Chapter 5
Analysis of Qualitative Data

Contents

5.1 Introduction 145
5.2 Essentials of Qualitative Data Analysis 147
5.2.1 Dealing with Data Errors 147
5.3 PivotChart or PivotTable Reports 151
5.3.1 An Example 151
5.3.2 PivotTables 153
5.3.3 PivotCharts 164
5.4 TiendaMía.com Example: Question 1 168
5.5 TiendaMía.com Example: Question 2 169
5.6 Summary 175
Key Terms 175
Problems and Exercises 176

5.1 Introduction

There are many business questions that require the collection and analysis of *qualitative* data. For example, how does a visitor's opinion of a commercial website relate to her purchases at the website? Does a positive opinion of the website, relative to a bad or mediocre, lead to higher sales? This type of information is often gathered in the form of an opinion and measured as a categorical response. Also, accompanying these opinions are some quantitative characteristics of the respondent; for example, their age or income. Thus, a data collection effort will include various forms of qualitative and quantitative data elements (fields). Should we be concerned with the type of data we collect? In the prior chapters we have answered this question with a resounding *yes*. It is the type of the data—categorical, interval, ratio, etc.—that dictates the form of analysis we can perform.

In this chapter, we examine some of the many useful Excel resources available to analyze qualitative data. This includes exploring the uses of *PivotTable* and *PivotChart* reports: a built-in Excel capability that permits quick and easy *cross-tabulation* analysis, sometimes referred to as crosstab analysis. Crosstabs permit us

H. Guerrero, *Excel Data Analysis*, https://doi.org/10.1007/978-3-030-01279-3_5

to determine how two or more variables in a set of data interact. Consider the auto sales data we introduced in Chap. 4, which now appear in Table 5.1. There are many questions a decision-maker might consider in examining these data. For example, is there a relationship between sales associates and the models of automobiles they sell? More specifically, is there a propensity for some of the sales associates to promote the sale of a particular automobile to a particular customer demographic[1]?

Although this type of analysis can be performed with sophisticated statistics, in this chapter we will use less rigorous numerical techniques to generate valuable insights. The simple, numerical information that results, may be all that is necessary for good decision-making. Returning to our decision-maker's question, if we find that associates are concentrating on the sale of higher priced station wagons to a small number of demographics, a decision-maker may want to take steps to change

Table 5.1 Auto sales data example

Auto sales data—01/01/2005—01/31/2005							
Record No.	Slsprn	Date	Make	Model	Amt paid	Rebates	Sales Com
1	Bill	01/02/05	Ford	Wgn	$24,000	$2500	$2150
2	Henry	01/02/05	Toyota	Sdn	26,500	1000	2550
3	Harriet	01/03/05	Audi	Sdn	34,000	0	3400
4	Ahmad	01/06/05	Audi	Cpe	37,000	0	5550
5	Ahmad	01/06/05	Ford	Sdn	17,500	2000	2325
6	Henry	01/08/05	Toyota	Trk	24,500	1500	2300
7	Lupe	01/10/05	Ford	Wgn	23,000	2500	2050
8	Piego	01/12/05	Ford	Sdn	14,500	500	1400
9	Kenji	01/13/05	Toyota	Trk	27,000	1200	2580
10	Ahmad	01/14/05	Audi	Cpe	38,000	0	5700
11	Kenji	01/16/05	Toyota	Trk	28,500	1500	2700
12	Bill	01/16/05	Toyota	Sdn	23,000	2000	2100
13	Kenji	01/18/05	Ford	Wgn	21,500	1500	2000
14	Ahmad	01/19/05	Audi	Sdn	38,000	0	5700
15	Bill	01/19/05	Ford	Wgn	23,000	1000	2200
16	Kenji	01/21/05	Toyota	Trk	26,500	1500	2500
17	Lupe	01/24/05	Ford	Sdn	13,500	500	1300
18	Piego	01/25/05	Ford	Sdn	12,500	500	1200
19	Bill	01/26/05	Toyota	Trk	22,000	1000	2100
20	Ahmad	01/29/05	Audi	Cpe	36,500	0	5475
21	Bill	01/31/05	Ford	Sdn	12,500	500	1200
22	Piego	01/31/05	Ford	Sdn	13,000	500	1250

[1] In marketing, the term *demographic* implies the grouping or segmentation of customers into groups with similar age, gender, family size, income, professions, education, religious affiliation, race, ethnicity, national origin, etc. The choice of the characteristics to include in a demographic is up to the decision-maker.

this focused selling. It is possible that other demographics will be interested in similar vehicles if we apply appropriate sales incentives.

In Chap. 6 we will focus on *statistical* analysis that can be performed with techniques appropriate for qualitative and quantitative data. Among the techniques that we will be examine are Analysis of Variance (ANOVA), tests of hypothesis with t-tests and z-tests, and chi-square tests. These statistical tools will allow us to study the effect of independent variables on dependent variables contained in a data set, and allow us to study the similarity or dissimilarity of data samples. Although these technical terms may sound a bit daunting, I will establish clear rules for their application, and certainly clear enough to permit a non-statistician to apply the techniques correctly. Now, back to the techniques we will study in this chapter.

5.2 Essentials of Qualitative Data Analysis

In Chap. 4 we discussed the essential steps to prepare, organize, and present qualitative data. The preparation of qualitative data for *presentation* should also lead to preparation for data analysis; thus, most of the work done in presentation will complement the work necessary for the data analysis stage. Yet, there are a number of problems relating to **data errors** that can occur: problems in data collection, transcription, and entry. These errors must be dealt with early in the data analysis process, or the analysis will inevitably lead to inaccurate and inexplicable results.

5.2.1 Dealing with Data Errors

Data sets, especially large ones, can and usually contain errors. Some errors can be uncovered, but others are simply absorbed, never to be detected. Errors can occur due to a variety of reasons: problems with manual keying or electronic transmission of data onto spreadsheets or databases, mistakes in the initial recording of data by a respondent or data collector, and many other sources too numerous to list. Thus, steps insuring the quality of the data entry process need to be taken. As we saw in the previous chapter, where we assumed direct data entry in worksheets, we can devise data entry mechanisms to facilitate entry and to protect against entry errors.

Now, let us assume a data set that has been transcribed onto an Excel worksheet from an outside source. We will focus on the rigorous inspection of the data for unexpected entries. This can include a broad range of data inspection activities, from sophisticated sampling of a subset of data elements, to exhaustive (100%) inspection of all data. If a low level of errors can be tolerated, then only a sample of the data need be reviewed for accuracy. This is usually the case when a data set is very large, and the cost of errors is low relative to the cost of verification. If the cost of errors is high, then 100% inspection of the data may be necessary. For example, data collected in the clinical trial of a new drug may require 100% inspection due to the gravity of the acceptance or rejection of the trail results.

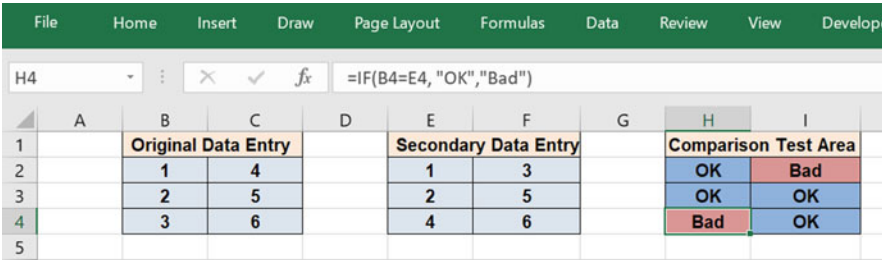
So, what capabilities does Excel provide to detect errors? In this section we examine a number of techniques for verification: (1) do two independent entries of similar data match, and (2) do data entries satisfy some range of characteristics? We will begin with a number of cell functions that permit the comparison of one data entry in a range to an entry in another range. Let us first assume a data collection effort for which data accuracy is of utmost importance. Thus, you employ two individuals to simultaneously key the data into two Excel worksheets. The entry is done independently, and the process is identical. Once the data is keyed into separate worksheets of a workbook, a third worksheet is used to compare each data entry. We will assume that if no differences are found in data entries, the data is without error. Of course, it is possible that two entries, though identical, can both be in error, but such a situation is likely to be a rare event.

This is an ideal opportunity to use the *logical IF* cell function to query whether a cell is identical to another cell. The *IF* function will be used to test the *equality* of entries in two cells, and return the cell value *OK* if the test results are equal, or *BAD* if the comparison is not equal. For simplicity's sake, assume that we have three ranges on a single worksheet where data is located—the first is data entry by Otto, the second is the identical data entry by Maribel, and the third is the test area to verify data *equality*. See Fig. 5.1 for the data entries and resulting test for equality.

Note Otto's data entry in cell B4 and the identical data element for Maribel in E4. The test function will appear in cell H4 as: *= IF (B4 = E4,“OK”,“BAD”)*. (Note that quotation marks must surround all text entries in Excel formulas.) The result of the **error checking** is a value of BAD since the entries are not equal (3< >4). The cell range of H2:I4 displays the results of all nine comparisons. The comparison determines that there are two disagreements in data entry. Of course, we are not in a position to suggest which entry is in error, but we are aware that the entries resulting in BAD must be investigated.

Regardless of the size of the data sets, the *IF* function can be written once and copied to a range equivalent in size and dimensions to the data entry ranges. Thus, if Otto's data entry occurs in range A1:H98, then the data entry range for Maribel could be A101:H198, and the *comparison area* could be in A201:H298. Our only restriction is that the dimension of the ranges containing data must be similar; that is, the number of rows and columns in the entry ranges must be the same.

A more direct approach to a comparison of data elements is the use of the Excel cell function **EXACT(text1, text2)**. As the title implies, the function compares two text data elements, and *if* an exact match is found it returns **TRUE** in the cell, *else* it



=IF(B4=E4, "OK","Bad")

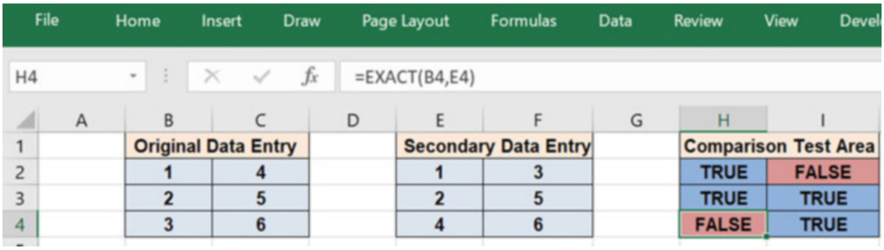
	A	B	C	D	E	F	G	H	I
1		Original Data Entry			Secondary Data Entry			Comparison Test Area	
2		1	4		1	3		OK	Bad
3		2	5		2	5		OK	OK
4		3	6		4	6		Bad	OK
5									

Fig. 5.1 If function error checking of data

returns **FALSE**. Figure 5.2 compares the six data items and performs error checking similar to that in Fig. 5.1. Note that the third data element of the first column and first of the second column are different (as before) and return a cell value result of *FALSE*. All other cells result in a value of *TRUE*.

It is also wise to test data for values outside the range of those that are anticipated. This is particularly true of numeric values. To perform statistical analysis, we sometimes convert qualitative variables (good, bad, male, female, etc.) to numeric values; thus, if the numeric values are incorrect, the analysis will also be incorrect. For example, it is easy to make a transcription error for data that must be converted from a *text* value (e.g. gender) to a *numeric* (male = 1, female = 2).

