



Using Data Science to Transform
Information into Insight

John W. Foreman

DATA SMART

"Data Smart makes modern statistic methods and algorithms understandable and easy to implement. Slogging through textbooks and academic papers is no longer required!" —**Patrick Crosby**, *Founder of StatHat & first CTO at OkCupid*

WILEY

Introduction

What Am I Doing Here?

You've probably heard the term *data science* floating around recently in the media, in business books and journals, and at conferences. Data science can call presidential races, reveal more about your buying habits than you'd dare tell your mother, and predict just how many years those chili cheese burritos have been shaving off your life.

Data scientists, the elite practitioners of this art, were even labeled “sexy” in a recent Harvard Business Review article, although there's apparently such a shortage that it's kind of like calling a unicorn sexy. There's just no way to verify the claim, but if you could see me as I type this book with my neck beard and the tired eyes of a parent of three boys, you'd know that sexy is a bit of an overstatement.

I digress. The point is that there's a buzz about data science these days, and that buzz is creating pressure on a lot of businesses. If you're not doing data science, you're gonna lose out to the competition. Someone's going to come along with some new product called the “BlahBlahBlahBigDataGraphThing” and destroy your business.

Take a deep breath.

The truth is most people are going about data science all wrong. They're starting with buying the tools and hiring the consultants. They're spending all their money before they even know what they want, because a purchase order seems to pass for actual progress in many companies these days.

By reading this book, you're gonna have a leg up on those jokers, because you're going to learn exactly what these techniques in data science are and how they're used. When it comes time to do the planning, and the hiring, and the buying, you'll already know how to identify the data science opportunities within your own organization.

The purpose of this book is to introduce you to the practice of data science in a comfortable and conversational way. When you're done, I hope that much of that data science anxiety you're feeling is replaced with excitement and with ideas about how you can use data to take your business to the next level.

A Workable Definition of Data Science

To an extent, *data science* is synonymous with or related to terms like *business analytics*, *operations research*, *business intelligence*, *competitive intelligence*, *data analysis and modeling*, and *knowledge extraction* (also called *knowledge discovery in databases* or *KDD*). It's just a new spin on something that people have been doing for a long time.

There's been a shift in technology since the heyday of those other terms. Advancements in hardware and software have made it easy and inexpensive to collect, store, and analyze large amounts of data whether that be sales and marketing data, HTTP requests from your website, customer support data, and so on. Small businesses and nonprofits can now engage in the kind of analytics that were previously the purview of large enterprises.

Of course, while data science is used as a catch-all buzzword for analytics today, data science is most often associated with data mining techniques such as artificial intelligence, clustering, and outlier detection. Thanks to the cheap technology-enabled proliferation of transactional

business data, these computational techniques have gained a foothold in business in recent years where previously they were too cumbersome to use in production settings.

In this book, I'm going to take a broad view of data science. Here's the definition I'll work from:

Data science is the transformation of data using mathematics and statistics into valuable insights, decisions, and products.

This is a *business-centric* definition. It's about a usable and valuable end product derived from data. Why? Because I'm not in this for research purposes or because I think data has aesthetic merit. I do data science to help my organization function better and create value; if you're reading this, I suspect you're after something similar.

With that definition in mind, this book will cover mainstay analytics techniques such as optimization, forecasting, and simulation, as well as more “hot” topics such as artificial intelligence, network graphs, clustering, and outlier detection.

Some of these techniques are as old as World War II. Others were introduced in the last 5 years. And you'll see that age has no bearing on difficulty or usefulness. All these techniques—whether or not they're currently the rage—are equally useful in the right business context.

And that's why you need to understand how they work, how to choose the right technique for the right problem, and how to prototype with them. There are a lot of folks out there who understand one or two of these techniques, but the rest aren't on their radar. If all I had in my toolbox was a hammer, I'd probably try to solve every problem by smacking it real hard. Not unlike my two-year-old.

Better to have a few other tools at your disposal.

But Wait, What about Big Data?

You've heard the term *big data* even more than *data science* most likely. Is this a book on big data?

That depends on how you define big data. If you define big data as computing simple summary statistics on unstructured garbage stored in massive, horizontally scalable, NoSQL databases, then no, this is not a book on big data.

If you define big data as turning transactional business data into decisions and insight using cutting-edge analytics (regardless of where that data is stored), then yes, this is a book about big data.

This is not a book that will be covering database technologies, like MongoDB and HBase. This is not a book that will be covering data science coding packages like Mahout, NumPy, various R libraries, and so on. There are other books out there for that stuff.

But that's a good thing. This book ignores the tools, the storage, and the code. Instead, it focuses as much as possible on the techniques. There are many folks out there who think that data storage and retrieval, with a little bit of cleanup and aggregation mixed in, constitutes all there is to know about big data.

