

## Using Excel's PivotTable to look at Bivariate Categorical Data

This handout briefly discusses MS Excel's PivotTable routine. PivotTables are very powerful tools to sort through data and make different summaries. They will prove useful for other than statistical analyses as well.

Consider the Housing Data. Let's suppose that we would like to examine whether basements are more common in Cape Cod, two-story or Ranch houses (both variables are clearly categorical). Our end result could look like the following *contingency table*:

Count of basement	Basement		Grand Total
	0	1	
0	14	25	39
1	0	25	25
2	3	41	44
Grand Total	17	91	108

*"Cape Cod"*  
*"Two-story"*  
*"Ranch"*

17      91  
*No*  
*Basement*    *Basement*

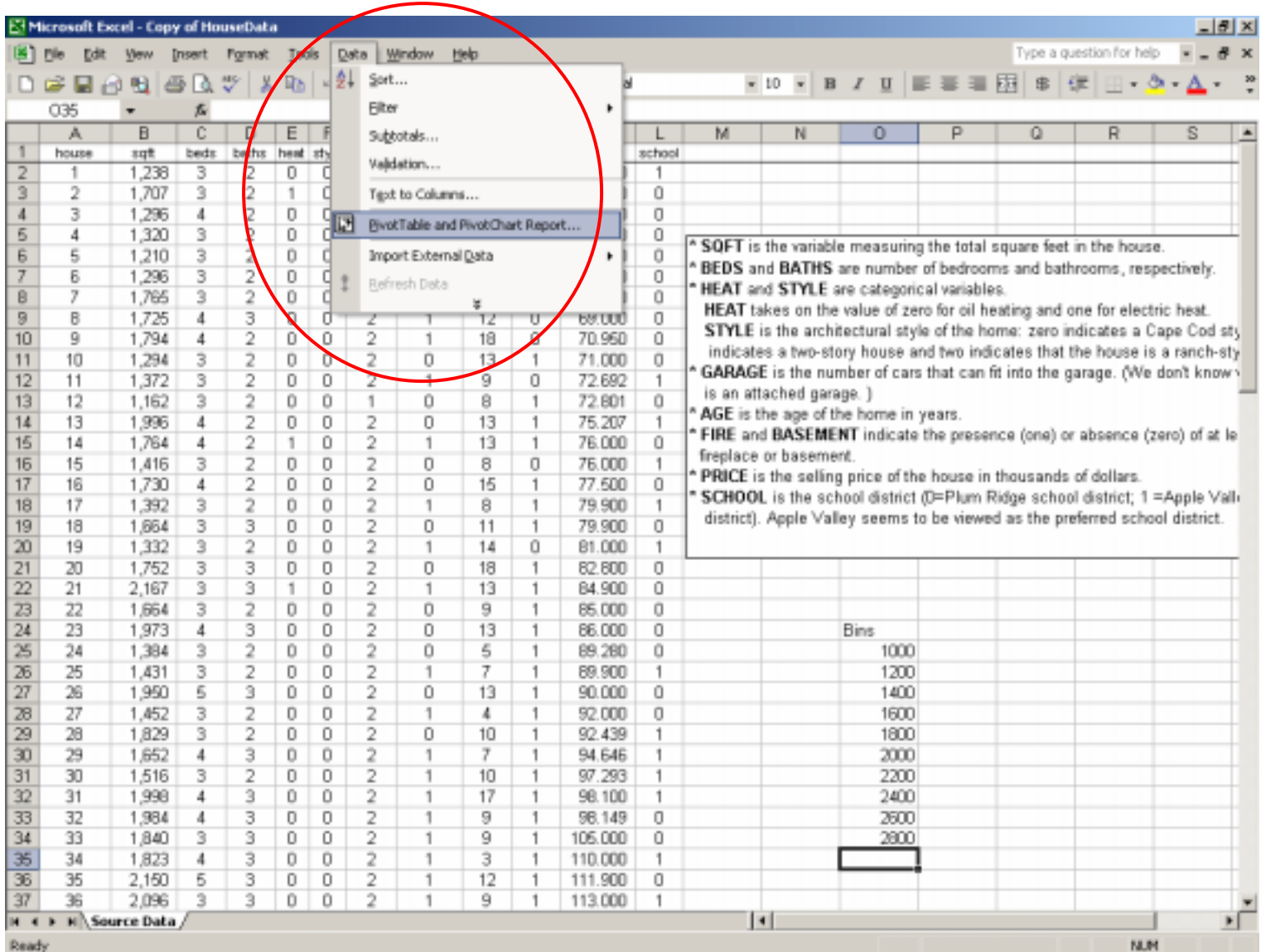
We could interpret the results as follows. Of all 17 houses without basements, 14 (over 82%) were in Cape Cod houses, as compared to only a little over 27% of the houses with basements in Cape Cod style. So, a Cape Cod style house is more likely to have no basement than having a basement. In contrast, all of the two-story homes had a basement. If the presence or absence of basements would be independent of the house style, we would roughly see 1/3 of the houses without (and with) basements for each house style, which is clearly not the case here. So, in conclusion, we can say that there *seems* to be a relationship between the categorical variables *house style* and the *presence of basements*.