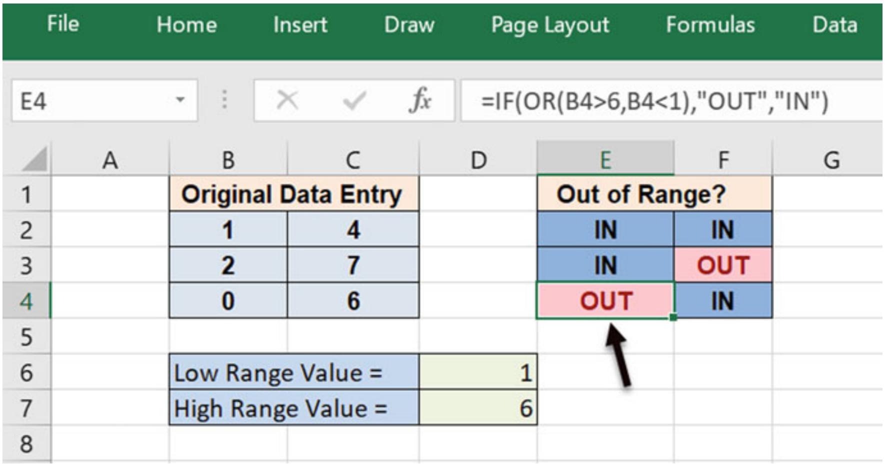
Consider the data table shown in Fig. 5.3. It consists of values that are anticipated to be in the range of 1–6. We can use the logical IF function to test the values occurring in the range of 1–6. But, rather than testing for each specific value (1, 2, 3,. . ., 6) by nesting multiple *IF* conditions and testing if the value is 1,2,. . .or 6, we can employ another logical function, the **OR**. It is used in Boolean logic, as are **AND**, **NOT**, **FALSE**, and **TRUE**. The combination of *IF* and *OR* can be used to test the data entries in a cell, say E4, by using the following logical conditions:



H4 =EXACT(B4,E4)

	A	B	C	D	E	F	G	H	I
1		Original Data Entry			Secondary Data Entry			Comparison Test Area	
2		1	4		1	3		TRUE	FALSE
3		2	5		2	5		TRUE	TRUE
4		3	6		4	6		FALSE	TRUE

Fig. 5.2 Exact function data entry comparison



E4 =IF(OR(B4>6,B4<1),"OUT","IN")

	A	B	C	D	E	F	G
1		Original Data Entry			Out of Range?		
2		1	4		IN	IN	
3		2	7		IN	OUT	
4		0	6		OUT	IN	
5							
6		Low Range Value =		1			
7		High Range Value =		6			
8							

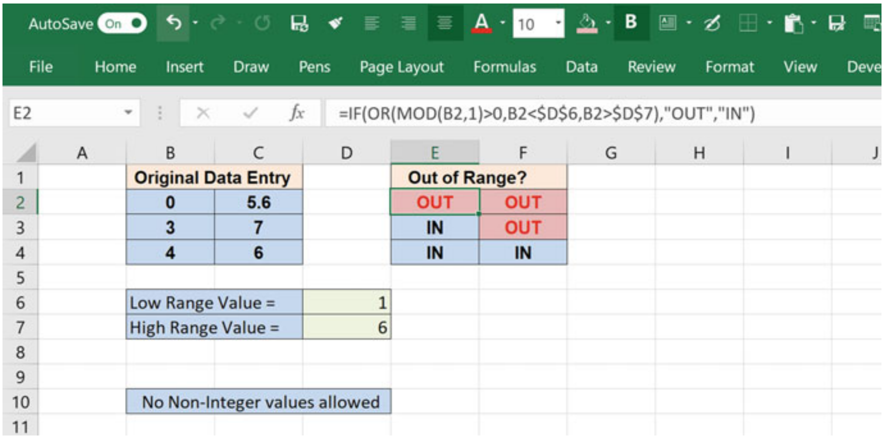
Fig. 5.3 Out of range data check

IF a value in cell B4 is less than 1 *OR* is greater than 6,
then return the text “*OUT*” to cell E4,
else return the text *“IN”* to cell E4.

Note that the results in Fig. 5.3 cells E4 and F3 are *OUT*, since the cell values B4 = 0 and C3 = 7 are outside the required range. Assuming the cell location B4 contains the data of interest in Fig. 5.3, the *IF* function used to perform the comparison in E4 is: = *IF (OR(B4 > 6, B4 < 1),“OUT”, “IN”)*. Of course, we could also replace the values 1 and 6 in the cell formula with cell references D6 and D7, respectively. This permits us to change the range of values if the need arises, without having to change cell formulas.

What happens when we are anticipating integer values from the data entry process and instead we encounter decimal values? The test above will not indicate that the value 5.6 is an incorrect entry. We can use another Excel Math and Trig cell function, the **MOD (number, divisor)** function, to logically test for a non-integer value. The *MOD* function returns only the remainder (also called the *modulus*) of the division of the *number* by the *divisor*—e.g. if the function argument *number* is 5.6 and the *divisor* is 1, the function will return the value 0.6. We can then include *MOD* as one of the tests in our *IF* function, just as we did with *OR*. It will test for integer values, while the other *OR* conditions test for values in the range of 1–6. The resulting function is now: = *IF(OR(MOD(B4,1) > 0,B4 > 6,B4 < 1),“OUT”, “IN”)*. The first *OR* condition, *MOD(B4,1) > 0*, divides B4 by 1, and returns the remainder. If that remainder is not 0, then the condition is *TRUE* (there is a remainder) and the text message *OUT* is returned. If the condition is *FALSE* (no remainder), then the next condition, B4 > 6, is examined. If the condition is found to be *TRUE*, then *OUT* is returned, and so on. Note that the *MOD* function simply becomes one of the three arguments for the *OR*. Thus, we have constructed a relatively complex *IF* function for error checking.

Figure 5.4 shows the results of both tests. As you can see, this is a convenient way to determine if a non-integer value is in our data, and the values 0 in B2, 5.6 in C2,



E2 =IF(OR(MOD(B2,1)>0,B2<$D$6,B2>$D$7),"OUT","IN")

	A	B	C	D	E	F	G	H	I	J
1		Original Data Entry			Out of Range?					
2		0	5.6		OUT	OUT				
3		3	7		IN	OUT				
4		4	6		IN	IN				
5										
6		Low Range Value =		1						
7		High Range Value =		6						
8										
9										
10		No Non-Integer values allowed								
11										

Fig. 5.4 Out of range and non-integer data error check

and 7 in C3 are identified as either out of range or non-integer. The value 0, satisfies both conditions, but is only detected by the out of range condition. This application shows the versatility of the *IF* function, as well as other related logical functions, such as *OR*, *AND*, *NOT*, *TRUE*, and *FALSE*. Any logical test that can be conceived can be handled by some combination of these logical functions.

Once we have verified the accuracy of our data, we are ready to perform several types of descriptive analyses, including **cross-tabulation** using Excel's ***PivotChart*** and ***PivotTable*** tools. *PivotChart*s and *PivotTable*s are frequently used tools for analyzing qualitative data—e.g. data contained in customer surveys, operations reports, opinion polls, etc. They are also used as exploratory techniques to guide us to more sophisticated types of analyses.

5.3 PivotChart or PivotTable Reports

Cross-tabulation provides a methodology for observing the interaction of several variables in a set of data. For example, consider an opinion survey that records demographic and financial variables for respondents. Among the variables recorded is age, which is organized into several mutually exclusive age categories (18–25, 26–34, 35–46, and 47 and older). Respondents are also queried for a response or opinion, *good* or *bad*, about some consumer product. Cross-tabulation permits the analyst to determine, for example, the number of respondents in the 35–46 age category that report the product as *good*. The analysis can also determine the number of respondents that fit the conditions (age and response) as a percentage of the total.

The *PivotTable* and *PivotChart* report function is found in the *Tables* group in the *Insert Ribbon*. Both reports are identical, except that the table provides numerical data in a table format, while the chart converts the numerical data into a graphical format. The best way to proceed with a discussion of *PivotTable* and *PivotChart* is to begin with an illustrative problem, one that will allow us to exercise many of the capabilities of these powerful functions.

5.3.1 An Example

Consider an example, a consumer survey, that will demonstrate the uses of *PivotTable*s and *PivotChart*s. The data for the example is shown in Table 5.2. A web-based business, TiendaMía.com[2] is interested in testing various web designs that customers will use to order products. The owners of TiendaMía.com hire a marketing firm to help them conduct a preliminary survey of 30 randomly selected customers to determine their preferences. Each of the customers is given a gift

[2] *TiendaMía* in Spanish translates to *My Store* in English

Table 5.2 Survey opinions on four webpage designs

Category					Opinion			
Case	Gender	Age	Income	Region	Product 1	Product 2	Product 3	Product 4
1	M	19	2500	East	Good	Good	Good	Bad
2	M	25	21,500	East	Good	Good	Bad	Bad
3	F	65	13,000	West	Good	Good	Good	Bad
4	M	43	64,500	North	Good	Good	Bad	Bad
5	F	20	14,500	East	Bad	Good	Bad	Good
6	F	41	35,000	North	Bad	Good	Bad	Bad
7	F	77	12,500	South	Good	Bad	Bad	Bad
8	M	54	123,000	South	Bad	Bad	Bad	Bad
9	F	31	43,500	South	Good	Good	Bad	Bad
10	M	37	48,000	East	Bad	Good	Good	Bad
11	M	41	51,500	West	Good	Good	Bad	Bad
12	F	29	26,500	West	Bad	Good	Bad	Bad
13	F	19	55,000,000	Outer space	Bad	Bad	Bad	Bad
14	F	32	41,000	North	Good	Bad	Good	Bad
15	M	45	76,500	East	Good	Bad	Good	Good
16	M	49	138,000	East	Bad	Bad	Bad	Bad
17	F	36	47,500	West	Bad	Bad	Bad	Bad
18	F	64	49,500	South	Bad	Good	Bad	Bad
19	M	26	35,000	North	Good	Good	Good	Bad
20	M	28	29,000	North	Good	Bad	Good	Bad
21	M	27	25,500	North	Good	Good	Good	Bad
22	M	54	103,000	South	Good	Bad	Good	Good
23	M	59	72,000	West	Good	Good	Good	Bad
24	F	30	39,500	West	Good	Bad	Good	Good
25	F	62	24,500	East	Good	Bad	Bad	Bad
26	M	62	36,000	East	Good	Bad	Bad	Good
27	M	37	94,000	North	Bad	Bad	Bad	Bad
28	F	71	23,500	South	Bad	Bad	Good	Bad
29	F	69	234,500	South	Bad	Bad	Bad	Bad
30	F	18	1500	East	Good	Good	Good	Bad

coupon to participate in the survey and is instructed to visit a website for a measured amount of time. The customers are then introduced to four web-page designs and asked to respond to a series of questions. The data are self-reported by the customers on the website, as they experience the four different webpage designs. The marketing firm has attempted to control each step of the survey to eliminate extraneous influences on the respondents. Although this is a simple example, it is typical of consumer opinion surveys and website tests.

The data collected from 30 respondents regarding their gender, age, income, and the region of the country where they live is shown in Table 5.2. Each respondent, often referred to as a *case*, has his data recorded in a row. Respondents have provided an *Opinion* on each of the four products in one section of the data and their demographic characteristics, *Category*, in another. As is often the case with data, there may be some data elements that are either out of range or simply ridiculous responses; for example, respondent number 13 claims to be a 19-year-old female that has an income of $55,000,000 and resides in outer space. This is one of the pitfalls of survey data: it is not unusual to receive information that is unreliable. In this case, it is relatively easy to see that our respondent is not providing information that we can accept as true. My position, and that of most analysts, on this respondent is to remove the record, or case, completely from the survey. In other cases, it will not be so easy to detect errors or unreliable information, but the validation techniques we developed in Chap. 4 might help identify such cases.

Now, let's consider a few questions that might be of business interest to the owners of TiendaMía:

1. Is there a webpage design that dominates others in terms of positive customer response? Of course, the opposite information might also be of interest. Finding a single dominant design could greatly simply the decision-makers efforts in choosing a new webpage design.
2. It is unlikely that we will find a single design that dominates. But, we can consider the various demographic and financial characteristics of the respondents, and how the characteristics relate to their webpage design preferences; that is, is there a particular demographic group(s) that responded with generally positive, negative, or neutral preferences to the particular webpage designs?

These questions cover a multitude of important issues that TiendaMía will want to explore. Let us assume that we have exercised most of the procedures for ensuring data accuracy discussed earlier, but we still have some **data scrubbing**[3] that needs to be done. We surely will eliminate respondent 13 in Table 5.2 who claims to be from outer space; thus, we now have data for 29 respondents to analyze.

As mentioned before, *PivotTable* and *PivotChart* Report tools can be found in the *Tables Group* of the *Insert Ribbon*. As is the case with other tools, the *PivotTable* and *PivotChart* have a wizard that guides the user in the design of the report. Before we begin to exercise the tool, I will describe the basic elements of a *PivotTable*. It is *best* practice to begin with the creation of a PivotTable, and then move to a *PivotChart* for visualization, although it is possible to create them simultaneously.

5.3.2 PivotTables

A *PivotTable* organizes large quantities of data into a 2-dimensional format. For a set of data, the combination of 2 dimensions will have an *intersection*. Recall that we are

[3] The term *scrubbing* refers to the process of removing or changing data elements that are contaminated or incorrect, or that are in the wrong format for analysis.

