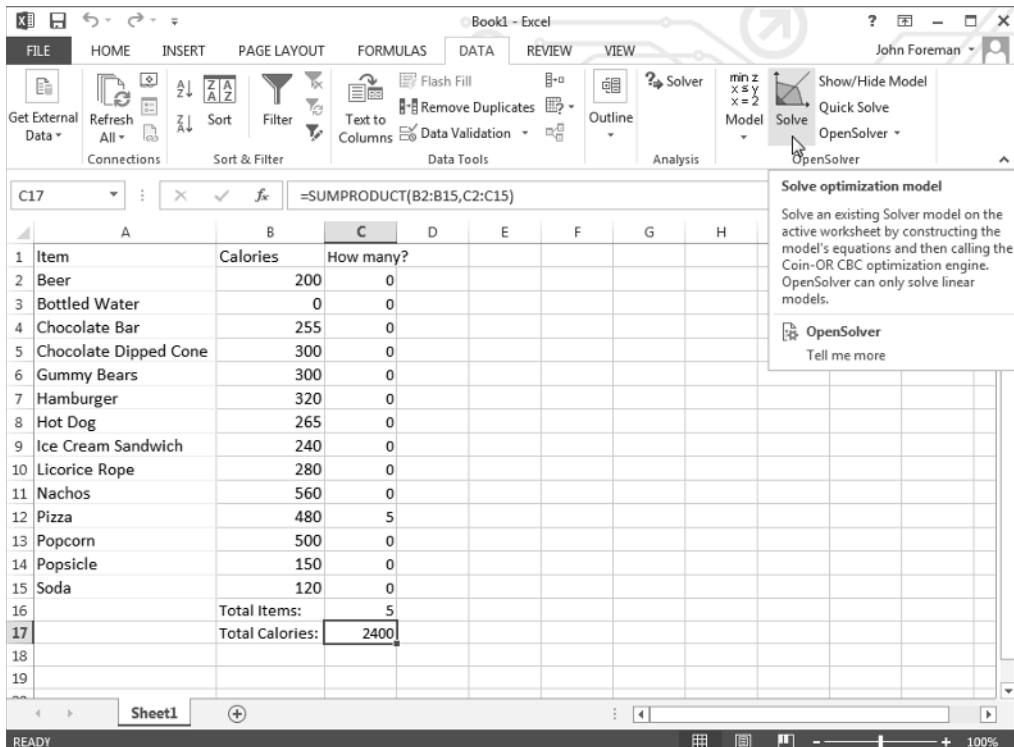


Figure 1-28: OpenSolver buys pizza like a madman

## Wrapping Up

All right, you've learned how to navigate and select ranges quickly, how to leverage absolute references, how to paste special values, how to use `VLOOKUP` and other matching formulas, how to sort and filter data, how to create PivotTables and charts, how to execute array formulas, and how and when to bust out Solver.

Here's either a depressing or fun fact depending on your perspective. I've known management consultants at prominent firms who earn excellent salaries by doing what I call the "consulting two-step":

1. Talk about nonsense with clients (sports, vacation, barbeque ... not that there's anything nonsensical about smoked meats).
2. Summarize data in Excel.

You may not know all there is to know about college football (I certainly don't), but if you internalize this chapter, you'll have point number two knocked out.

But you're not here to become a management consultant. You're here to drive deep into data science, and that starts in the next chapter where we'll get started with a little bit of unsupervised machine learning.