We will also show how we could generate a contingency table with *percentages* rather than absolute counts, such as the examples in Tables 2.11, 2.12 or 2.13 in the textbook, and for the Housing Data may look like this:

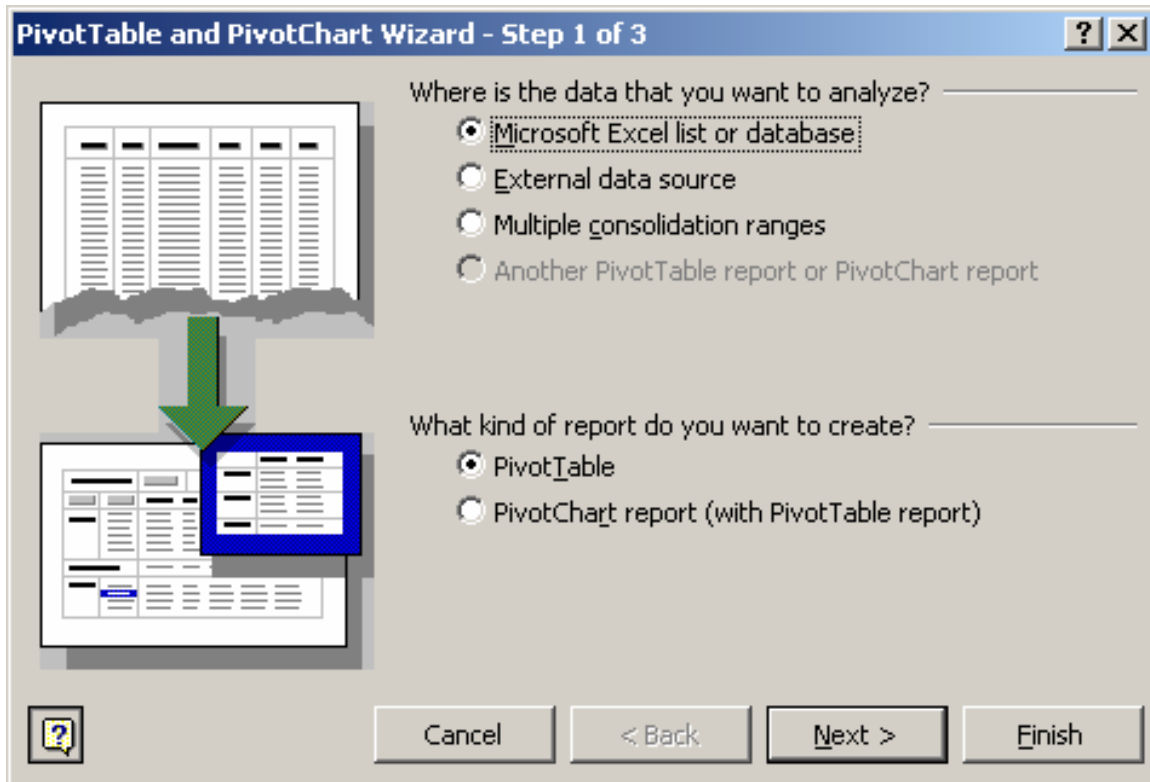
Style	Basement	
	No	Yes
<b>Cape Cod</b>	82.35%	27.47%
<b>Two-story</b>	0.00%	27.47%
<b>Ranch</b>	17.65%	45.05%
Grand Total	100.00%	100.00%

Again, we may interpret this table by saying that of all the houses without a basement, over 82% are Cape Cod style houses, which again gives evidence that there is a relationship between house style and the presence or absence of a basement.