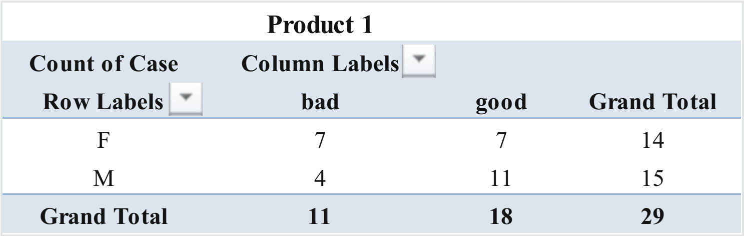
interested in the respondents that satisfy some set of conditions: a position or place within two dimensions. For example, in our survey data the dimension Gender (male or female) and Product 1 (good or bad) can intersect in how the two categories of gender rate their preference for Product 1, a *count* of either *good* or *bad*. Thus, we can identify all females that choose *good* as an opinion for the webpage design Product 1; or similarly, all males that choose *bad* as an opinion for Product 1. Table 5.3 shows the cross-tabulation of *Gender* and preference for *Product 1*. The table accounts for all 29 respondents, with the 29 respondents distributed into the four mutually exclusive and collectively exhaustive categories—7 in Female/ Bad, 7 in Female/Good, 4 in Male/Bad, and 11 in Male/Good. The categories are **mutually exclusive** in that a respondent can only belong to one of the 4 Gender/ Opinion categories; they are **collectively exhaustive** in that the four Gender/Opinion categories contain all possible respondents. Note we could also construct a similar cross-tabulation for each of the three remaining webpage designs (Product 2, 3, 4) by examining the data and *counting* the respondents that meet the conditions in each cell. Obviously, this could be a tedious chore, especially if a large data set is involved. That is why we depend on *PivotTables* and *PivotCharts*: they automate the process of creating cross-tabulations and we dictate the structure.

In Excel, the internal area of the cross-tabulation table is referred to as the **data area**, and the data elements captured within the data area represent a **count** of respondents (7, 7, 4, 11). The dimensions (see Table 5.3) are referred to as the *row* and *column*. On the margins of the table we can also see the totals for the various values of each dimension. For example, the *data area* contains 11 total *bad* respondent preferences and 18 *good*, regardless of the gender of the respondent. Also, there are 14 total females and 15 males, regardless of their preferences.

The data area and marginal dimensions are selected by the user and can range from *count*, *sum*, *average*, *min*, *max*, etc. In the *data area* we currently display a *count* of respondents, but there are other values we could use; for example, the respondents *average* age or the *sum* of income for respondents meeting the *row* and *column* criteria. There are many other values that could be selected. We will provide more detail on this topic later in the chapter.

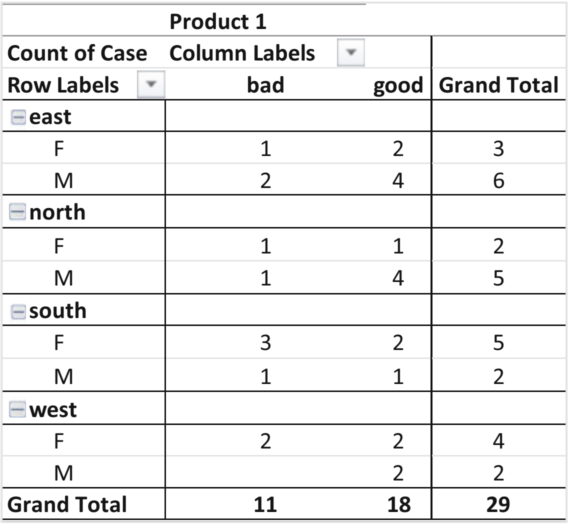
We can expand the number of data elements along one of the dimensions, *row* or *column*, to provide a more detailed and layered view of the data. Previously, we had

Table 5.3 Cross-tabulation of gender and product 1 preference in terms of respondent count



	Product 1		
Count of Case	Column Labels ▾		
Row Labels ▾	bad	good	Grand Total
F	7	7	14
M	4	11	15
Grand Total	**11**	**18**	**29**

Table 5.4 Cross-tabulation of gender/region and product 1 preference in terms of respondent count



	Product 1		
Count of Case	Column Labels ▾		
Row Labels ▾	bad	good	Grand Total
⊟east			
F	1	2	3
M	2	4	6
⊟north			
F	1	1	2
M	1	4	5
⊟south			
F	3	2	5
M	1	1	2
⊟west			
F	2	2	4
M		2	2
Grand Total	**11**	**18**	**29**

only *Gender* on the *row* dimension. Consider a new combination of *Region* and *Gender*. *Region* has 4 associated categories and *Gender* has 2; thus, we will have 8 (4 × 2) rows of data plus subtotal rows, if we choose. Table 5.4 shows the expanded and more detailed cross-tabulation of data. This new breakdown provides detail for *Male* and *Female* by region. For example, there are 3 females and 6 males in the East region. There is no reason why we could not continue to add other dimensions, either to the *row* or *column*, but from a practical point of view, adding more dimensions can lead to visual overload of information. Therefore, we are careful to consider the confusion that might result from the indiscriminate addition to dimensions. In general, two characteristics on a row or column are usually a maximum for easily understood presentations.

You can see that adding dimensions to either the row or column results in a data reorganization and different presentation of Table 5.2; that is, rather than organizing based on all respondents (observations), we organize based on the specific categories of the dimensions that are selected. All our original data remains intact. If we add all the counts found in the totals column for females in Table 5.4, we still have a count of 14 females (marginal column totals. . .3 + 2 + 5 + 4 = 14). By not including a dimension, such as Age, we ignore age differences in the data. The same is true for Income. More precisely, we are not ignoring Age or Income, but we are simply not concerned with distinguishing between the various categories of these two dimensions.

So, what are the preliminary results of the cross-tabulation that is performed in Tables 5.3 and 5.4? Overall there appears to be more *good* than *bad* evaluations of Product 1, with 11 *bad* and 18 *good*. This is an indication of the relative strength of the product, but if we dig a bit more deeply into the data and consider the gender preferences in each region, we can see that females are far less enthusiastic about Product 1, with 7 bad and 7 good. Males on the other hand, seem far more enthusiastic, with 4 *bad* and 11 *good*. The information that is available in Table 5.4 also permits us to see the regional differences. If we consider the South region, we see that both males and females have a mixed view of Product 1, although the number of respondents in the South is relatively small.

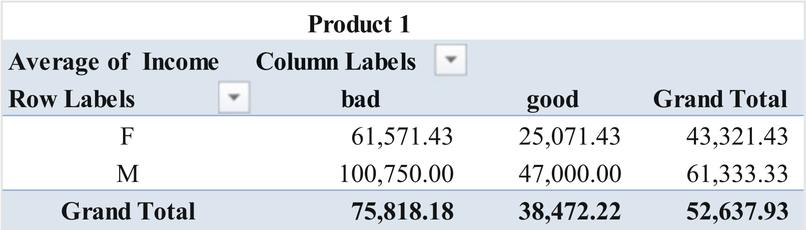
Thus far, we have only considered count for the data field of the cross-tabulation table; that is, we have counted the respondents that fall into the various intersections of categories—e.g. two Female observations in the West have a *bad* opinion of Product 1. There are many other alternatives for how we can present data, depending on our goal for the analysis. Suppose that rather than using a *count* of respondents, we decide to present the *average* income of respondents for each cell of our data area. Other options for the income data could include the sum, min, max, or standard deviation of the respondent's income in each cell. Additionally, we can calculate the percentage represented by a count in a cell relative to the total respondents. There are many interesting and useful possibilities available.

Consider the cross-tabulation in Table 5.3 that was presented for respondent counts. If we replace the respondent count with the average of their Income, the data will change to that shown in Table 5.5. The value is $100,750 for the combination of Male/Bad in the cross-tabulation table. This is the average[4] of the four respondents found in Table 5.2: #8–$123,000, #10–$48,000, #16–$138,000, and #27–$94,000. TiendaMía.com might be very concerned that these males with an average of substantial spending power do not have a good opinion of Product 1.

Now, let us turn to the *PivotTable* and *PivotChart* tool in Excel to produce the results we have seen in Tables 5.3, 5.4, and 5.5. The steps for constructing the tables follow:

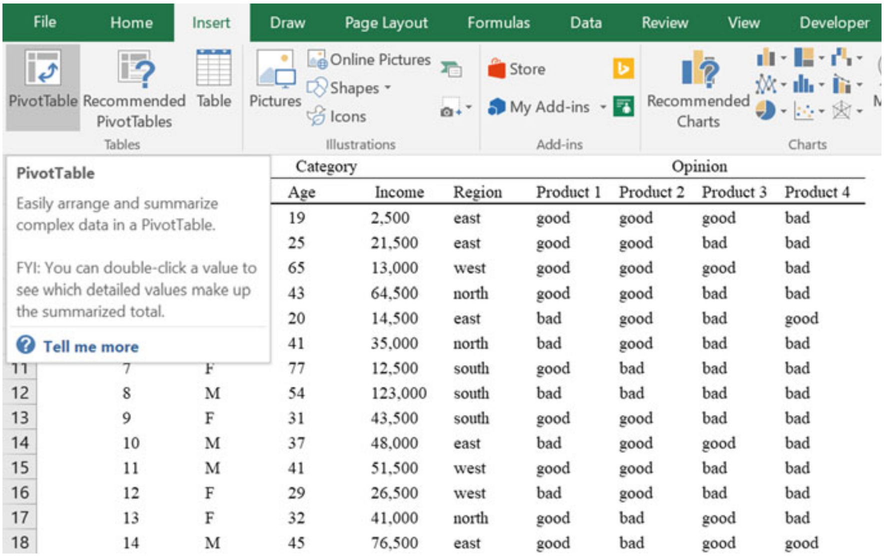
1. Figure 5.5 shows the *Tables Group* found in the *Insert Ribbon*. In this step, you can choose between a *PivotTable* and a *PivotChart*. Excel makes the selection of

Table 5.5 Cross-tabulation of gender and product 1 preference in terms of average income



	Product 1		
Average of Income	**Column Labels** ▾		
Row Labels ▾	**bad**	**good**	**Grand Total**
F	61,571.43	25,071.43	43,321.43
M	100,750.00	47,000.00	61,333.33
Grand Total	**75,818.18**	**38,472.22**	**52,637.93**

[4] (123,000 + 48,000 + 138,000 + 94,000)/4 = $100,750.



PivotTable

Easily arrange and summarize complex data in a PivotTable.

FYI: You can double-click a value to see which detailed values make up the summarized total.

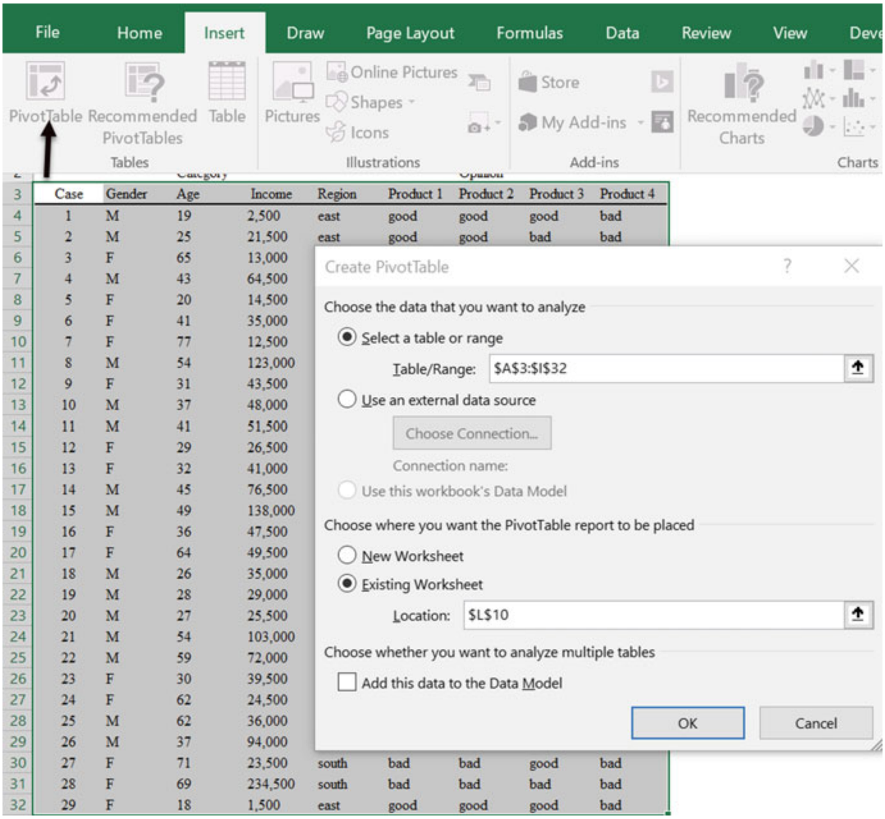
Tell me more

			Category			Opinion			
			Age	Income	Region	Product 1	Product 2	Product 3	Product 4
			19	2,500	east	good	good	good	bad
			25	21,500	east	good	good	bad	bad
			65	13,000	west	good	good	good	bad
			43	64,500	north	good	good	bad	bad
			20	14,500	east	bad	good	bad	good
			41	35,000	north	bad	good	bad	bad
11	7	F	77	12,500	south	good	bad	bad	bad
12	8	M	54	123,000	south	bad	bad	bad	bad
13	9	F	31	43,500	south	good	good	bad	bad
14	10	M	37	48,000	east	bad	good	good	bad
15	11	M	41	51,500	west	good	good	bad	bad
16	12	F	29	26,500	west	bad	good	bad	bad
17	13	F	32	41,000	north	good	bad	good	bad
18	14	M	45	76,500	east	good	bad	good	good

Fig. 5.5 Insert *PivotTable* command

PivotTable the default, although a *PivotChart* can be selected with a few more keystrokes. As noted earlier, they contain the same information.