They're wrong. This book will take you beyond the spiel you've been hearing from the big data software sales reps and bloggers to show you what's really possible with your data. And the cool thing is that for many of these techniques, your dataset can be any size, small or large. You don't have to have a petabyte of data and the expenses that come along with it in order to predict the interests of your customer base. If you have a massive dataset, that's great, but there are some businesses that don't have it,

need it, and will likely never generate it. Like my local butcher. But that doesn't mean his e-mail marketing couldn't benefit from a little bacon versus sausage cluster detection.

If data science books were workouts, this book would be all calisthenics—no machine weights, no ergs. Once you understand how to implement the techniques with even the most barebones of tools, you'll find yourself free to implement them in a variety of technologies, prototype with them with ease, buy the correct data science products from consultants, delegate the correct approach to your developers, and so on.

Who Am I?

Let me pause a moment to tell you my story. It'll go a long way to explaining why I teach data science the way I do. Many moons ago, I was a management consultant. I worked on analytics problems for organizations such as the FBI, DoD, the Coca-Cola Company, Intercontinental Hotels Group, and Royal Caribbean International. And through all these experiences I walked away having learned one thing—more people than just the scientists need to understand data science.

I worked with managers who bought simulations when they needed an optimization model. I worked with analysts who only understood Gantt charts, so everything needed to be solved with Gantt charts. As a consultant, it wasn't hard to win over a customer with any old white paper and a slick PowerPoint deck, because they couldn't tell AI from BI or BI from BS.

The point of this book is to broaden the audience of who understands and can implement data science techniques. I'm not trying to turn you into a data scientist against your

will. I just want you to be able to integrate data science as best as you can into the role you're already good at.

And that brings me to who you are.

Who Are You?

No, I haven't been using data science to spy on you. I have no idea who you are, but thanks for shelling out some money for this book. Or supporting your local library. You can do that, too.

Here are some archetypes (or *personas* for you marketing folks) I had in mind when writing this book. Maybe you are:

- The vice president of marketing who wants to use her transactional business data more strategically to price products and segment customers. But she doesn't understand the approaches her software developers and overpriced consultants are recommending she try.
- The demand forecasting analyst who knows his organization's historical purchase data holds more insight about his customers than just the next quarter's projections. But he doesn't know how to extract that insight.
- The CEO of an online retail start-up who wants to predict when a customer is likely to be interested in buying an item based on their past purchases.
- The business intelligence analyst who sees money going down the tubes from the infrastructure and supply chain costs her organization is accruing, but doesn't know how to systematically make cost-saving decisions.
- The online marketer who wants to do more with his company's free text customer interactions taking place

in e-mail, Facebook, and Twitter, but right now they're just being read and saved.

I have in mind that you are a reader who would benefit directly from knowing more about data science but hasn't found a way to get a foothold into all the techniques. The purpose of this book is to strip away all the distractions around data science (the code, the tools, and the hype) and teach the techniques using practical use cases that someone with a semester of linear algebra or calculus in college can understand. Assuming you didn't fail that semester. If you did, just read slower and use Wikipedia liberally.

No Regrets. Spreadsheets Forever

This is not a book about coding. In fact, I'm giving you my “no code” guarantee (until Chapter 10 at least). Why?

Because I don't want to spend a hundred pages at the beginning of this book messing with Git, setting environment variables, and doing the dance of Emacs versus Vi.

If you run Windows and Microsoft Office almost exclusively. If you work for the government, and they don't let you download and install random open source stuff on your box. Even if MATLAB or your TI-83 scared the hell out of you in college, you need not be afraid.

Do you need to know how to write code to put most of these techniques in automated, production settings? Absolutely! Or at least someone you work with needs to be able to handle code and storage technologies.

Do you need to know how to write code in order to understand, distinguish between, and prototype with these techniques? Absolutely not!

This is why I go over every technique in spreadsheet software.

Now, this is all a bit of a lie. The final chapter in this book is actually on moving to the data science-focused programming language, R. It's for those of you that want to use this book as a jumping-off point to deeper things.

But Spreadsheets Are So Démodé!

Spreadsheets are not the sexiest tools around. In fact, they're the Wilford-Brimley-selling-Colonial-Penn of the analytics tool world. Completely unsexy. Sorry, Wilford.

But that's the point. Spreadsheets stay out of the way. They allow you to see the data and to touch (or at least click on) the data. There's a freedom there. In order to learn these techniques, you need something vanilla, something everyone understands, but nonetheless, something that will let you move fast and light as you learn. That's a spreadsheet.

Say it with me: “I am a human. I have dignity. I should not have to write a map-reduce job in order to learn data science.”