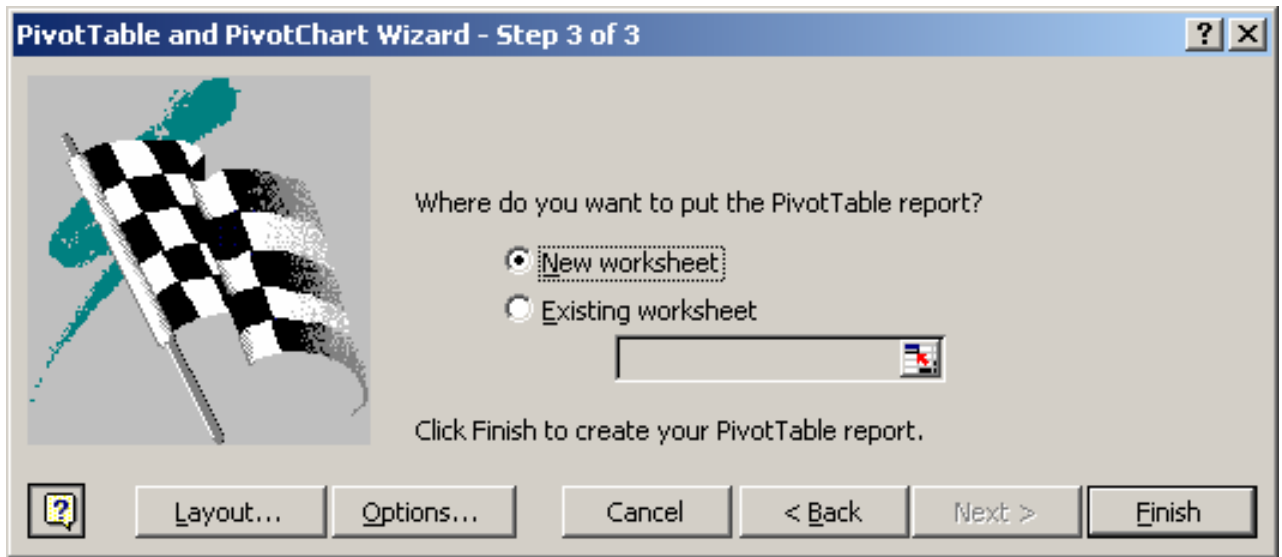
Before we look at the *bivariate* case, it may be easier to start with a one-way table that summarizes the categorical data for just *one* variable. The output of this one-way table may then easily be used with graphical displays such as Pie Charts, Bar Charts or Pareto diagrams. Here are the steps to accomplish this. From the Command Menu, choose *Data*, then *PivotTable and PivotChart Report*:



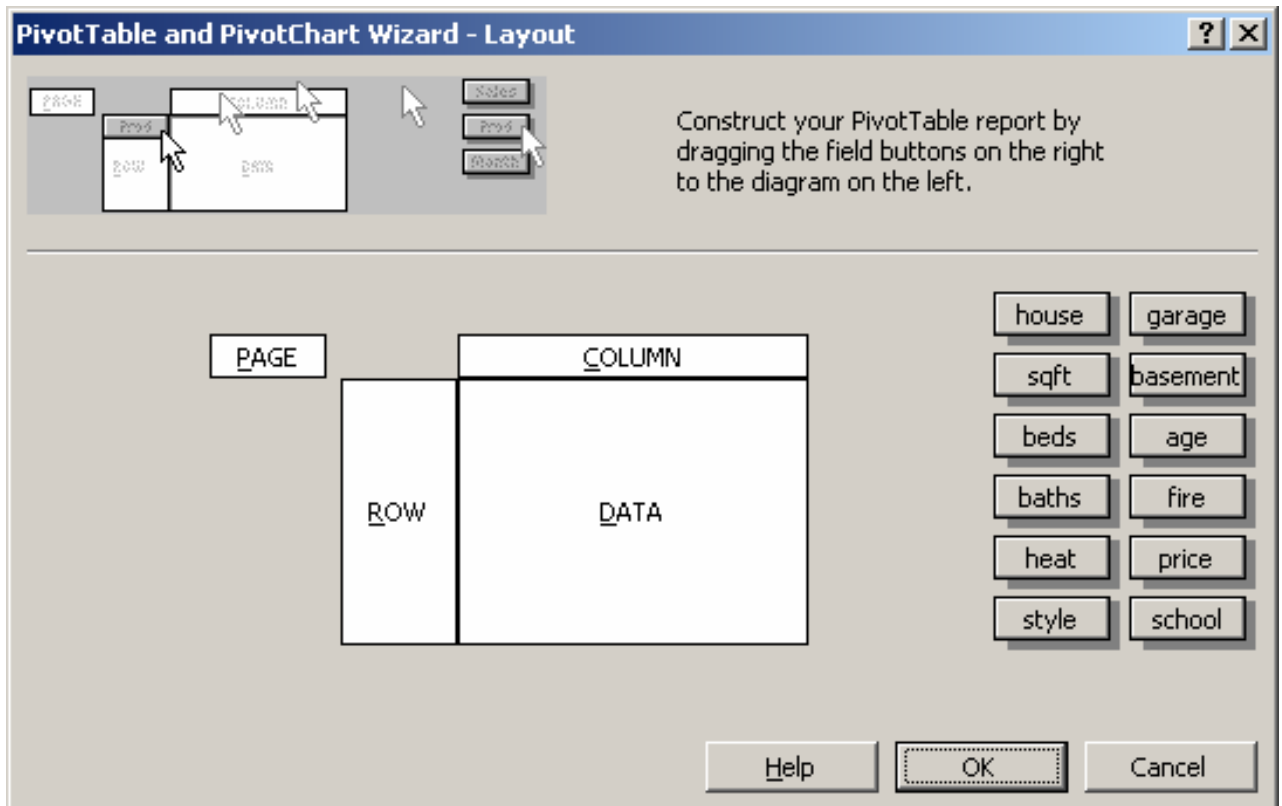
The PivotTable Wizard gets started and the first dialog box shows. We will use the default settings for this Step. Clicking the *Next* button takes us to Step 2.