2. Next, and shown in Fig. 5.6, the wizard opens a dialogue box that asks you to identify the data range you will use in the analysis—in our case $A$3:$I$32. Note that I have included the titles (dimension labels such as *Gender*, *Age*, etc.) of the fields just as we did in the data sorting and filtering process. This permits a title for each data field—*Case*, *Gender*, *Income*, etc. The dialogue box also asks where you would like to locate the *PivotTable*. We choose to locate the table in the same sheet as the data, cell $L$10, but you can also select a new worksheet.
3. A convenient form of display will enable the drop-and-drag capability for the table. Once the table appears, or you have populated it with row and column fields, right click and select *Pivot Table Options* from the pull-down menu. Under the *Display* Tab, select *Classic Pivot Table Layout*. See Fig. 5.7.
4. Figure 5.7 shows the general layout of the *PivotTable*. Note that there are four fields that form the table and that require an input—**Filters**, **Column**, **Row**, and **Values**. Except for the *Filters*, the layout of the cross-tabulation table is similar to our previous Tables 5.3, 5.4, and 5.5. On the right (*PivotTable Fields*), you see six (nine if you include Product 2–4) buttons that represent the dimensions that we identified earlier as the titles of columns in our data table. You can drag-and-drop the buttons into the four fields/regions shown below—*Filters*, *Columns*, *Rows*, and *Values*. Figure 5.8 shows the *Row* area populated with *Gender*. Of the four fields, the *Filters* field is generally the field that is not populated, at least initially. Note that the *Filters* field has the effect of providing a third dimension for the *PivotTable*.



	Case	Gender	Age	Income	Region	Product 1	Product 2	Product 3	Product 4
4	1	M	19	2,500	east	good	good	good	bad
5	2	M	25	21,500	east	good	good	bad	bad
6	3	F	65	13,000					
7	4	M	43	64,500					
8	5	F	20	14,500					
9	6	F	41	35,000					
10	7	F	77	12,500					
11	8	M	54	123,000					
12	9	F	31	43,500					
13	10	M	37	48,000					
14	11	M	41	51,500					
15	12	F	29	26,500					
16	13	F	32	41,000					
17	14	M	45	76,500					
18	15	M	49	138,000					
19	16	F	36	47,500					
20	17	F	64	49,500					
21	18	M	26	35,000					
22	19	M	28	29,000					
23	20	M	27	25,500					
24	21	M	54	103,000					
25	22	M	59	72,000					
26	23	F	30	39,500					
27	24	F	62	24,500					
28	25	M	62	36,000					
29	26	M	37	94,000					
30	27	F	71	23,500	south	bad	bad	good	bad
31	28	F	69	234,500	south	bad	bad	bad	bad
32	29	F	18	1,500	east	good	good	good	bad

Create PivotTable ? ×

Choose the data that you want to analyze

● Select a table or range

Table/Range: $A$3:$I$32

○ Use an external data source

Choose Connection...

Connection name:

○ Use this workbook's Data Model

Choose where you want the PivotTable report to be placed

○ New Worksheet

● Existing Worksheet

Location: $L$10

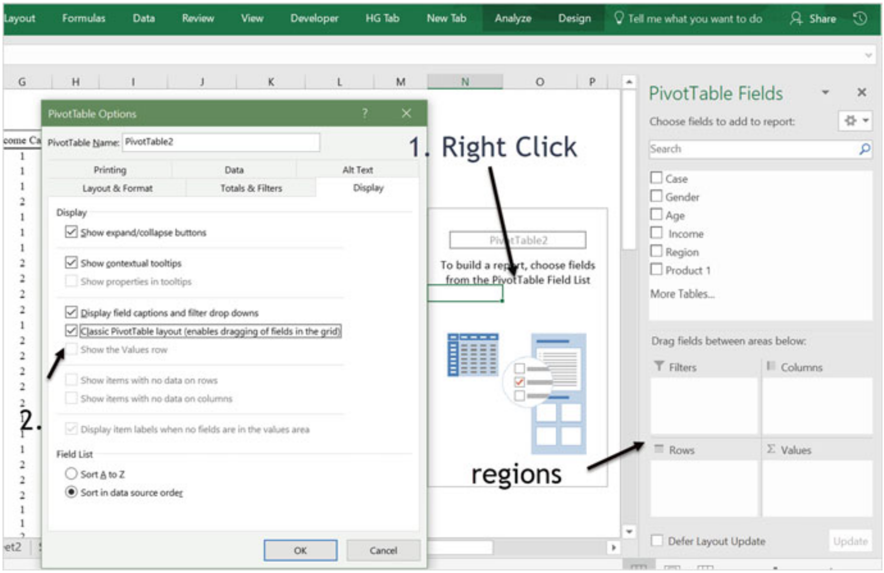
Choose whether you want to analyze multiple tables

□ Add this data to the Data Model

OK Cancel

Fig. 5.6 Create *PivotTable*

5. Figure 5.9 shows the constructed table. I have selected *Gender* as the row, *Product 1* as the column, *Region* as a filter, and *Case* as the values fields. Additionally, I have selected *count* as the measure for *Case* in the data field. By selecting the pull-down menu for the *Values* field (see Fig. 5.10) or by right clicking a value in the table you can change the measure to one of many possibilities—**Sum**, **Count**, **Average**, **Min**, **Max**, etc. Even greater flexibility is provided in the *Show Values As* menu tab. For example, the *Value Field Settings* dialogue box allows you to select additional characteristics of how the count will be presented—e.g. as a *% of Grand Total*. See Fig. 5.11.
6. In Fig. 5.12 you can see one of the extremely valuable features of *PivotTables*. A pull-down menu is available for the *Filters*, *Row*, and *Column* fields. These correspond to *Region*, *Gender*, and *Product 1*, respectively. These menus will allow you to change the data views by limiting the categories within dimensions, and to do so without reengaging the wizard. For example, currently *all Region* categories (East, West, etc.) are included in the table shown in Fig. 5.12, but you can use the pull-down menu to limit the regions to *East* only. Figure 5.13 shows



PivotTable Options

PivotTable Name: PivotTable2

Printing | Data | Alt Text
Layout & Format | Totals & Filters | Display

Display
- ☑ Show expand/collapse buttons
- ☑ Show contextual tooltips
- □ Show properties in tooltips
- ☑ Display field captions and filter drop downs
- ☑ Classic PivotTable layout (enables dragging of fields in the grid)
- □ Show the Values row
- □ Show items with no data on rows
- □ Show items with no data on columns
- ☑ Display item labels when no fields are in the values area

Field List
- ○ Sort A to Z
- ● Sort in data source order

OK | Cancel

1. Right Click

2.

regions

PivotTable Fields

Choose fields to add to report:

Search

- □ Case
- □ Gender
- □ Age
- □ Income
- □ Region
- □ Product 1

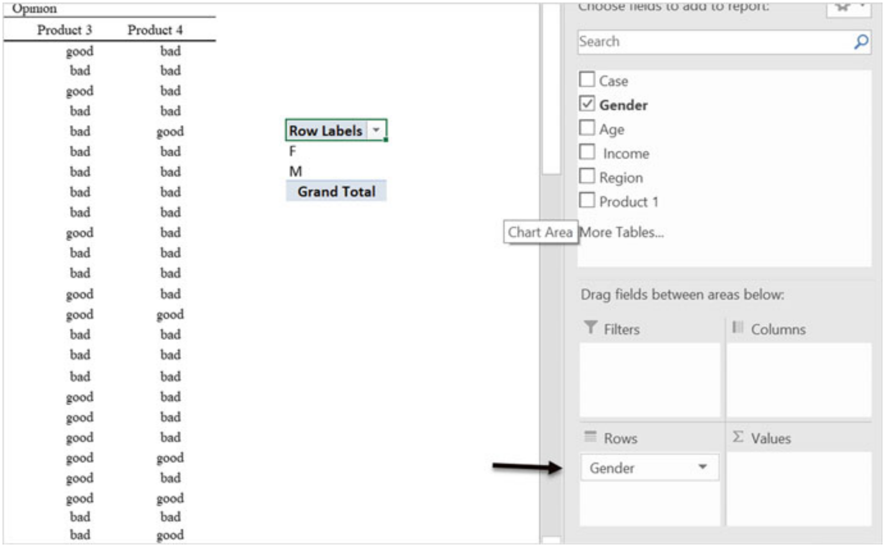
More Tables...

Drag fields between areas below:

Filters | Columns | Rows | Values

□ Defer Layout Update | Update

Fig. 5.7 *PivotTable* field entry



Opinion	
Product 3	Product 4
good	bad
bad	bad
good	bad
bad	bad
bad	good
bad	bad
bad	bad
bad	bad
bad	bad
good	bad
bad	bad
bad	bad
good	bad
good	good
bad	bad
bad	bad
bad	bad
good	bad
good	bad
good	bad
good	good
good	bad
good	good
bad	bad
bad	good

Row Labels
F
M
Grand Total

Choose fields to add to report:

Search

- □ Case
- ☑ Gender
- □ Age
- □ Income
- □ Region
- □ Product 1

More Tables...

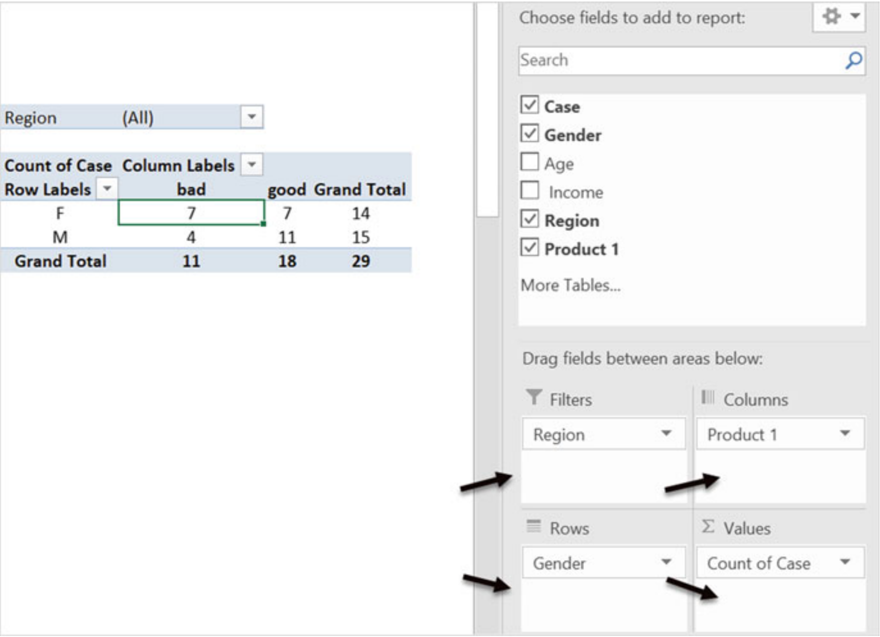
Drag fields between areas below:

