Creating a Simple Database: An Excel Strawman

ACTIVITY 6

Most students are familiar with an Excel spreadsheet. Data in a spreadsheet – in the form of rows and columns – are a simple form of a database. This activity aims to expose students to database concepts using this as a strawman. Since Excel format is binary and proprietary, data cannot be read without Excel or other software. The comma separated values (CSV) file is a more common way that spreadsheet-like data can be exchanged or provided between source and analysis program. Since it is not formatted for a specific software type and is ASCII text file, it can be read by most software. In this file format, the comma separates values between columns of data. The first row is typically, but not always the names of the columns. The comma is a called a "delimiter". Note that characters other than commas can be used as a delimiter such as ";" or "|" or other text characters.

In this activity you are given two files that are excerpts from TriMet's stop-level bus AVL system data. Don't worry too much about what data these files represent, we will work on that later. They have been edited for demonstration and learning purposes. They are:

- **stop_level.csv** (sample of TriMet's Bus Dispatch System data)
- **stop_names.csv** (from the RLIS regional database this is table information about stops on the TriMet system)

Purpose

The purpose of this activity is to expose the student to a simple model of relational database using Excel as the strawman.

Learning Objective

- Using Excel as a simple model of database, be able to define tables, columns, rows, data types, and do simple joins.
- Using filters and joins, answer a set of simple count-type questions.

Required Resources

- Microsoft Excel
- Files posted on class site: **stop_level.csv** & **stop_names.csv**
- TriMet metadata description posted on the class website

Time Allocated	90 minutes in class

Tasks

A. Working with CSV files

Retrieve the **stop_names.csv** from the class web site and open the csv file in a text editor, such as Notepad or Textpad. A good text editor is going to be very useful. Notepad or Wordpad has no tools or options and has a row limitation (on the number of records it can read in). My suggestion is to use Textpad. You can find it under "General Applications".

1. How many rows of data do there appear to be (Hint: use the status bar in Textpad)?

Excel can also easily read the csv files into a worksheet:

- Open a blank workbook in Excel. Create 2 tabs, stops and stop_names.
- To read the files into Excel there are a couple of options:
  1) go to the ***Data*** tab and ***Get External Data***, select ***From Text*** and follow the prompts.
  2) select the data from the text file, paste in Excel, then use the ***Data → Text to Columns*** option.

  Feel free to try both.

B. A Simple Database

A relational database allows data to be stored in a more compact format and can be a useful in doing analysis. The most common relationship is what is called a *one-to-many* relationship.

An easy way to think of this is represented below. In one table, a column has entries in each cell for data elements. The data elements in the column are often repeated. In the example below, the table SALES is a log of all customer transactions, the field CUSTOMERID can be related to another TABLE (CUSTOMER) that contains all the information about the customer. In this way, there is no need to repetitively store all of the information about each customer. Rather, this information can be extracted at any time by joining (i.e., linking) the two tables on CUSTOMERID column.