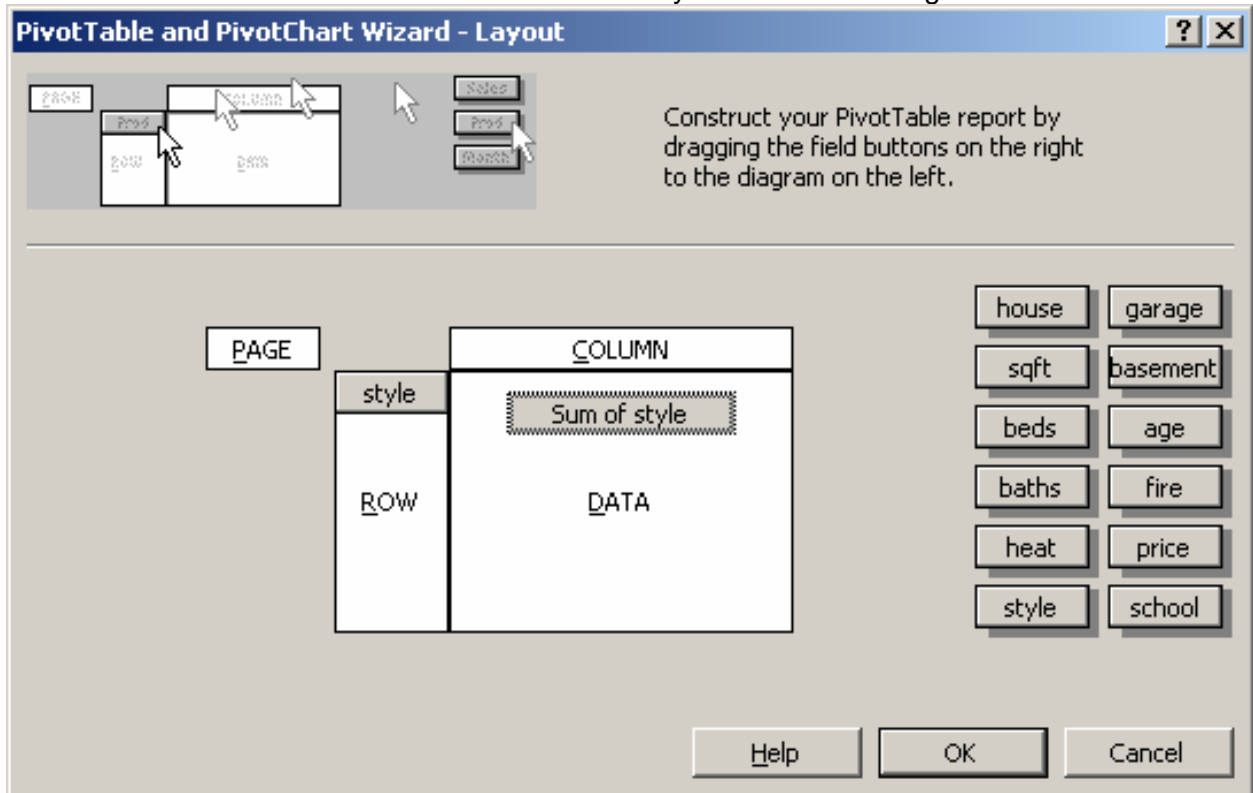
In this Step we choose the data range we want to use for our PivotTable. Excel makes a clever guess and in most instances finds the range of your data automatically and all you have to do is click the *Next* button. In some exceptional cases you may have to enter the range yourself, or select with the mouse after pressing the *red arrow* button on the right. Do not worry that Excel chooses a range that is too big (with, e.g., numerical as well as categorical variables). It comes in handy later when you can do whatever analysis you want (and didn't think of when creating the PivotTable) since all variables are available. The last Step of the PivotTable wizard comes next.