Filters | Columns

Rows: Gender | Values

Fig. 5.8 Populating *PivotTable* fields

the results and the valuable nature of this built-in capability. Note that the number of respondents for the *PivotTable* for the *East* region results in only 9 respondents, whereas, the number of respondents for all regions is 29, as seen in Fig. 5.12. Also, note the appearance of a funnel in the pull-down menu indicating filtered data. Combining this capability with the changes we can perform for the *Values*



Region (All)

Count of Case	Column Labels		
Row Labels	bad	good	Grand Total
F	7	7	14
M	4	11	15
Grand Total	11	18	29

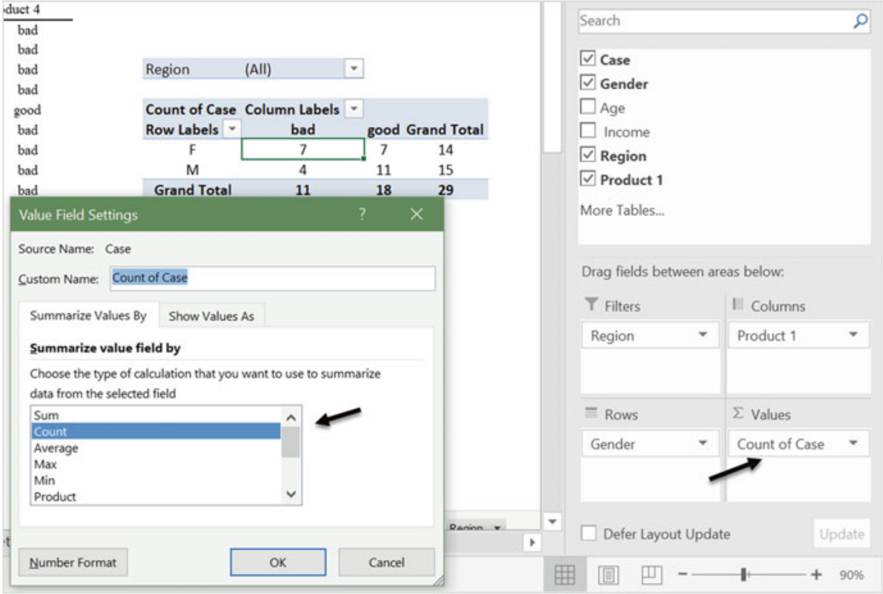
Choose fields to add to report:

- ☑ Case
- ☑ Gender
- □ Age
- □ Income
- ☑ Region
- ☑ Product 1

Drag fields between areas below:

- Filters: Region
- Columns: Product 1
- Rows: Gender
- Values: Count of Case

Fig. 5.9 Complete *PivotTable* layout



duct 4
bad
bad
bad
bad
good
bad
bad
bad
bad

Region (All)

Count of Case	Column Labels		
Row Labels	bad	good	Grand Total
F	7	7	14
M	4	11	15
Grand Total	11	18	29

Value Field Settings

Source Name: Case

Custom Name: Count of Case

Summarize value field by

Choose the type of calculation that you want to use to summarize data from the selected field

- Sum
- Count
- Average
- Max
- Min
- Product

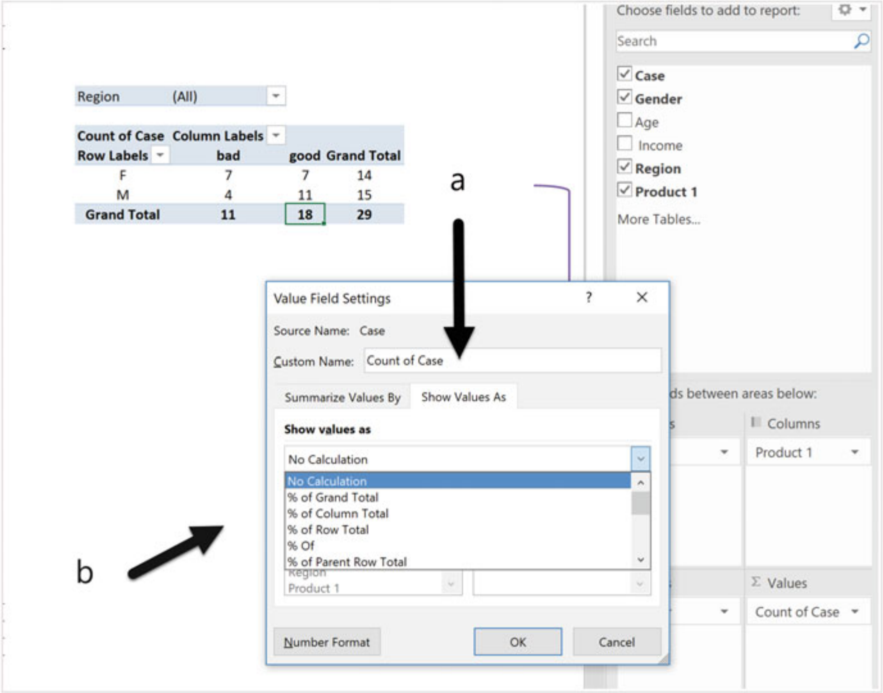
- ☑ Case
- ☑ Gender
- □ Age
- □ Income
- ☑ Region
- ☑ Product 1

Drag fields between areas below:

- Filters: Region
- Columns: Product 1
- Rows: Gender
- Values: Count of Case

□ Defer Layout Update

Fig. 5.10 Case count summary selection



Region (All)

Count of Case	Column Labels		
Row Labels	bad	good	Grand Total
F	7	7	14
M	4	11	15
Grand Total	11	18	29

Value Field Settings

Source Name: Case

Custom Name: Count of Case

Show values as: No Calculation

No Calculation
% of Grand Total
% of Column Total
% of Row Total
% Of
% of Parent Row Total

Choose fields to add to report:

☑ Case
☑ Gender
□ Age
□ Income
☑ Region
☑ Product 1

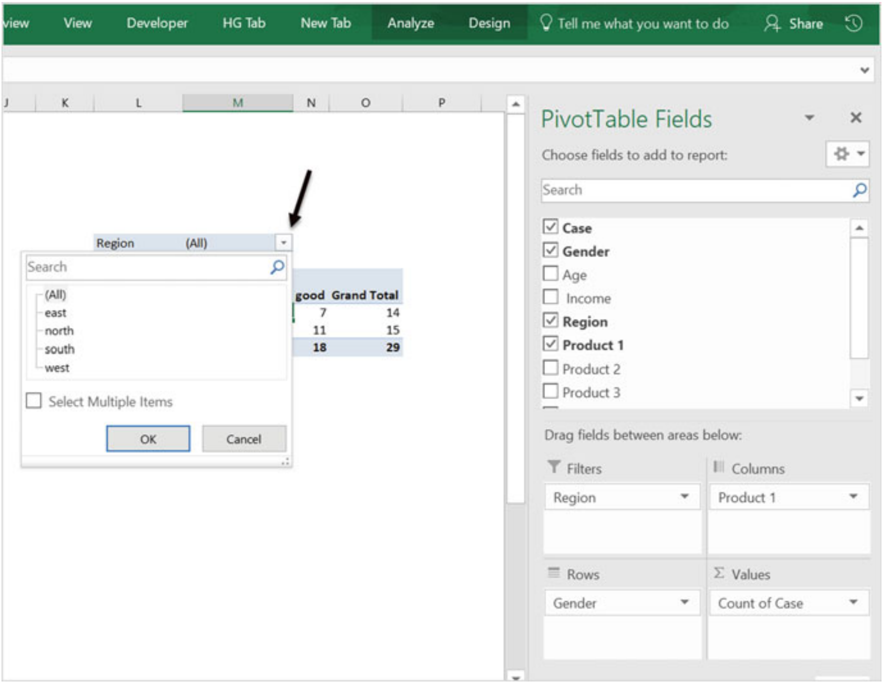
Columns: Product 1

Values: Count of Case

Fig. 5.11 Case count summary selection as % of total count

field, we can literally view the data in our table from a nearly unlimited number of perspectives.

7. Figure 5.14 demonstrates the change proposed in (5) above. We have changed the *Value Field Settings* from *count* to *% of Grand Total*, just one of the dozens of views that is available. The choice of views will depend on your goals for the data presentation.
8. We can also extend the chart quite easily to include *Region* in the *Row* field. This is accomplished by first pointing to any cell in the table and right clicking to show *Field List*. Once available, you can drag and drop the *Region* button into the *Row Label*. This action adds a new layer to the row dimension. The converted table will resemble Table 5.4. This action also provides subtotals for the various regions and gender combinations—*East-M*, *West-F*, etc. Figure 5.15 shows the results of the extension. Note that I have also selected Age as the new *Filter*; thus, we could filter for example to find records that consist of ages over 40 years.
9. One final feature of special note is the capability to identify the specific respondents' records that are in the cells of the data field. Simply double-click the cell of interest in the Pivot Table, and a separate sheet is created with the records of interest. This allows you to *drill-down* into the records of the specific respondents



Region (All)

Search

(All)
east
north
south
west

□ Select Multiple Items

OK Cancel

good	Grand Total
7	14
11	15
18	29

PivotTable Fields

Choose fields to add to report:

Search

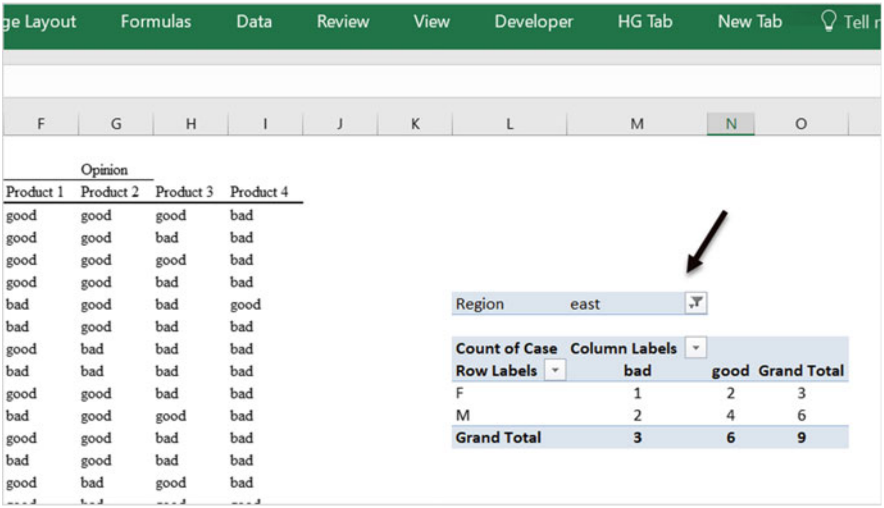
☑ Case
☑ Gender
□ Age
□ Income
☑ Region
☑ Product 1
□ Product 2
□ Product 3

Drag fields between areas below:

Filters	Columns
Region	Product 1

Rows	Values
Gender	Count of Case

Fig. 5.12 *PivotTable* drop-down menus

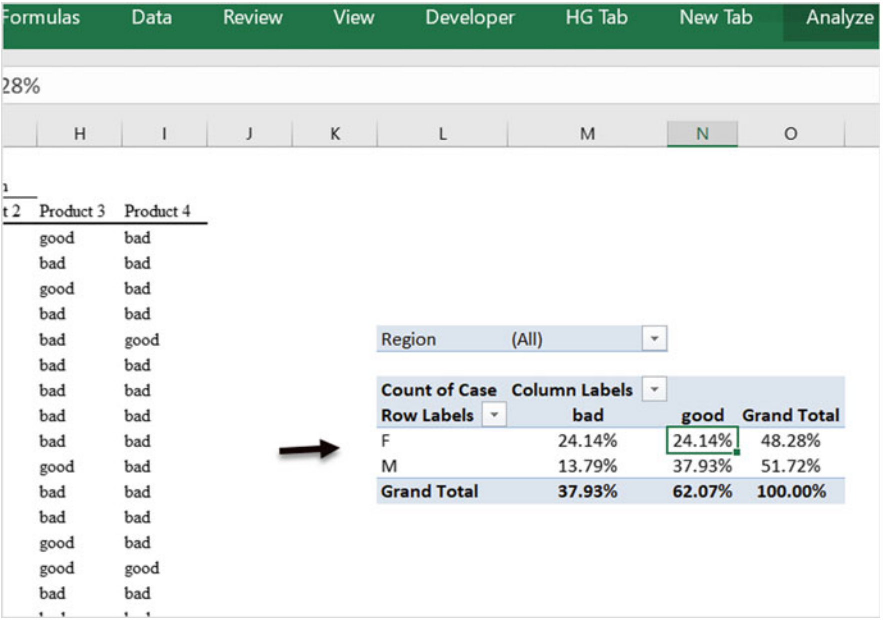


	Opinion		
Product 1	Product 2	Product 3	Product 4
good	good	good	bad
good	good	bad	bad
good	good	good	bad
good	good	bad	bad
bad	good	bad	good
bad	good	bad	bad
good	bad	bad	bad
bad	bad	bad	bad
good	good	bad	bad
bad	good	good	bad
good	good	bad	bad
bad	good	bad	bad
good	bad	good	bad