And spreadsheets are great for prototyping! You're not running a production AI model for your online retail business out of Excel, but that doesn't mean you can't look at purchase data, experiment with features that predict product interest, and prototype a targeting model. In fact, it's the perfect place to do just that.

Use Excel or LibreOffice

All the examples you're going to work through will be visualized in the book in Excel.

On the book's website (www.wiley.com/go/datasmart) are posted companion spreadsheets for each chapter so that you can

follow along. If you're really adventurous, you can clear out all but the starting data in the spreadsheet and replicate all the work yourself.

This book is compatible with Excel versions 2007, 2010, 2011 for Mac, and 2013. Chapter 1 will discuss the version differences most in depth.

Most of you have access to Excel, and you probably already use it for reporting or recordkeeping at work. But if for some reason you don't have a copy of Excel, you can either buy it or go for LibreOffice (www.libreoffice.org) instead.

What About Google Drive?

Now, some of you might be wondering whether you can use Google Drive. It's an appealing option since Google Drive is in the cloud and can run on your mobile devices as well as your beige box. But it just won't work.

Google Drive is great for simple spreadsheets, but for where you're going, Google just can't hang. Adding rows and columns in Drive is a constant annoyance, the implementation of Solver is dreadful, and the charts don't even have trendlines. I wish it were otherwise.

LibreOffice is open source, free, and has nearly all of the same functionality as Excel. I think its native solver is actual preferable to Excel's. So if you want to go that route for this book, feel free.

Conventions

To help you get the most from the text and keep track of what's happening, I've used a number of conventions throughout the book.

Sidebars

Sidebars, like the one you just read about Google Drive, touch upon some side issue related to the text in detail.

Warning

Warnings hold important, not-to-be-forgotten information that is directly relevant to the surrounding text.

Note

Notes cover tips, hints, tricks, or asides to the current discussion.

Frequently in this text I'll reference little snippets of Excel code like this:

```
=CONCATENATE("THIS IS A FORMULA", " IN EXCEL!")
```

We *highlight* new terms and important words when we introduce them. We show file names, URLs, and formulas within the text like so:

<http://www.john-foreman.com>.

Let's Get Going

In the first chapter, I'm going to fill in a few holes in your Excel knowledge. After that, you'll move right into use cases. By the end of this book, you'll not only know about

but actually have experience implementing from scratch the following techniques:

- Optimization using linear and integer programming
- Working with time series data, detecting trends and seasonal patterns, and forecasting with exponential smoothing
- Using Monte Carlo simulation in optimization and forecasting scenarios to quantify and address risk
- Artificial intelligence using the general linear model, logistic link functions, ensemble methods, and naïve Bayes
- Measuring distances between customers using cosine similarity, creating kNN graphs, calculating modularity, and clustering customers
- Detecting outliers in a single dimension with Tukey fences or in multiple dimensions with local outlier factors
- Using R packages to “stand on the shoulders” of other analysts in conducting these tasks

If any of that sounds exciting, read on! If any of that sounds scary, I promise to keep things as clear and enjoyable as possible.

In fact, I prefer clarity well above mathematical correctness, so if you're an academician reading this, there may be times where you should close your eyes and think of England. Without further ado, then, let's get number-crunching.

Chapter 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.

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 Treats	\$ 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 Treats	\$ 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](#)).