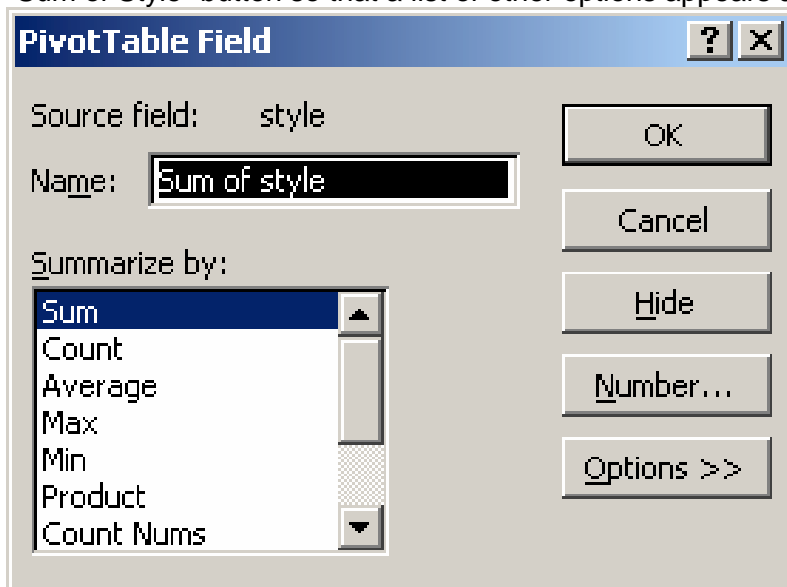
Here we need to specify where the output (our PivotTable) will go. It's always a good idea to generate the results in a new worksheet. Before pressing the *Finish* button, we would like to specify the *Layout* now, so when pressing this button, we get:



Now we can drag and drop the categorical variables to either rows or columns, and then specify what we would like to compute for each value of this variable. Suppose we are interested in knowing how many houses are in each style, then for the Rows we would drag the *Style* button into the ROW rectangle as well as to the DATA rectangle where we want to *count* the number of occurrences of each style. So far the dialog box looks like:



We don't want to "sum" the styles, but count them, so we need to double-click on the "Sum of Style" button so that a list of other options appears as follows:



Choose “Count” from the “Summarize by” list. <sup>1</sup> Click *OK* on this dialog box as well as on the previous one, and the *Finish* button on the Step 3 dialog box. You should now see the following:

The screenshot shows Microsoft Excel with a PivotTable and the PivotTable Field List task pane. The PivotTable is located in the range B3:D8 and has the following data:

style	Total
0	39
1	25
2	44
<b>Grand Total</b>	<b>108</b>

The PivotTable Field List task pane on the right shows the following fields:

- house
- sqft
- beds
- baths
- heat
- style**
- garage
- basement
- age
- fire
- price
- school

The PivotTable dialog box is also visible, showing the PivotTable name and various options.

The output should speak for itself. Try dragging the “style” button in cell A4 to B3. The table “pivots”! Hence the name PivotTable.

<sup>1</sup> The curious reader may want to try the following to enhance his/her understanding of what PivotTable can do. Instead of dragging the “style” button to the DATA rectangle, drag the “price” button there, and double-click. From the “Summarize by” list, choose the “Average” option. Now the PivotTable will compute the average sales price of a house by style!

The screenshot shows Microsoft Excel with a PivotTable and two floating menus. The PivotTable is located in cells B3:E5 and has the following data:

Count of style	style	0	1	2	Grand Total
Total		39	25	44	108

The PivotTable Fields task pane on the right shows a list of fields: house, sqft, beds, baths, heat, style, and garage. The 'style' field is currently selected. Below the PivotTable, a floating menu titled 'PivotTable' is visible, and another floating menu titled 'PivotTable Fields' is also present.

