Airline passenger data analysis

This data is from U.S. Department of Transportation. It was downloaded from this searchable database that is regularly updated:

<http://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=258&DB_Short_Name=Air%20Carriers>

The slice we are working with shows the number of passengers, each month in 2013, broken down by the airline and the route. So for example, one record would indicate the total passengers who flew on Delta from Minneapolis-St. Paul to Detroit in the month of January. And another record would show how many took the reverse route -- Detroit to MSP -- on Delta that month.

Let's start with some housekeeping.

Notice that the data doesn't indicate what year this came from. I like to make sure that information is stored within the data, just in case I need it in the future. (And, for something like this, you might be interested in adding more data to it in the future)

So let's add a new field, called "Year" and populate it with "2013" for all the records

Next, let's create a new field that combines the origin and destination so we can use that for a pivot table to add up the total passengers for each route. Name the new field "Origin-Dest"

=TRIM(D2)&"-"&TRIM(F2)

Next, let's flag all the records that have either an origin or a destination as MSP. Name the new field "MSP flag"

To do that, we'll use an IF THEN statement

=IF(D2="MSP", D2, IF(F2="MSP", F2, ""))

One of my first questions is: How many passengers flew on each route for the year in total (regardless of carrier)?

Create a pivot table. Put the new "origin-Dest" field in the ROWS

Put "Passengers" in VALUES and make sure it is summing that field

Then sort by the sum of Passengers field, descending

You'll see that the usual suspects are at the top -- LAX to SFO (and vice versa), JFL-LAX (and vice versa)

What about MSP? I'd like to know which routes are the most heavily traveled in and out of MSP. An easy way to do that is to add the "MSP flag" field to the FILTERS box in your pivot table and then choose "MSP" from the dropdown that appears at the top of the page. This will filter your pivot table so it only displays records where MSP is in that "MSP flag" field.

That shows us that Chicago O'Hare to MSP is the most heavily traveled (no surprise there), followed by Atlanta and Denver.

One of the things I don't like about the pivot table we just made is that there are two records for each route -- i.e. O'Hare to MSP and MSP to O'Hare. Wouldn't it be great to total up the passengers regardless of which direction they flew?

We can do that by a little trick to re-create that "origin-dest" field so that we alphabetize the airport codes, setting it so that it always puts them in the same order.

Create a new field called "route"

Again, we'll use IF THEN statement -- this time we'll use two of them and also use the concatenation that we did for the first "origin-dest" field.

=IF(D2<F2, D2, F2)&"-"&IF(D2<F2, F2, D2)

This says: if the origin (D2) is "less than" (i.e. comes first in the alphabet) the destination (F2), put D2 first. If not, put F2 first. Then there's an ampersand to connect in a dash and another ampersand to connect in the final airport code, which is the reverse of the first IF statement.

NOTE: We need to use the airport codes for this because of places like Chicago where there is more than one airport. The origin city name and destination city name fields are the same for those airports.

Repeat the pivot table we made earlier, but this time use the Route field. Also add the MSP filter.

MaryJo Webster

@MaryJoWebster

mjwebster71@gmail.com