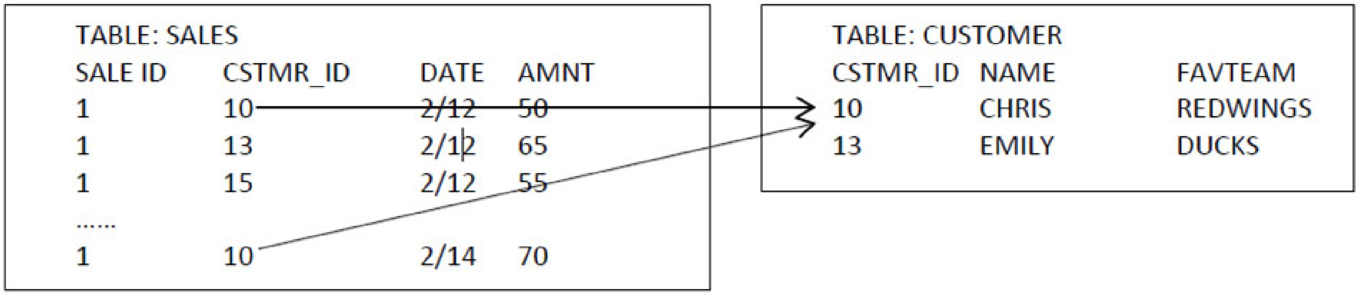


TABLE: SALES

SALE ID	CSTMR_ID	DATE	AMNT
1	10	2/12	50
1	13	2/12	65
1	15	2/12	55
……			
1	10	2/14	70

TABLE: CUSTOMER

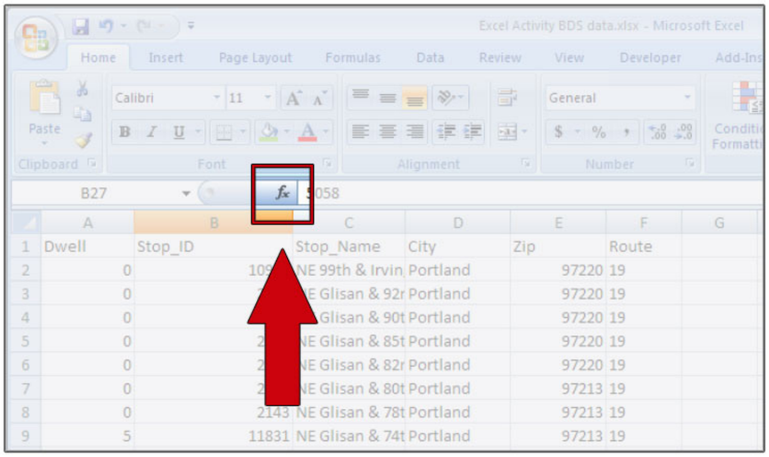
CSTMR_ID	NAME	FAVTEAM
10	CHRIS	REDWINGS
13	EMILY	DUCKS

Figure 4 Sample Database

Let's try a simple example of linking the two tables you have in Excel. First, look at the two tabs.

2. Which column is the obvious one to make the join?

Excel really isn't a database, though it has some functionality that is similar to a database. Note: Microsoft's desktop database is Access. One way to make the join, is to use the ***VLOOKUP*** function. TIP: Use the function insert tool in Excel to help you learn about ***VLOOKUP*** (see the arrow pointing to it in Figure 5).



	A	B	C	D	E	F	G
1	Dwell	Stop_ID	Stop_Name	City	Zip	Route	
2	0	109	NE 99th & Irvin	Portland	97220	19	
3	0	[illegible]	E Glisan & 92r	Portland	97220	19	
4	0	[illegible]	Glisan & 90t	Portland	97220	19	
5	0	2	NE Glisan & 85t	Portland	97220	19	
6	0	2	NE Glisan & 82r	Portland	97220	19	
7	0	2	NE Glisan & 80t	Portland	97213	19	
8	0	2143	NE Glisan & 78t	Portland	97213	19	
9	5	11831	NE Glisan & 74t	Portland	97213	19	

Figure 5 Excel Screen Capture of Stop Table

Produce a third table that has the one column from the main stop table and all of the columns from the joined table.

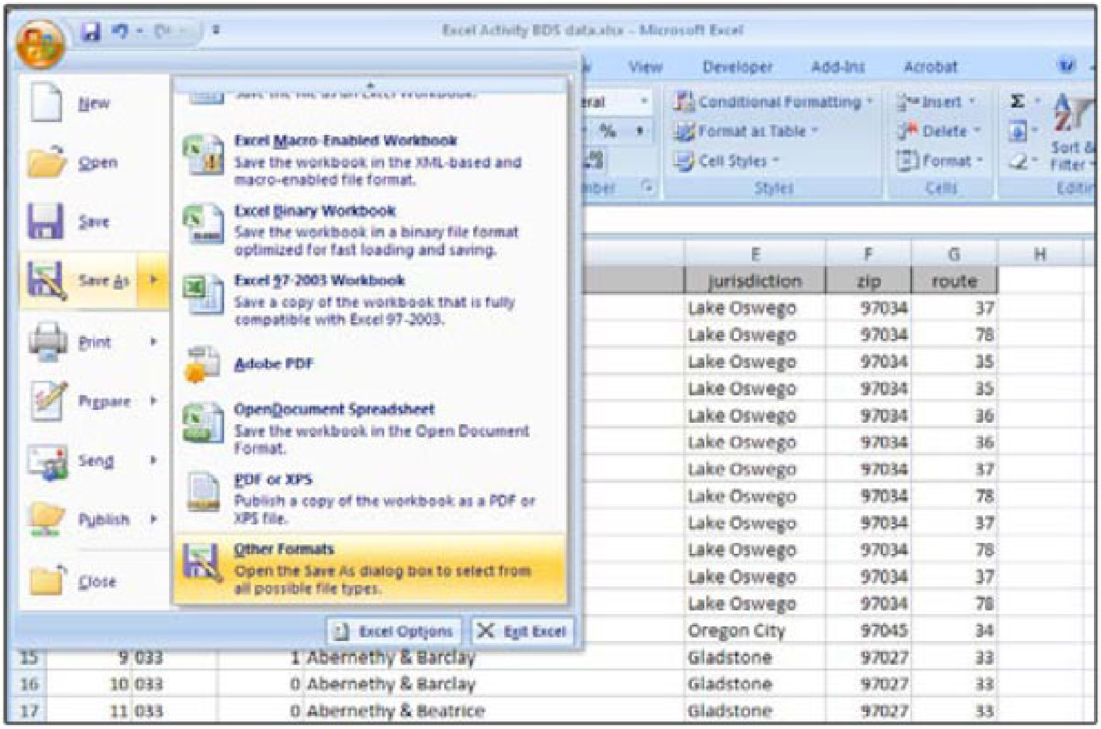
It should be clear what fields do this "join" on