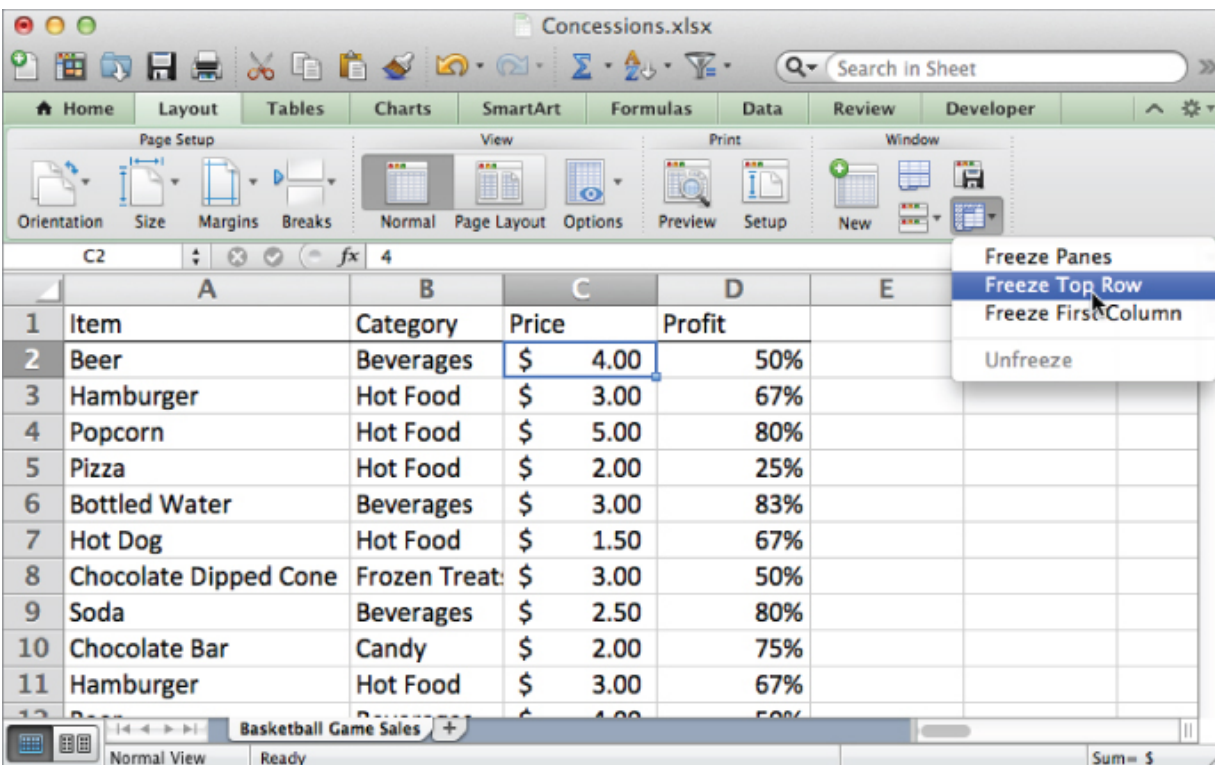


Figure 1.2 Freezing the top row

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.