Region east

Count of Case	Column Labels		
Row Labels	bad	good	Grand Total
F	1	2	3
M	2	4	6
Grand Total	**3**	**6**	**9**

Fig. 5.13 Restricting page to east region

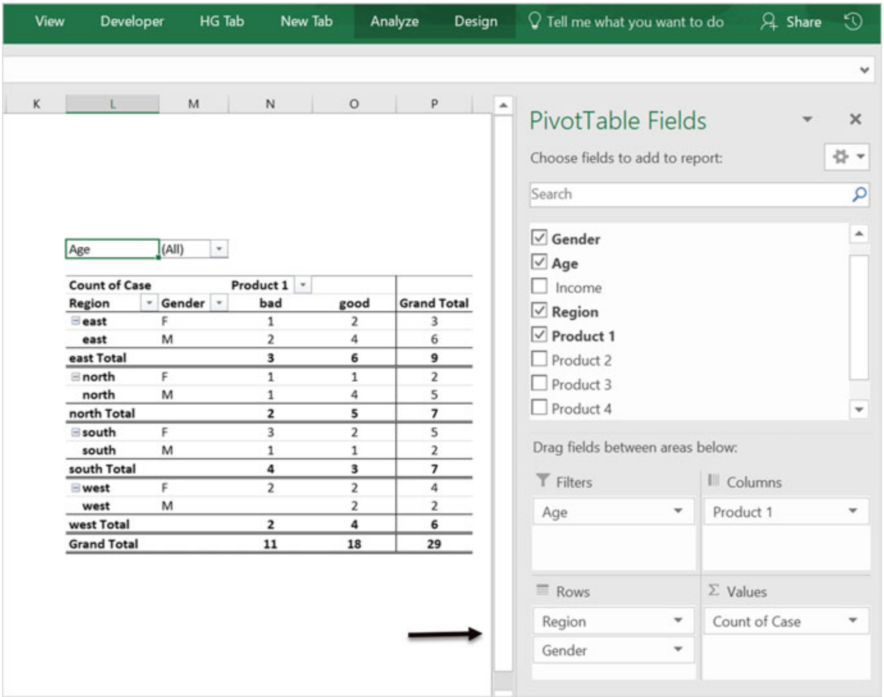


t 2	Product 3	Product 4
	good	bad
	bad	bad
	good	bad
	bad	bad
	bad	good
	bad	bad
	bad	bad
	bad	bad
	bad	bad
	good	bad
	bad	bad
	bad	bad
	good	bad
	good	good
	bad	bad

Region (All)

Count of Case	Column Labels		
Row Labels	bad	good	Grand Total
F	24.14%	24.14%	48.28%
M	13.79%	37.93%	51.72%
Grand Total	37.93%	62.07%	100.00%

Fig. 5.14 Region limited to east- resulting table


View
Developer
HG Tab
New Tab
Analyze
Design
Tell me what you want to do
Share
K
L
M
N
O
P
PivotTable Fields
Choose fields to add to report:
Search
Drag fields between areas below:

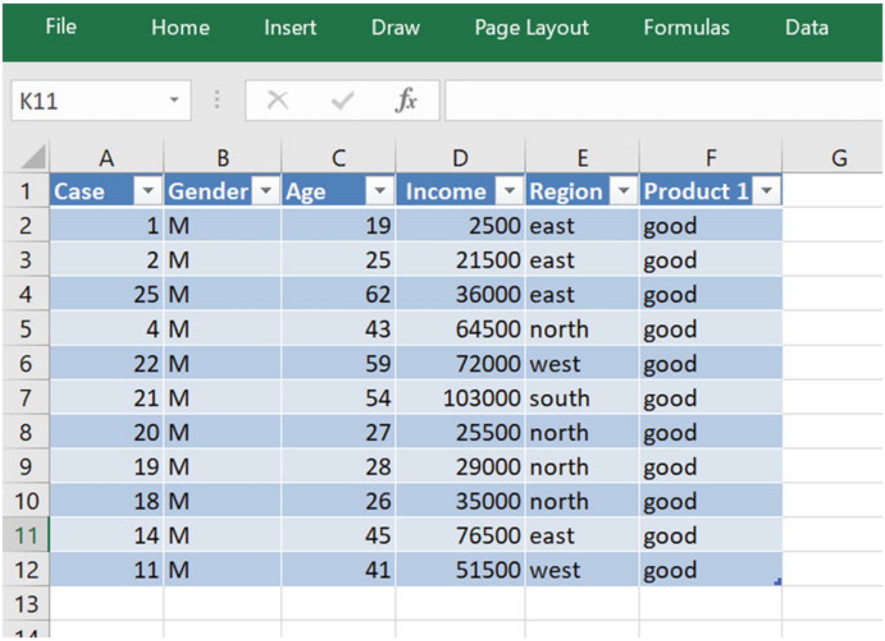
Age (All)

Count of Case		Product 1		
Region	Gender	bad	good	Grand Total
⊟ east	F	1	2	3
east	M	2	4	6
east Total		3	6	9
⊟ north	F	1	1	2
north	M	1	4	5
north Total		2	5	7
⊟ south	F	3	2	5
south	M	1	1	2
south Total		4	3	7
⊟ west	F	2	2	4
west	M		2	2
west Total		2	4	6
Grand Total		11	18	29

☑ Gender
☑ Age
☐ Income
☑ Region
☑ Product 1
☐ Product 2
☐ Product 3
☐ Product 4

Filters: Age
Columns: Product 1
Rows: Region, Gender
Values: Count of Case

Fig. 5.15 Extension of row field- resulting table



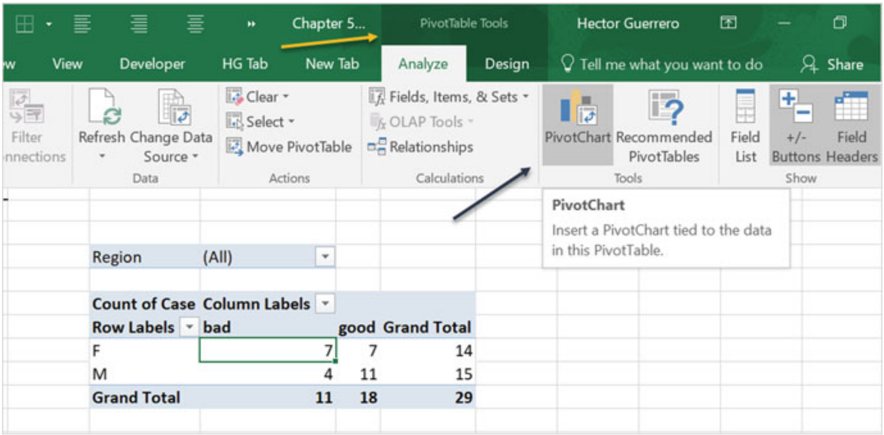
	A	B	C	D	E	F	G
1	Case	Gender	Age	Income	Region	Product 1	
2	1	M	19	2500	east	good	
3	2	M	25	21500	east	good	
4	25	M	62	36000	east	good	
5	4	M	43	64500	north	good	
6	22	M	59	72000	west	good	
7	21	M	54	103000	south	good	
8	20	M	27	25500	north	good	
9	19	M	28	29000	north	good	
10	18	M	26	35000	north	good	
11	14	M	45	76500	east	good	
12	11	M	41	51500	west	good	
13							

Fig. 5.16 Identify the respondents associated with a table cell

in the cell count. In Fig. 5.16 the 11 males that responded *Good* to *Product 1* are shown; they are respondents 1, 2, 25, 4, 22, 21, 20, 19, 18, 14, and 11. If you return to Fig. 5.9, you can see the cell of interest (the cell contains the count 11) to double-click. These records are now ready for further analysis and visualization. This is particularly useful in data that contains thousands of records.

5.3.3 PivotCharts

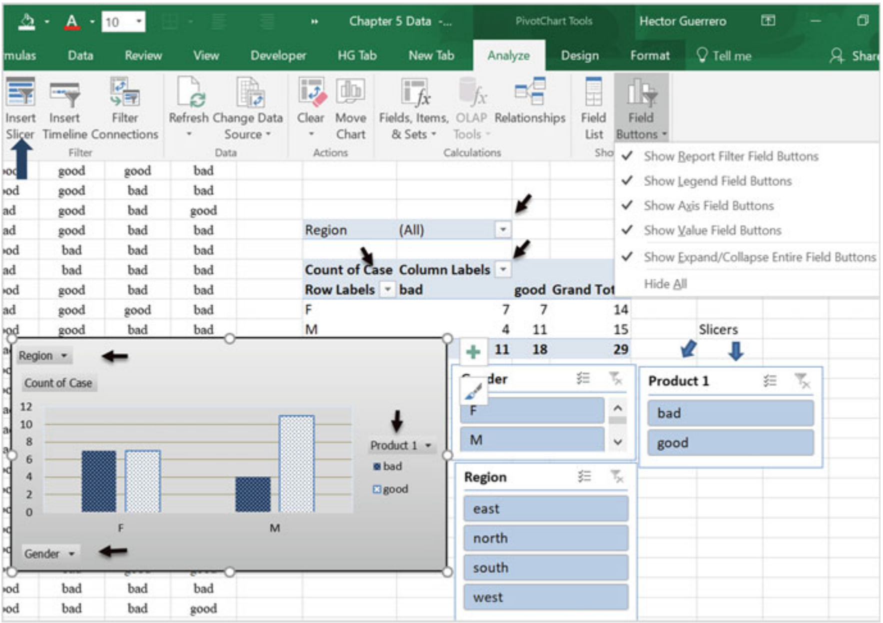
Now, let us repeat the process steps described above to construct a *PivotChart*. There is little difference in the steps to construct a table versus a chart. In fact, the process of constructing a *PivotChart* leads to the simultaneous construction of a *PivotTable*. The obvious difference is in step 1: rather than select the *PivotTable*, select the *PivotChart* option. The process of creating a *PivotChart* can be difficult; we will not invest a great deal of effort on the details. It is wise to always begin by creating a *PivotTable*. Creating a *PivotChart* then follows more easily. By simply selecting the *PivotTable*, a tab will appear above the row of ribbons—*PivotTable Tools*. In this ribbon you will find a tools group that permits conversion of a *PivotTable* to a *PivotChart*. See Fig. 5.17.



PivotChart
Insert a PivotChart tied to the data in this PivotTable.

