Baseball in Thin Air: Cleaning Data and Analyzing With Pivot Tables

Robert Gebeloff

[rgebeloff@nytimes.com](mailto:rgebeloff@nytimes.com)

In American baseball, teams are often known for the unique characteristics of their stadiums. In Boston, the Red Sox play in Fenway Park, which is known for a giant outfield wall known as “The Green Monster”. In Chicago, the Cubs play in Wrigley Field, known for walls covered in ivy.

The stadium in Denver, however, is known not so much for its physical characteristics but the atmosphere. The elevation of Coors Field is 5,200 feet above sea level. The air is so thin, the ball flies farther. And that in turn leads to more runs.

Unfortunately for the Rockies, the air is thin for BOTH teams. In this exercise, we’ll explore just how much the air impacts the team. We’ll look at a log of games played by the team in 2015, and calculate how the team performs in their home stadium vs games played on the road.

**Importing the Data**

Let’s take a look. The game log data is on the Web. You can open files on the Web directly in Excel. Go to File/Open URL and enter this address:

<http://bit.ly/1NQ1pSV>

The text import wizard will appear. In step 1, It guesses that we’re trying to open a text file where the columns are delimited by something ( commas), which is correct, so we can click “next”.

In step 2, we confirm that our delimiter is in fact a comma. In many cases, data will also come wrapped in “” marks. This is not true in this case, but it’s important to be aware of this possibility.

Step 3 is where we can make sure our data is optimized for analysis. One of the big mistakes you can make here involves data types. Excel tries to guess whether each column is a number, date or text but it’s smart to review the choices.

Scanning the data we’re about to import, you can see one possible problem, the column labeled “W-L” – that’s the won-loss record of the team after every game. The dash is a problem – Excel will likely interpret that as a minus sign or a date signifier, so we need to specify during our import that this column is text.

Once all of the column types look ok, we can go ahead and finish the wizard and our data is now in the spreadsheet. Keeping the good practice of saving this as the original data and making a working copy, hit cmnd+A to select everything, cmnd+c to copy it, make a new worksheet, and hit cmnd+v to paste it in. Let’s name our sheets “raw data” and “prepared data” and save the file to your computer.

**Cleaning the Data**

Most data you download from the Web requires some cleaning before it’s ready to be analyzed. But before you can clean it, it’s important to think through as many of your possible “interview” questions first.

So let’s start there. Some of the things we’ll want to know are:

1. What is the team’s record at home and on the road?
2. How many runs does Colorado score and allow at home and on the road?
3. What is the average scoring on the home and road?
4. What is the team’s record by month? Day of the week? Day game vs. night game?
5. Are there any interesting patterns with attendance?
6. What is the team’s longest winning streak/losing streak?

Ok, so now let’s look at our data and see what we have to clean. In this case, I’m going to work from left to right on the spreadsheet and clean up anything that I think needs cleaning.

The first candidate is the “Date” column. The date in this case is in a single column, and formatted so it’s easy for a **human** to read but not a **computer**. We’re going to need to “group” our data by month, for example, and Excel will only be able to determine the month if we reformat the data into something it can understand.

The good news is that the date is in a **consistent** format in that each element is separated by a single space: day of week, month and day. Note that for two days – May 6 and September 1, there are two games.

Let’s use Excel’s “Text to Columns” took to put each into its own, separate column.

1. Click on column D and then on the menus, Insert/Columns. Hit cmd-Y on a Mac, f4 on a PC to repeat three more times to create four blank columns in all.
2. Copy column C into column D – this will preserve our original data in column C while we slice up column D.
3. Now on the menu click Data/Text to Columns and a wizard will pop up. In this step 1, we tell Excel that our data is “delimited”. Hit next.
4. Step 2 is where we tell Excel that our delimiter is a space. Once we click that box, Excel should propose dividing lines for our data. If you scroll down, notice how it accounts for the double-header games as well. Hit next.
5. Step 3 gives us the option of forcing our data into a particular format. This would be important if we had, say, a postal code that began with a zero; in this step, we could tell Excel to preserve the zero by forcing the column to text format. For now, it looks good, so go ahead and click Finish.
6. Let’s put headers on our new columns – Pivot tables require them. Let’s use Day, Month, Daynumber, and DH

Moving on, column H is extraneous --it’s the word “boxscore” that probably had a link to detailed results for the game. We don’t need it, let’s select this column and delete it.

Column I is missing a header and is hard to understand. The @ sign is commonly used in baseball to designate “away game”, so let’s clear that up. Let’s put a header on that column, homeaway. But now let’s click on that column and insert a new one and call it homeaway\_new. In this column we’re going to enter a formula to translate the symbol into something easier to understand.

In cell J2, enter the formula:

=IF(I2="@","Away","Home")

This makes use of Excel’s powerful “IF” function. It literally means: If cell I2 equals the @ sign, then return the value “Away”, otherwise, return the value “Home”. Once you have the formula working, copy it down the rest of the column.

Column L is an important one – it tells us whether the Rockies one or lost that game. But it also in some cases has an extra designation, “wo” – this notes a game where the winning team ended the game by scoring, and therefore “walked off” with a win. (In baseball, a team wins if they have the most runs after 9 innings, so often the winning team ends the game while on defense.)

We don’t care about this detail for this analysis, so let’s make a won-loss column with just the “W” or “L”. We’ll use a different Excel function. Click on column M and insert a new column. Then in cell M2 enter the following formula:

=LEFT(L2,1)

This literally means look in cell L2 and return the 1st character on the left. Once you have it right, copy it down the rest of the column.

I know by now you’re dying to get to the analysis but there’s a couple of more cleaning tasks. Let’s get the winning/losing streak data into better shape using both of the functions we just learned. Let’s put a new label in column AA called “streak type” and another in column AB called “streak length”.

In cell AA2, enter the following formula:

=IF(LEFT(Z2,1)="+","winning","losing")

Can you interpret what this means? In cell AB2, we’ll use the LEN function to count how long the streak was on that particular day:

=LEN(z2)

Copy those formulas down the columns.

By now, you’re thinking – We must be ready to analyze!. Almost… Scroll down the spreadsheet vertically and you’ll see after game 50 that the headers repeat themselves. We need to get them out of our data, and for this job, we’ll use a filter.

Select Row 1 and click on Data/Filter. Each column will now have drop-down arrows. Use the drop-down on column A, and a filtering dialog box will appear. Unselect “Select All”, and then scroll down until you see the value “RK” – check that box, and then close the filtering dialog box.

You’ll see rows 52 and 103 as the offenders. You can delete them, and then go back to Data/Filter to turn your filters off.

You are ready to pivot! For this part of the exercise, I will leave you to apply your pivot skills learned in previous exercises. Remember, you want to select all of the data, then go to Data/Pivot Table, and use the builder to lay out your results table. And just as a reminder, here are some of the questions you would want to answer:

1. What is the team’s record at home and on the road?
2. How many runs does Colorado score and allow at home and on the road?
3. What is the average scoring on the home and road?
4. What is the team’s record by month? Day of the week? Day game vs. night game?
5. Are there any interesting patterns with attendance?
6. What is the team’s longest winning streak/losing streak?

Some sample pivot tables are provided on the class file page, but please don’t look until you’ve tried getting through the list yourself.