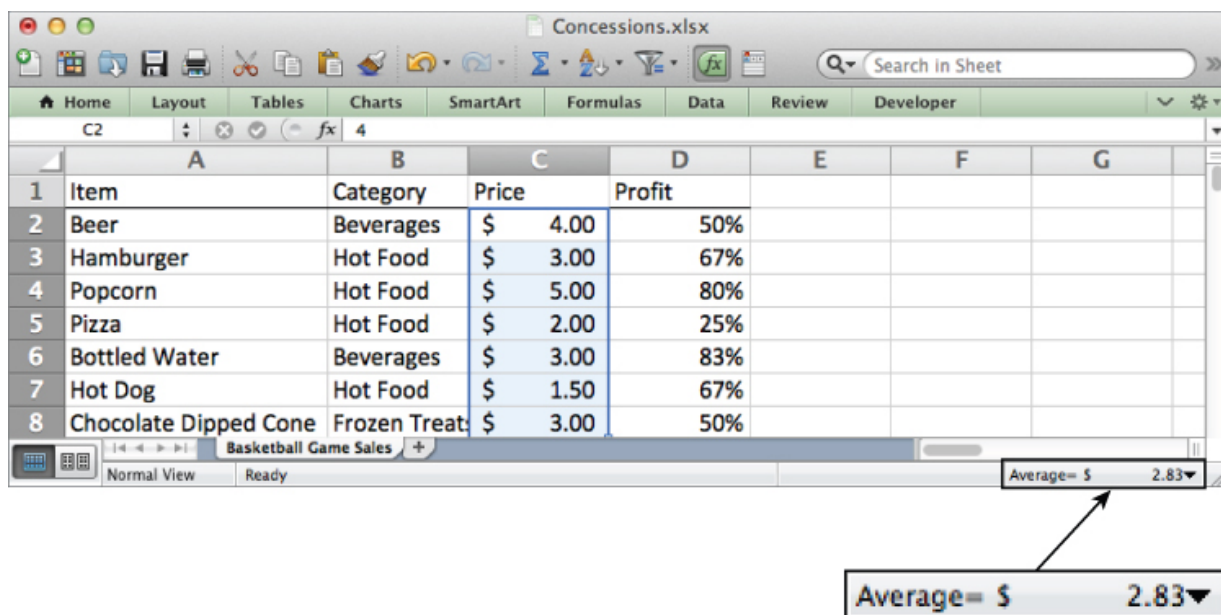


Figure 1.3 Average of the price column in the status bar

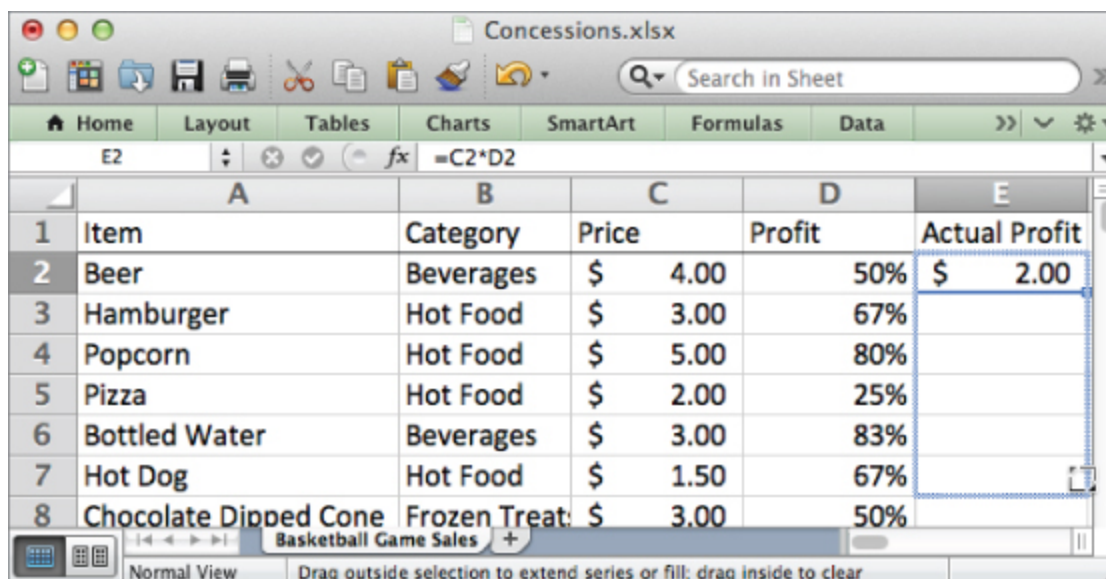
Copying Formulas and Data Quickly

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.



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

Figure 1.4 Filling in a formula by dragging the corner

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`

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.