Region	(All)		
Count of Case	Column Labels		
Row Labels	bad	good	Grand Total
F	7	7	14
M	4	11	15
Grand Total	11	18	29

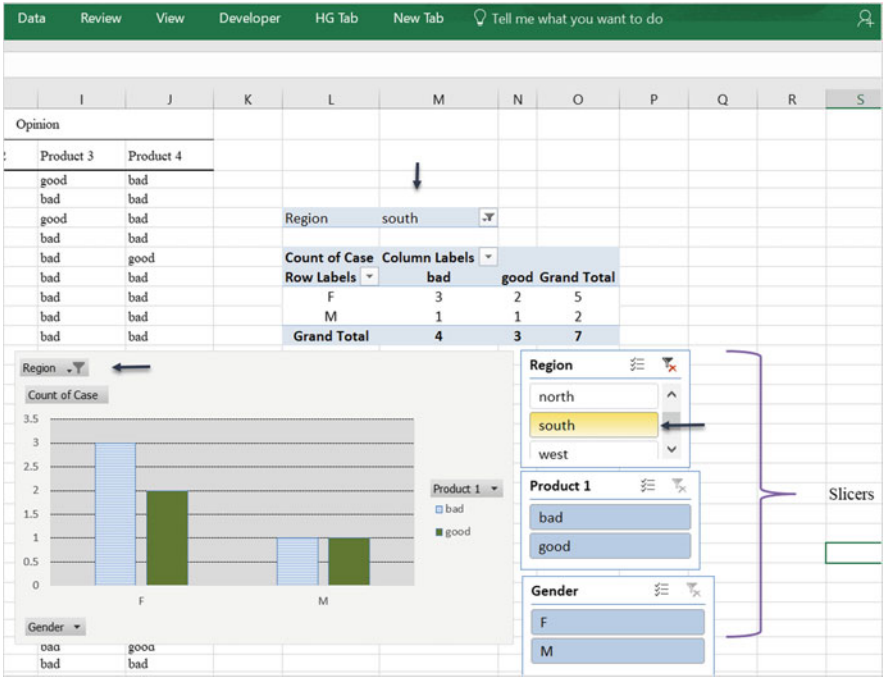
Fig. 5.17 *PivotTable* ribbon



Region	(All)		
Count of Case	Column Labels		
Row Labels	bad	good	Grand Tot
F	7	7	14
M	4	11	15
	11	18	29

Fig. 5.18 Pivot Chart equivalent of Pivot Table

The table we constructed in Fig. 5.9 is presented in Fig. 5.18, but as a chart. Note that a *PivotChart Field Buttons* is available in the *Analyze Ribbon* when the chart is selected. Just as before, it is possible to filter the data that is viewed by manipulating



Opinion

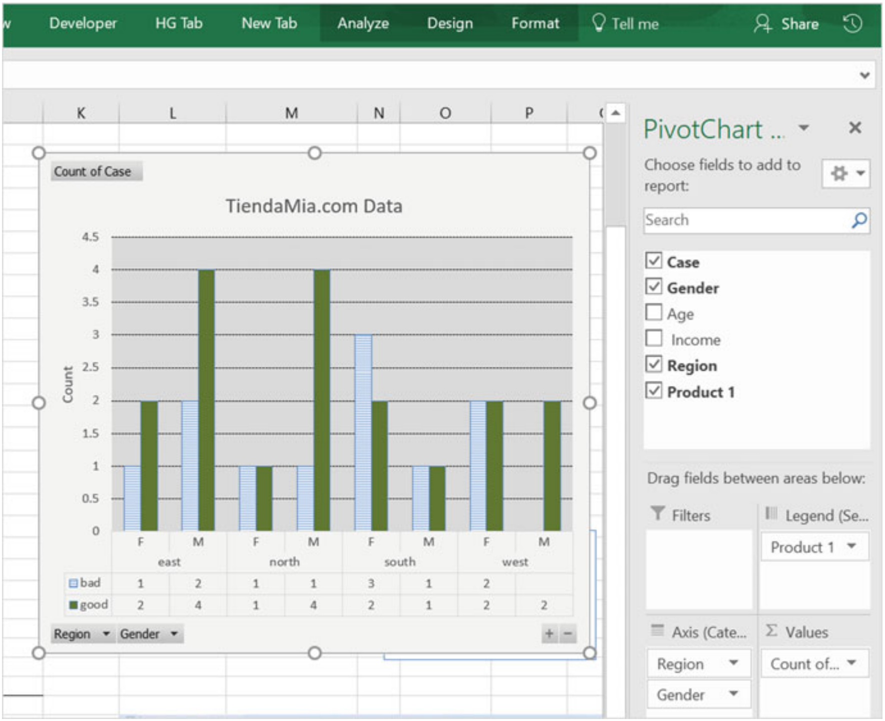
Product 3	Product 4
good	bad
bad	bad
good	bad
bad	bad
bad	good
bad	bad
bad	bad
bad	bad
bad	bad
bad	good
bad	bad

Region	south		
Count of Case	Column Labels		
Row Labels	bad	good	Grand Total
F	3	2	5
M	1	1	2
Grand Total	4	3	7

Fig. 5.19 Pivot Chart for south region only

the field buttons for *Gender*, *Product* and *Region* directly on the graph. You can also place *Slicers* on the spreadsheet but contacting the chart and selecting *Insert Slicers* in the *Analyze* ribbon, as shown in the Figs. 5.18 and 5.19. Thus, there are three possible methods of filtering: *Slicers*, *PivotTable* menus, and *PivotChart* menus. Figure 5.19 shows the result of changing the *Filters* field from all regions to only the South. From the chart, it is apparent that there are seven respondents that are contained in the south region, and of the seven, three females responded that *Product 1* was *bad*. Additionally, the chart can be extended to include multiple fields, as shown by the addition of *Region* to the *Row* field, along with *Gender*. See Fig. 5.20. This is equivalent to the *PivotTable* in Fig. 5.15.

As with any Excel chart, the user can specify the type of chart and other options for presentation. In Figs. 5.20 and 5.21 we show the data table associated with the chart; thus, the viewer has access to the chart and a table, simultaneously. Charts are powerful visual aids for presenting analysis, and are often more appealing and accessible than tables of numbers. The choice of a table versus a chart is a matter of preference. Figure 5.21 also shows the many options available for chart visualization.



Choose fields to add to report:

☑ Case
☑ Gender
□ Age
□ Income
☑ Region
☑ Product 1

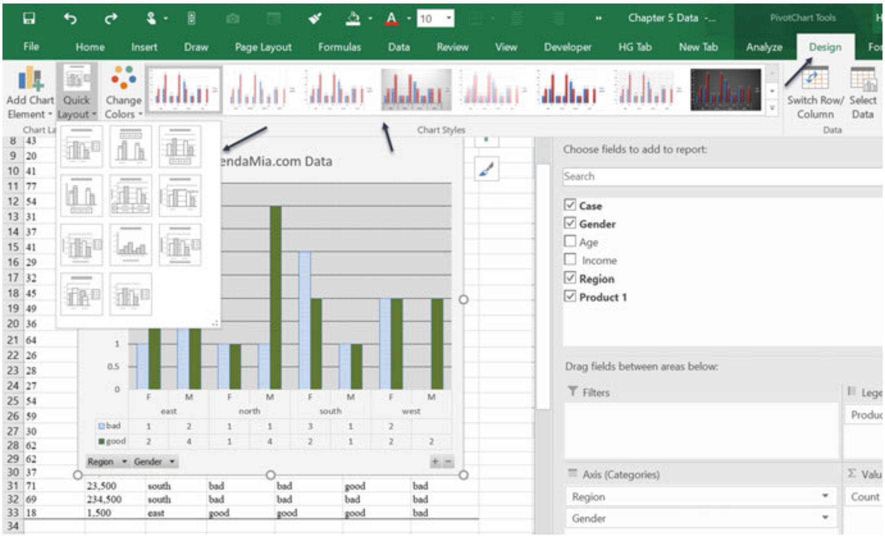
Legend (Se...: Product 1

Axis (Cate...: Region, Gender

Values: Count of...

	east		north		south		west	
	F	M	F	M	F	M	F	M
bad	1	2	1	1	3	1	2	
good	2	4	1	4	2	1	2	2

Fig. 5.20 Extended axis field to include gender and region



Choose fields to add to report:

☑ Case
☑ Gender
□ Age
□ Income
☑ Region
☑ Product 1

Axis (Categories): Region, Gender

	east		north		south		west	
	F	M	F	M	F	M	F	M
bad	1	2	1	1	3	1	2	
good	2	4	1	4	2	1	2	2

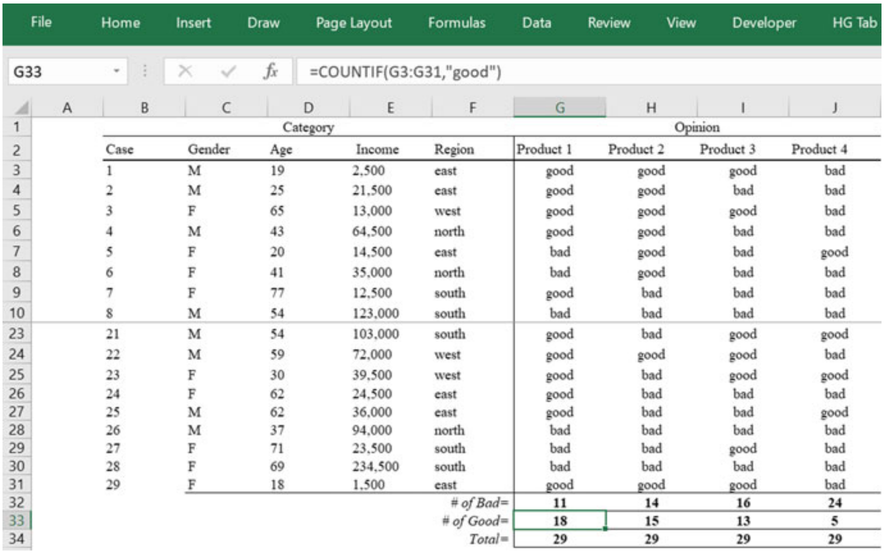
8	43						
9	20						
10	41						
11	77						
12	54						
13	31						
14	37						
15	41						
16	29						
17	32						
18	45						
19	49						
20	36						
21	64						
22	26						
23	28						
24	27						
25	54						
26	59						
27	30						
28	62						
29	62						
30	37						
31	71	23,500	south	bad	bad	good	bad
32	69	234,500	south	bad	bad	bad	bad
33	18	1,500	east	good	good	good	bad
34							

Fig. 5.21 Data table options for PivotCharts

5.4 TiendaMía.com Example: Question 1

Now, back to the questions that the owners of TiendaMía.com asked earlier. But before we begin with the cross-tabulation analysis, a warning is in order. As with previous examples, this example has a relatively small number of respondents (29). It is dangerous to infer that the result of a small sample (29) is indicative of the entire population of TiendaMía.com customers. We will say more about sample size in the next chapter, but generally larger samples provide greater comfort in the generalization of results. For now, we can assume that this study is intended as a preliminary analysis, and leading to more rigorous study later. Thus, we will also assume that our sample is large enough to be meaningful. Now, we consider the first question—Is there a webpage design that dominates others in terms of positive customer response?

To answer this question, we need not use cross-tabulation analysis. Cross-tabulation provides insight into how the various characteristics of the respondents relate to preferences; our question is one that is concerned with summary data for respondents without regard to detailed characteristics. So, let us focus on how many respondents prefer each webpage design? Let's use the **COUNTIF(range, criteria)** cell function to count the number of *bad* and *good* responses that are found in our data table. For Product 1 in Fig. 5.22, the formula in cell G33 is *COUNTIF(G3:G31,"good")*. Thus, the counter will count a cell value if it corresponds to the criterion that is provided in the cell formula, in this case *good*. Note that a split screen is used in this figure (hiding most of the data), but it is obvious that *Product 1*, with 18 *good*, dominates all products. *Product 2* and *Product 3* are relatively close



=COUNTIF(G3:G31,"good")

	A	B	C	D	E	F	G	H	I	J
1				Category				Opinion		
2		Case	Gender	Age	Income	Region	Product 1	Product 2	Product 3	Product 4
3		1	M	19	2,500	east	good	good	good	bad
4		2	M	25	21,500	east	good	good	bad	bad
5		3	F	65	13,000	west	good	good	good	bad
6		4	M	43	64,500	north	good	good	bad	bad
7		5	F	20	14,500	east	bad	good	bad	good
8		6	F	41	35,000	north	bad	good	bad	bad
9		7	F	77	12,500	south	good	bad	bad	bad
10		8	M	54	123,000	south	bad	bad	bad	bad
23		21	M	54	103,000	south	good	bad	good	good
24		22	M	59	72,000	west	good	good	good	bad
25		23	F	30	39,500	west	good	bad	good	good
26		24	F	62	24,500	east	good	bad	bad	bad
27		25	M	62	36,000	east	good	bad	bad	good
28		26	M	37	94,000	north	bad	bad	bad	bad
29		27	F	71	23,500	south	bad	bad	good	bad
30		28	F	69	234,500	south	bad	bad	bad	bad
31		29	F	18	1,500	east	good	good	good	bad
32						*# of Bad=*	**11**	**14**	**16**	**24**
33						*# of Good=*	**18**	**15**	**13**	**5**
34						*Total=*	**29**	**29**	**29**	**29**

Fig. 5.22 Summary analysis of product preference

(15 and 13) to each other, but *Product 4* is significantly different, with only 5 *good* responses. Again, recall that this result is based on a relatively small sample size, so we must be careful to understand that if we require a high degree of assurance about the results, we need a much larger sample size than 29 respondents.

The strength of the preferences in our data is recorded as a simple dichotomous choice—*good* or *bad*. In designing the survey, there are other possible data collection options that could have been used for preferences. For example, the respondents could be asked to rank the webpages from best to worse. This would provide information of the relative position (ordinal data) of the webpages, but it would not determine if they were acceptable or unacceptable, as is the case with *good* and *bad* categories. An approach that brings both types of data together could create a scale with one extreme representing a *highly favorable* webpage, the center value a *neutral* position, and the other extreme as *highly unfavorable*. Thus, we can determine if a design is acceptable and determine the relative position of one design versus the others. For now, we can see that relative to Products 1, 2, and 3, Product 4 is by far the least acceptable option.