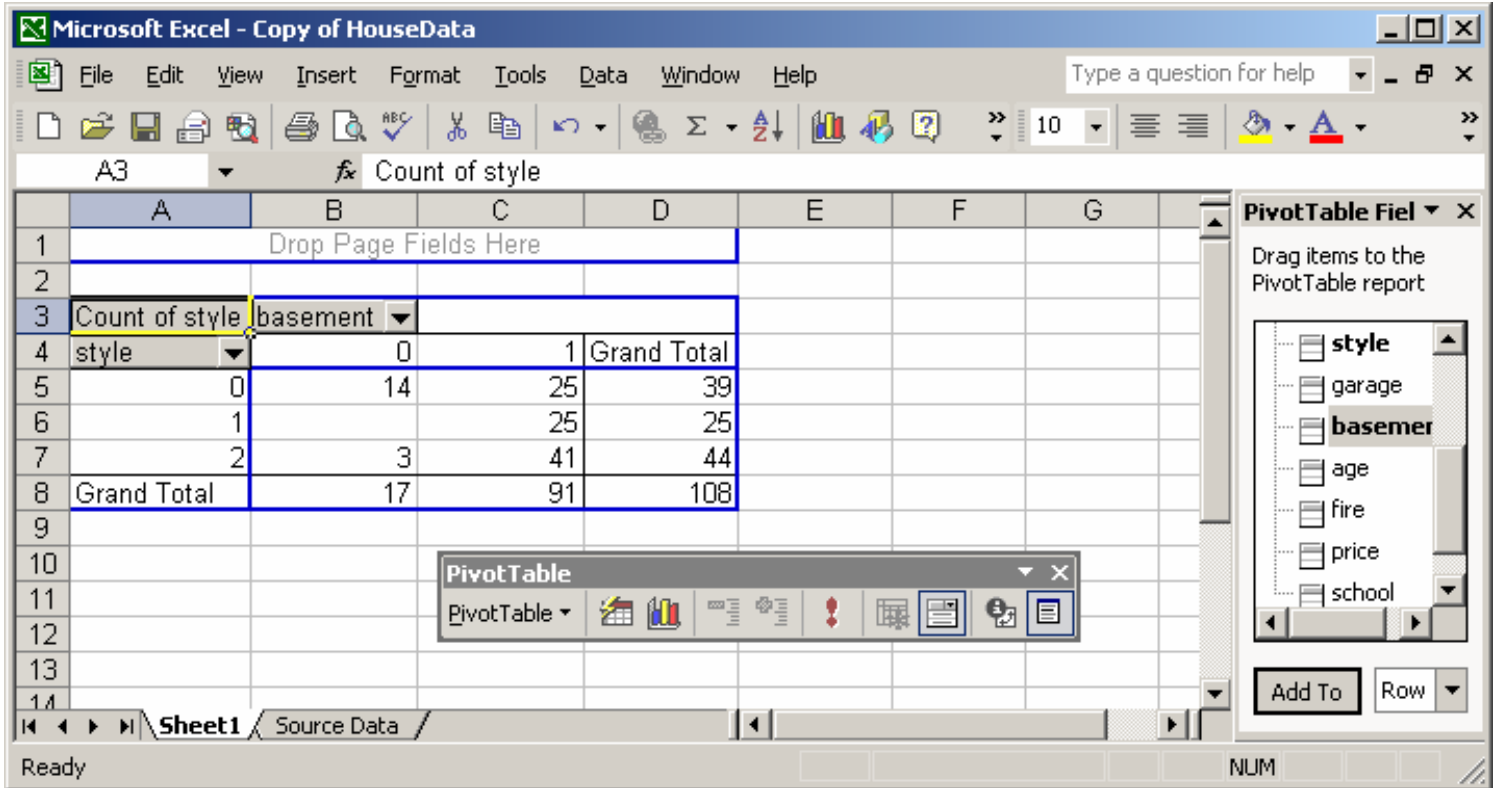
Note that we have two floating menus: one below the PivotTable (labeled “PivotTable” as well), and one on the right (labeled “PivotTable Fields”). Now we can generate new analyses and views of the data *on-the-fly!* For example, drag the “school” button in the “PivotTable Fields” list to the cell A1 (where it says “Drop Page Fields here”). Now we can pick from the list in cell B1 whether we want the table to include the houses of the Plum Ridge school district (value of school = 0), the Apple Valley school district (value = 1) or both. Choosing the first option, the PivotTable looks like the screen snapshot below. It seems almost as if we can display multiple pages of tables, for different values of the “school” variable. If we don’t need this option any longer, we can just drag the “school” button from cell A1 back to “PivotTable Fields” list.

The screenshot shows Microsoft Excel with a PivotTable and the PivotTable Field List. The PivotTable is located in the range B3:D8. The PivotTable Field List is on the right side of the window, showing the following fields: style, garage, basement, age, fire, price, and school. The PivotTable is currently set to show the count of houses per style, with the 'Total' field selected in the Values area.

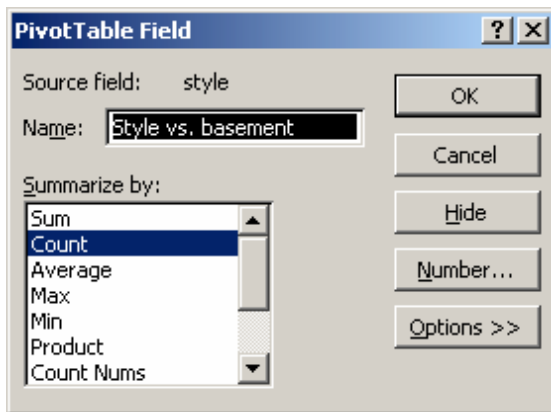
	style	Total
Count of style	0	24
style	0	24
style	1	11
style	2	8
Grand Total		43