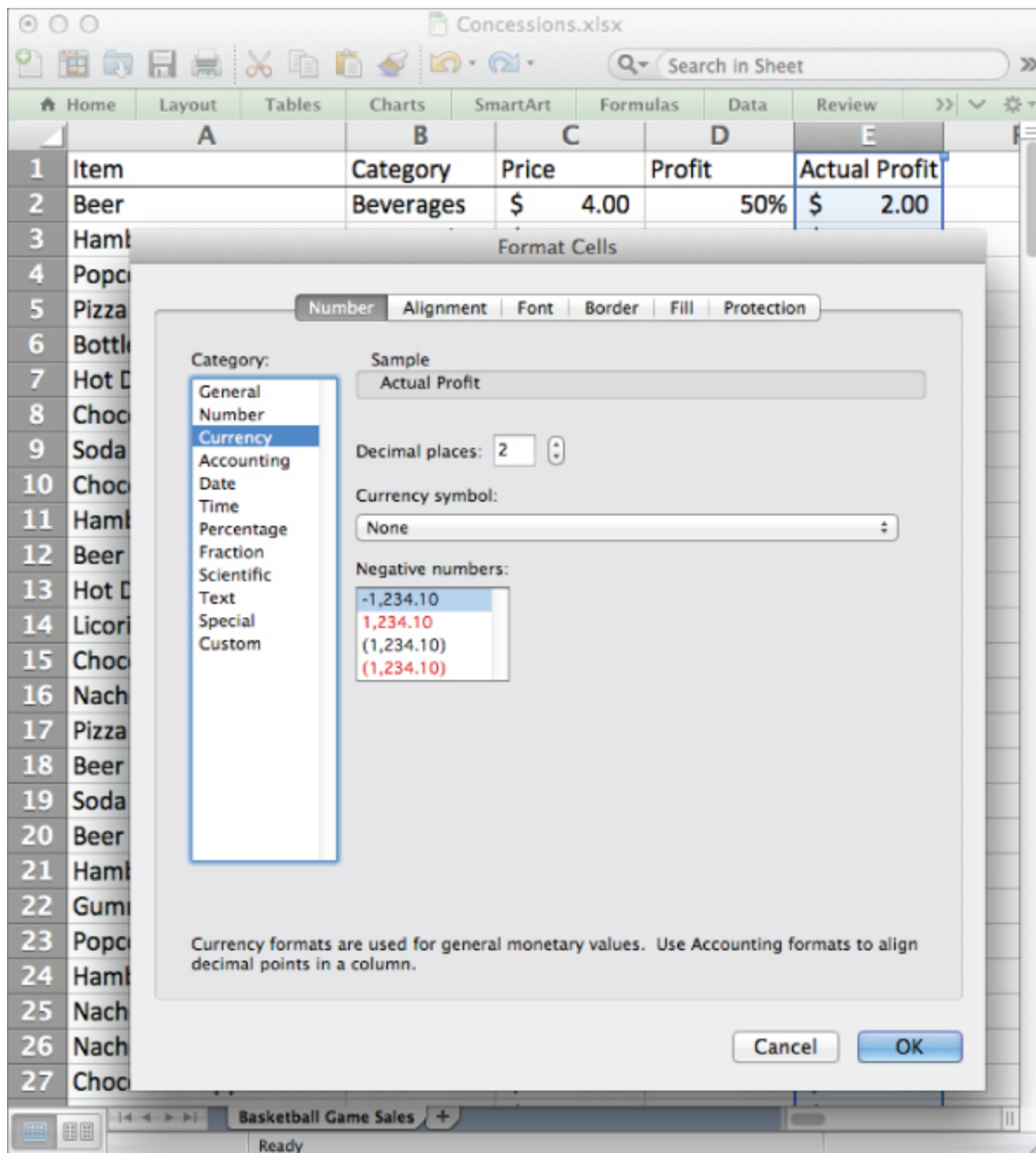


Figure 1.5 The Format Cells menu

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.