Now, let's take a minute to explore the "stops_name" tab in the spreadsheet. In Figure 4, the table "Customers" is like the "stops" data. It contains information about each stop. There is one important difference about the stops table

3. What is it?

4. Now look at the “joined” table you created. What looks wrong about this? Hint: look at the Rte_No in the stop level data. How would you fix this?

Export this tab to a CSV format. Before you do that, save the workbook that you are currently working in. To export, use the ***FILE*** → ***SAVE AS*** and select other format (see Figure 6).


Excel Activity BDS data.xlsx - Microsoft Excel
View
Developer
Add-Ins
Acrobat
Conditional Formatting
Format as Table
Cell Styles
Styles
Insert
Delete
Format
Cells
Sort & Filter
Editing
New
Open
Save
Save As
Print
Prepare
Send
Publish
Close
Excel Macro-Enabled Workbook
Save the workbook in the XML-based and macro-enabled file format.
Excel Binary Workbook
Save the workbook in a binary file format optimized for fast loading and saving.
Excel 97-2003 Workbook
Save a copy of the workbook that is fully compatible with Excel 97-2003.
Adobe PDF
OpenDocument Spreadsheet
Save the workbook in the Open Document Format.
PDF or XPS
Publish a copy of the workbook as a PDF or XPS file.
Other Formats
Open the Save As dialog box to select from all possible file types.
Excel Options
Exit Excel

					E	F	G
					jurisdiction	zip	route
					Lake Oswego	97034	37
					Lake Oswego	97034	78
					Lake Oswego	97034	35
					Lake Oswego	97034	35
					Lake Oswego	97034	36
					Lake Oswego	97034	36
					Lake Oswego	97034	37
					Lake Oswego	97034	78
					Lake Oswego	97034	37
					Lake Oswego	97034	78
					Lake Oswego	97034	37
					Lake Oswego	97034	78
					Oregon City	97045	34
15	9	033	1	Abernethy & Barclay	Gladstone	97027	33
16	10	033	0	Abernethy & Barclay	Gladstone	97027	33
17	11	033	0	Abernethy & Beatrice	Gladstone	97027	33

Figure 6 Exporting the Excel file to a CSV format

C. Tabulating Some Answers

Let’s explore the Excel “database” in a little more detail. Use the filter option in the SORT & FILTER tools to answer these following questions.

5. How many stops were made at STOP_ID 805?

6. How many stops were made at STOP_ID 2001?

7. How many stops were made where the STOP_ID was greater than 1500 in the table **stop_level**?

8. How many stops were made at time points (use the TriMet data dictionary ***http://www.gcu.pdx.edu/data/dictionary.htm***)?

9. How many routes serve Cornelius?

10. Which routes serve the intersection of W Arlington & Barton?

11. What is the total amount of dwell time that occurred with a stop on Glisan? (*Hint: use text filters for cells that contain Glisan.*)

Deliverable

Provide a short-typed text or Word document with your answers to questions 1-11 due at the end of this class session. Upload a PDF of the document to the class dropbox

Assessment

Participation for this activity is based on the submission of answers and text files.

Student Notes