Now back to our initial problem. We wanted to count the number of houses per style (in the rows) versus the presence or absence of a basement. How to accomplish this? The answer is just *too easy*: drag the “basement” button to B3! Your spreadsheet should then look like:

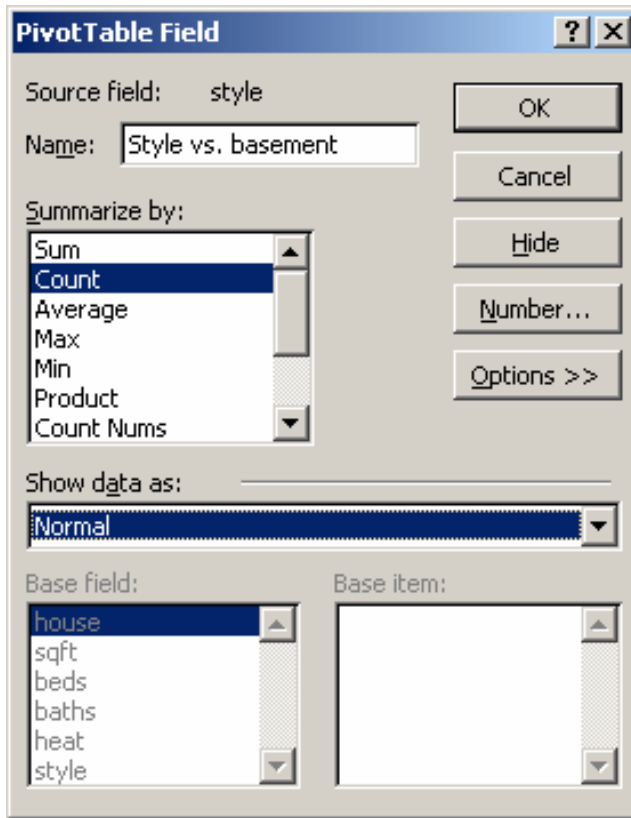




This is pretty much what we set out to do! Now, if we would like the data inside the table to be percentages (percent of grand total, percent of row total, percent of column total), we proceed as follows. Right-click inside the body of the table, and choose the "Field Settings" option from this list. The following dialog box appears:



Pressing the "Options >>" button in the lower right corner displays the expanded list:



When you pull down the “Show data as:” list, the options “% of row”, “% of column” and “% of total” appear. Choosing the “% of column” option then transforms our pivot table into the contingency table on the next page. This was the table we were after in the first place.

The screenshot shows Microsoft Excel with a PivotTable and the PivotTable Fields task pane. The PivotTable is located in the range B3:D8. The PivotTable Fields task pane is on the right, showing a list of fields: house, sqft, beds, baths, heat, and style. The PivotTable shows the following data:

style	0	1	Grand Total
0	82.35%	27.47%	36.11%
1	0.00%	27.47%	23.15%
2	17.65%	45.05%	40.74%
Grand Total	100.00%	100.00%	100.00%

The PivotTable Fields task pane shows the following fields:

- house
- sqft
- beds
- baths
- heat
- style

The PivotTable Fields task pane also has an "Add To" button and a "Row" dropdown menu.