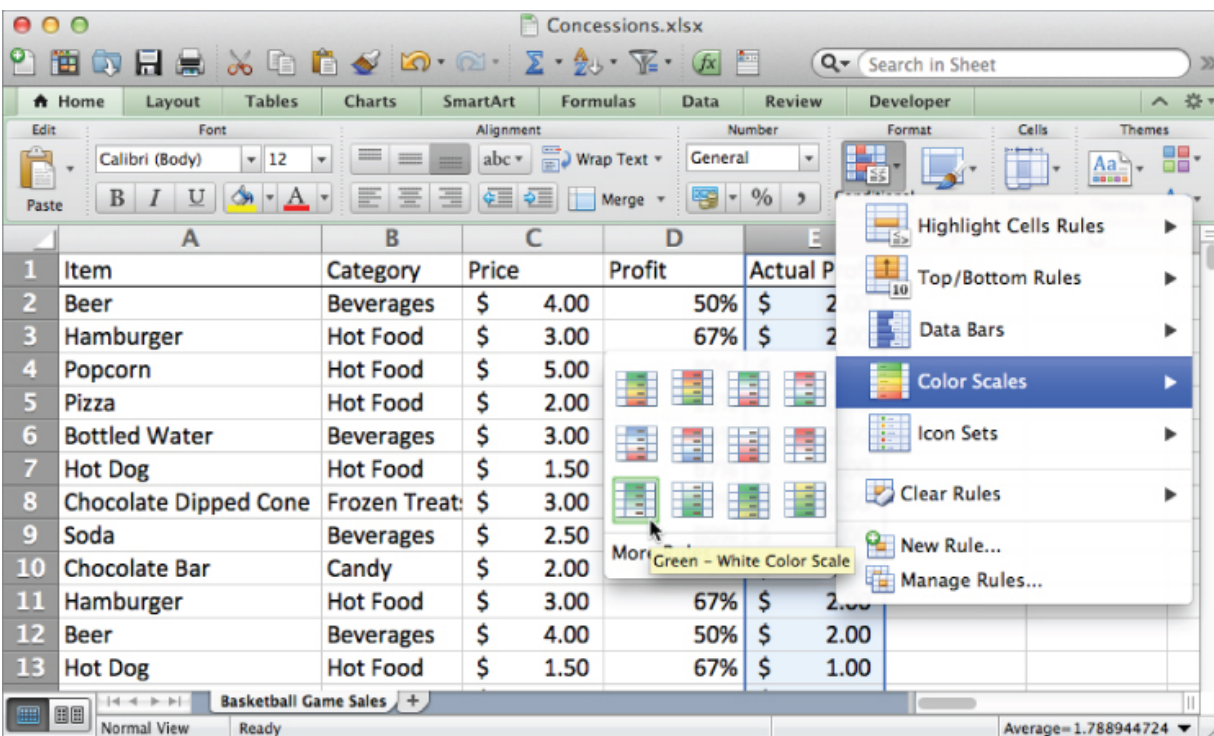


Figure 1.6 Applying conditional formatting to the profit

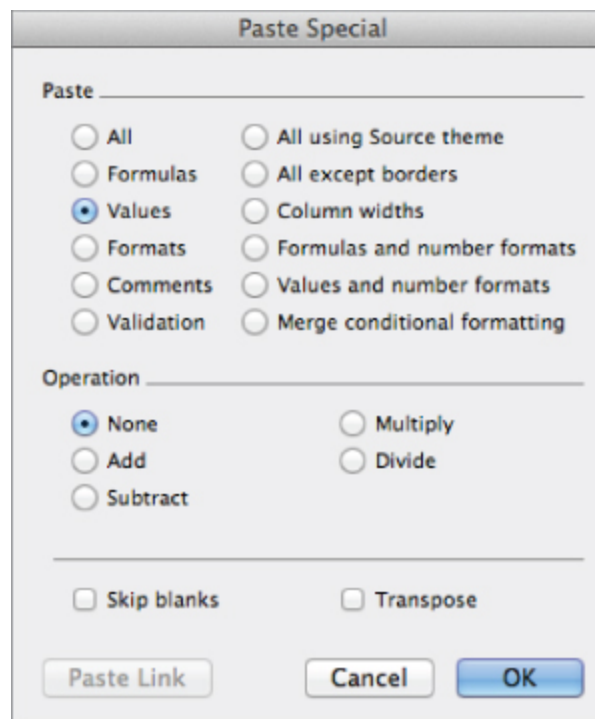
To remove conditional formatting, use the Clear Rules options under the Conditional Formatting menu.

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](#).

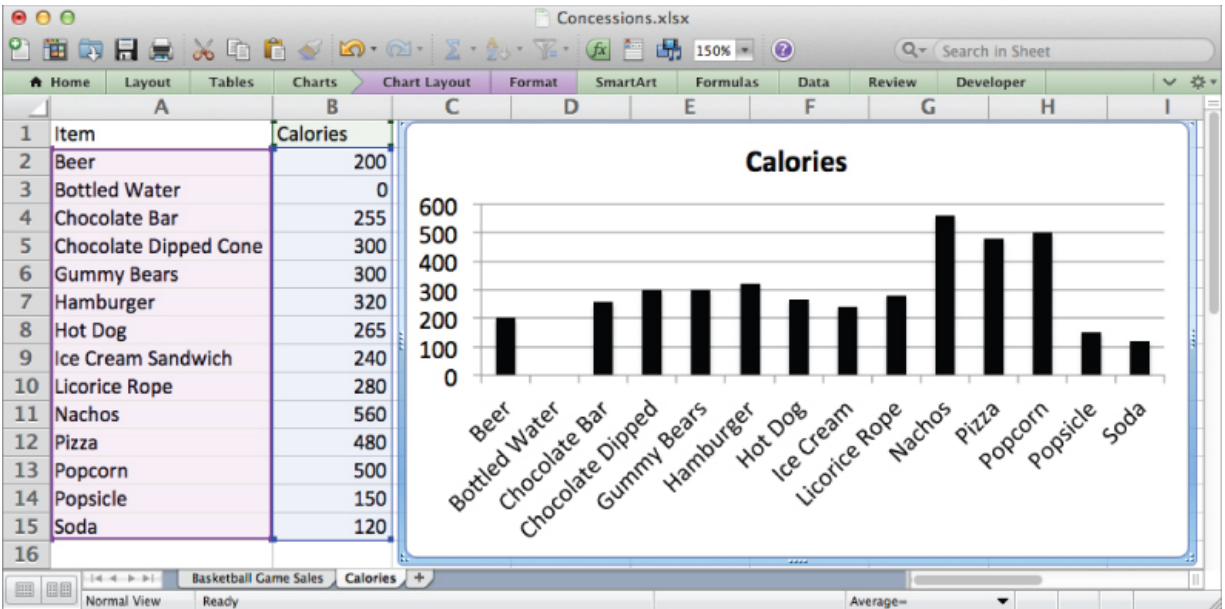


Figure 1.8 Inserting a calories column chart

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.

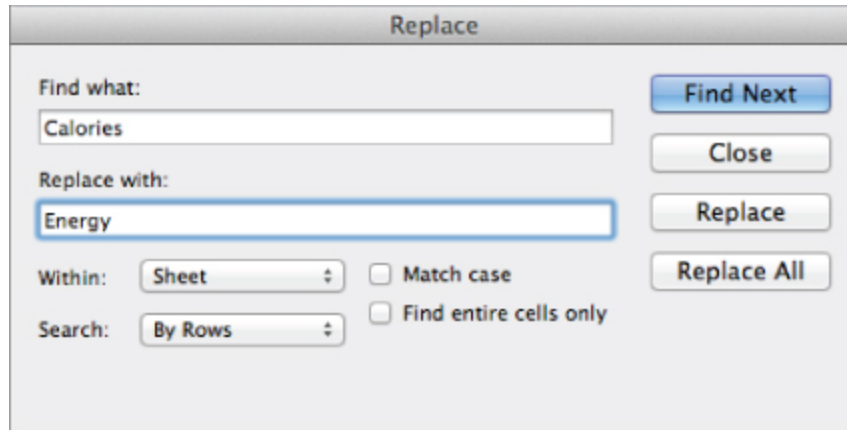


Figure 1.9 Running a Find and Replace

Formulas for Locating and Pulling Values

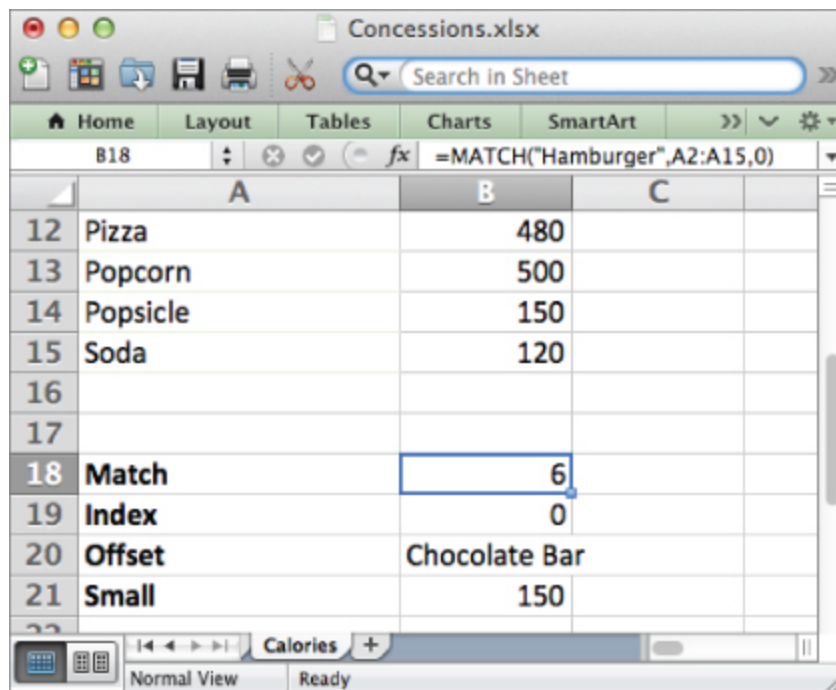
If I didn't assume you at least knew some formulas in Excel (SUM, MAX, MIN, PERCENTILE, and so on), we'd be here all day. And I want to get started. But there are some formulas used a lot in this book that you've probably not used unless you've dug deep into the wonderful world of spreadsheets. These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



	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

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.

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](#).

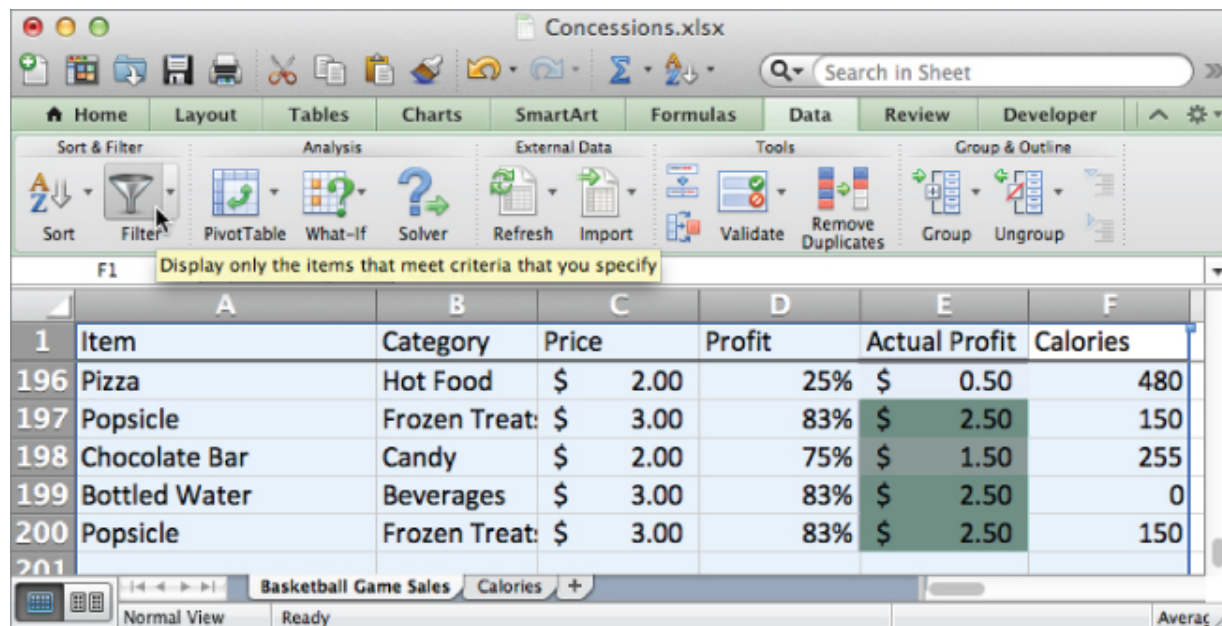


Figure 1.12 Place auto-filter on a selected range

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](#).