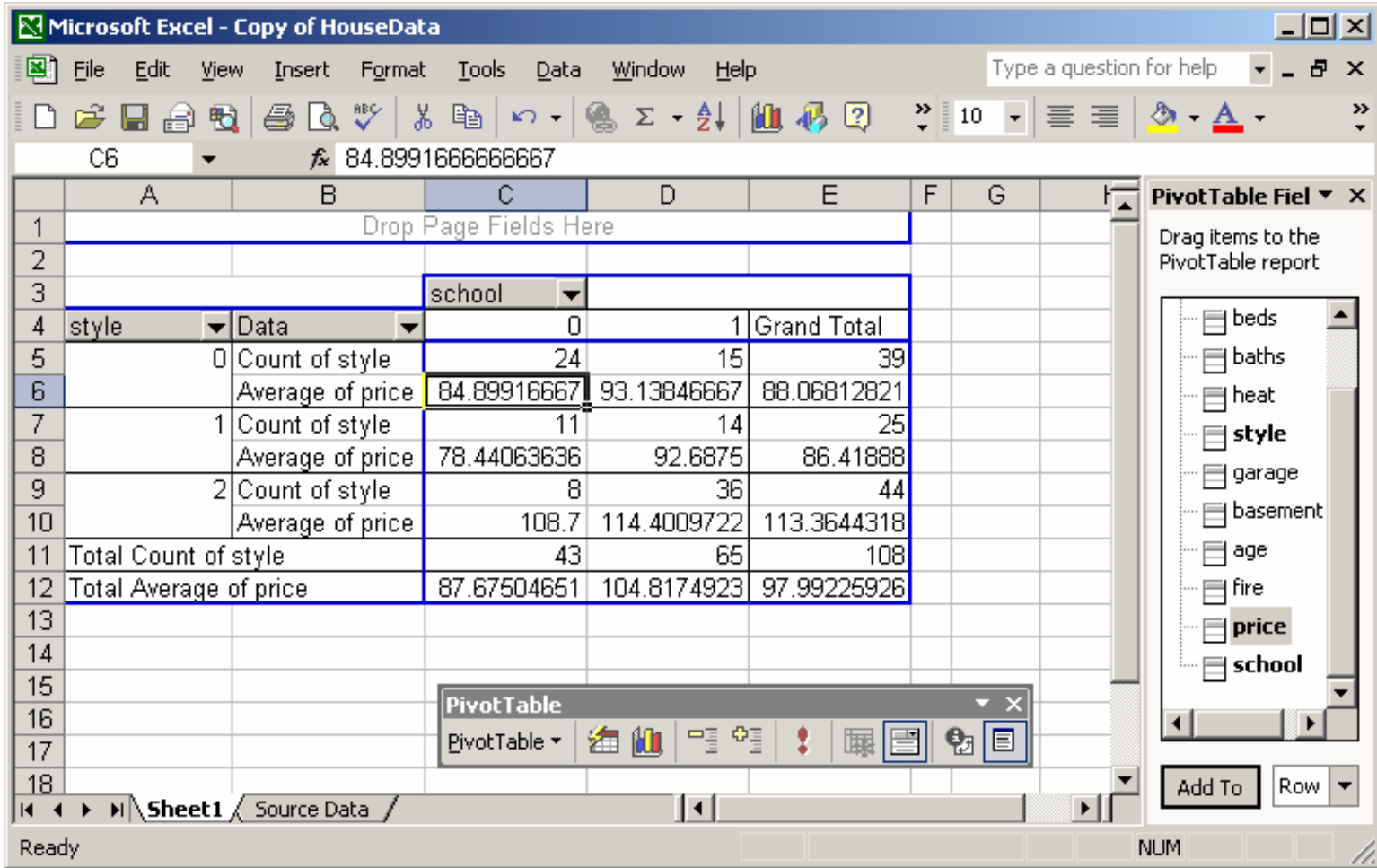
The reader should feel free to explore some more options and try new things. For example, note that we could even see the relationship between more than two variables. For example, we could see whether the presence of a basement may be explained not only by the style of the house, but by its location (i.e., the school district the house is located in) as well. I.e., we could divide the houses by school district *within* each house style. This is how we proceed. Drag the “school” button from “PivotTable Fields” list to column A, somewhere between row 4 and 8. Depending on how far you dragged it, your spreadsheet may look like:

The screenshot shows Microsoft Excel with a PivotTable and the PivotTable Fields task pane. The PivotTable is located in the range B4:E14. The PivotTable Fields task pane is on the right, showing a list of fields: beds, baths, heat, style, garage, basemer, age, fire, price, and school. The PivotTable is currently set to show the count of style, with 'style' in the Row Labels and 'school' in the Column Labels. The data is as follows:

style	school	0	1	Grand Total
0	0	11	13	24
0	1	3	12	15
0 Total		14	25	39
1	0		11	11
1	1		14	14
1 Total			25	25
2	0		8	8
2	1	3	33	36
2 Total		3	41	44
Grand Total		17	91	108

or the order may be “School”, followed by “Style” if you dragged “School” all the way to the left. Try dragging “School” or “Style” to the COLUMN section (row 3, columns C, D or E) and see what happens.

We can also do many more analyses that are related to statistics, but are extremely valuable for managers. Instead of just looking at the *frequencies*, we may be interested what the average sales *price* is as a function of *house style* and *school district*. This is how you do it. First, drag the field *price* from the “PivotTable Fields” to cell A3 (the upper left corner of the PivotTable). If the data displayed in the table isn’t the *average* price (but, e.g., the sum of all sales), then right-click in the DATA section (i.e., the “body”) of the table, and a pop-up menu will appear. From there, choose the *Field settings* option and you’ll again see the available computations for the field *price*. Choosing “Average” from that list, your final spreadsheet will look as follows:



There is now a new list labeled “Data” where we can select which cells we want displayed.

PivotTables are a very powerful tool to sift through a lot of data from different dimensions quickly, allowing you to compute different things for each dimension rather easily. Although cumbersome at first, PivotTables become easy after just a little bit of practice.