5.5 TiendaMía.com Example: Question 2

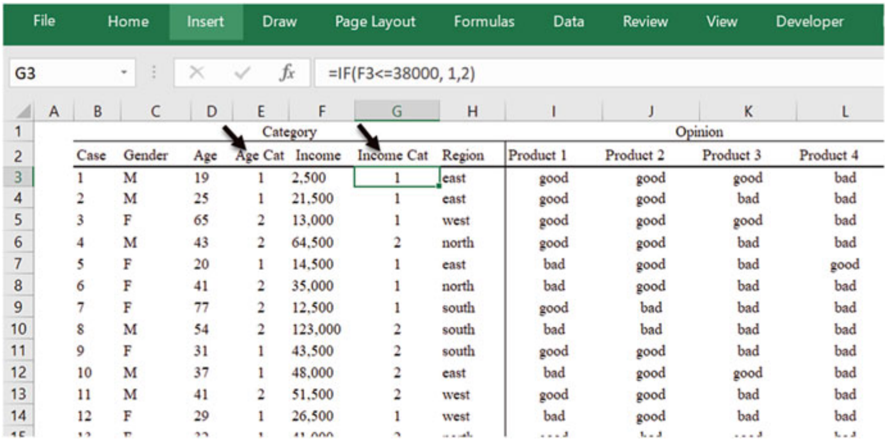
Question 2 asks how the demographic data of the respondents relates to preferences. This is precisely the type of question that can easily be handled by using cross-tabulation analysis. Our demographic characteristics are represented by *Gender*, *Age*, *Income*, and *Region*. *Gender* and *Region* have two and four variable levels (categories), respectively, which are a manageable number of values. But, *Income* and *Age* are a different matter. The data has been collected in increments of $500 for income and units of years for age, resulting in many possible values for these variables. What is the value of having such detailed information? Is it absolutely necessary to have such detail for our goal of analyzing the connection between these demographic characteristics and preferences for webpages? Can we simplify the data by creating categories of contiguous values of the data and still answer our questions with some level of precision? These are important questions.

Survey studies often group individuals into age categories spanning multiple years (e.g. 17–21, 22–29, 30–37, etc.) that permit easier cross-tabulation analysis, with minor loss of important detail. The same is true of income. We often find with quantitative data that it is advantageous, from a data management point of view, to create a limited number of categories. This can be done *after* the initial collection of detailed data. Thus, the data in Table 5.2 would be collected then *conditioned*, or *scrubbed*, to reflect categories for both *Age* and *Income*. In an earlier chapter, we introduced the idea of collecting data that would serve multiple purposes, and even unanticipated purposes. Table 5.2 data is a perfect example of such data.

So, let us create categories for *Age* and *Income* that are easy[5] to work with and simple to understand. For *Age*, we will use the following categories: 18–37; 38-older. Let us assume that they represent groups of consumers that figure similar behavior: purchasing characteristics, visits to the site, level of expenditures per visit, etc. For *Income*, we will use $0–38,000; $38,001-above. Again, assume that we have captured similar financial behavior for respondents in these categories. Note that we have 2 categories for each dimension and we will apply a numeric value to the categories—*1* for values in the lowest range and *2* in the highest. The changes resulting for the initial Table 5.2 data are shown in Table 5.6. The conversion to these categories can be accomplished with an *IF* statement. For example, *IF (F3 < =38,000, 1,2)* returns *1* if the income of the first respondent is less than or equal to $38,000, otherwise *2* is returned.

Generally, the selection of the categories should be based on the expertise of the data collector (TiendaMía.com) or their advisors. There are commonly accepted categories in many industries and can be found by reading research studies or the popular press associated with a business sector—e.g. the U.S. Census Bureau often uses the following age categories for income studies—below 15, 15–24, 25–34, 35–44, etc. Other sources producing studies in specific areas of industry or business, such as industry trade associations, can be invaluable as sources of demographic/financial category standards.

Table 5.6 Age and income category extension



G3 =IF(F3<=38000, 1,2)

	A	B	C	D	E	F	G	H	I	J	K	L
1					Category					Opinion		
2		Case	Gender	Age	Age Cat	Income	Income Cat	Region	Product 1	Product 2	Product 3	Product 4
3		1	M	19	1	2,500	1	east	good	good	good	bad
4		2	M	25	1	21,500	1	east	good	good	bad	bad
5		3	F	65	2	13,000	1	west	good	good	good	bad
6		4	M	43	2	64,500	2	north	good	good	bad	bad
7		5	F	20	1	14,500	1	east	bad	good	bad	good
8		6	F	41	2	35,000	1	north	bad	good	bad	bad
9		7	F	77	2	12,500	1	south	good	bad	bad	bad
10		8	M	54	2	123,000	2	south	bad	bad	bad	bad
11		9	F	31	1	43,500	2	south	good	good	bad	bad
12		10	M	37	1	48,000	2	east	bad	good	good	bad
13		11	M	41	2	51,500	2	west	good	good	bad	bad
14		12	F	29	1	26,500	1	west	bad	good	bad	bad

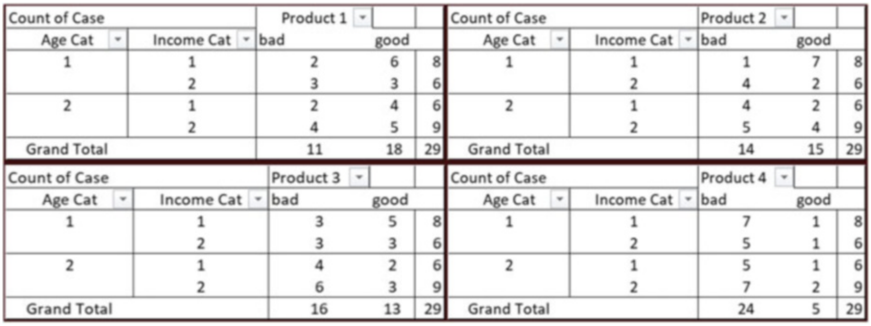
[5] Since we are working with a very small sample, the categories have been chosen to reflect differences in the relationships between demographic/financial characteristics and preferences. In other words, I have made sure the selection of categories results in interesting findings for this simple example.

Now, back to the question related to the respondent's demographic characteristics and how those characteristics relate to preferences. TiendaMía.com is interested in targeting particular customers with particular products, and doing so with a particular web design. Great product offerings are not always enough to entice customers to buy. TiendaMía.com understands that a great web design can influence customers to buy more items and more expensive products. This is why they are concerned with the attitudes of respondents toward the set of four webpage designs. So, let us examine *which* respondents prefer *which* webpages.

Assume that our management team at TiendaMía.com believes that *Income* and *Age* are the characteristics of greatest importance; *Gender* plays a small part in preferences and *Region* plays an even lesser role. We construct a set of four *PivotTables* that contain the cross-tabulations for comparison of all the products and respondents in our study. Results for all four products are combined in Fig. 5.23, beginning with *Product 1* in the Northwest corner and *Product 4* in the Southeast—note the titles in the column field identifying the four products. One common characteristic of data in each cross-tabulation is the number of individuals that populate each combination of demographic/financial categories—e.g. there are 8 individuals in the combination of the 18–37 *Age* range and 0-$$38,000 *Income* category; there are 6 that are in the 18–37 and $38,001–above categories, etc. These numbers are in the **Grand Totals** in each *PivotTable*.

To facilitate the formal analysis, let us introduce a shorthand designation for identifying categories: *AgeCategory\IncomeCategory*. We will use the category values introduced earlier to shorten and simplify the Age-Income combinations. Thus, 1\1 is the 18–37 *Age* and the 0-$38,000 *Income* combination. Now here are some observations that can be reached by examining Fig. 5.23:

1. Category 1\1 has strong opinions about products. They are positive to very positive regarding *Products 1, 2,* and *3* and they strongly dislike *Product 4*. For example, for *Product 1*, category 1\1 rated it *bad* = 2 and *good* = 6.
2. Category 1\2 is neutral about *Products 1, 2,* and *3*, but strongly negative on *Product 4*. It may be argued that they are not neutral on *Product 2*. This is an important



Count of Case		Product 1 ▾		
Age Cat ▾	Income Cat ▾	bad	good	
1	1	2	6	8
	2	3	3	6
2	1	2	4	6
	2	4	5	9
Grand Total		11	18	29

Count of Case		Product 2 ▾		
Age Cat ▾	Income Cat ▾	bad	good	
1	1	1	7	8
	2	4	2	6
2	1	4	2	6
	2	5	4	9
Grand Total		14	15	29

Count of Case		Product 3 ▾		
Age Cat ▾	Income Cat ▾	bad	good	
1	1	3	5	8
	2	3	3	6
2	1	4	2	6
	2	6	3	9
Grand Total		16	13	29

Count of Case		Product 4 ▾		
Age Cat ▾	Income Cat ▾	bad	good	
1	1	7	1	8
	2	5	1	6
2	1	5	1	6
	2	7	2	9
Grand Total		24	5	29

Fig. 5.23 Age and income category extension

category due to their higher income and therefore their higher potential for spending. For example, for *Product 4*, category 1\2 rated it *bad* = 5 and *good* = 1.
3. Category 2\1 takes slightly stronger positions than 1\2, and they are only positive about *Product 1*. They also take opposite positions than 1\1 on *Products 2* and *3*, but agree on *Products 1* and *4*.
4. Category 2\2 is relatively neutral on *Product 1* and *2* and negative on *Product 3* and *4*. Thus, 2\2 is not particularly impressed with any of the products, but the category is certainly unimpressed with Products 3 and 4.
5. Clearly, there is universal disapproval for *Product 4*, and the disapproval is quite strong. Ratings by 1\1, 1\2, 2\1, 2\2 are far more negative than positive: 24 out 29 respondents rated it *bad*.

There is no clear consensus for a webpage design that is acceptable to all categories, but clearly Product 4 is a disaster. If TiendaMía.com decides to use a single webpage design, which one is the most practical design to use? This is not a simple question given out results. TiendaMía.com may decide that the question requires further study. Why? Here are general reasons for further study:

1. if the conclusions from the data analysis are inconclusive
2. if the size of the sample is deemed to be too small, then the preferences reflected may not merit a generalization of results to the population of website users
3. if the study reveals new questions of interest or guides us in new directions that might lead us to eventual answers for these questions.

Let us consider number (3) above. We will perform a slightly different analysis by asking the following new question—is there a single measure that permits an overall ranking of products? The answer is yes. We can summarize the preferences shown in Fig. 5.23 in terms of a new measure—*favorable rating*.

Table 5.7 organizes the respondent categories and their *favorable rating*—the ratio[6] of *good* responses relative to the total of *all responses.* From Fig. 5.23 you see that respondents in category 1\1 have an 87.5% (7 of 8 rated the product *good*) favorable rating for *Product 2*. This is written as P-2 (87.5%). Similarly, category 2\1 has a favorable rating for P-2 and P-3 of 33.3% (2 of 6). To facilitate comparison, the favorable ratings are arranged on the vertical axis of the table, with highest near the top of the table and lowest near the bottom, with a corresponding scale from acceptable to neutral to unacceptable. (Note this is simply a table and not a *PivotTable.*)

A casual analysis of the table suggests that *Product 1* shows a 50% or greater favorable rating for all categories. No other product can equal this favorable rating: *Product 1* is the top choice of all respondent categories except for 1\1 (75%) and it is tied with *Product 3* for category 1\2. Although this casual analysis suggests a clear choice, we can now do more formal analyses to select a single website design.

First, we will calculate the average of all *favorable ratings* for each category (1\1, 1\2, 2\1, 2\2) as a single composite score. This is a simple calculation and it provides

[6] [number *good*] ÷ [number *good* + number *bad*].

Table 5.7 Detailed view of respondent favorable ratings

Respondent category	1\1	1\2	2\1	2\2
Acceptable	P-2 (87.5%)			
	P-1 (75.0%)			
			P-1 (66.7%)	
	P-3 (62.5%)			
				P-1 (55.6%)
Neutral		P-1&3 (50%)		
				P-2 (44.4%)
		P-2 (33.3%)	P-2&3 (33.3%)	P-3 (33.3%)
				P-4 (22.2%)
		P-4 (16.7%)	P-4 (16.7%)	
Unacceptable	P-4 (12.5%)			

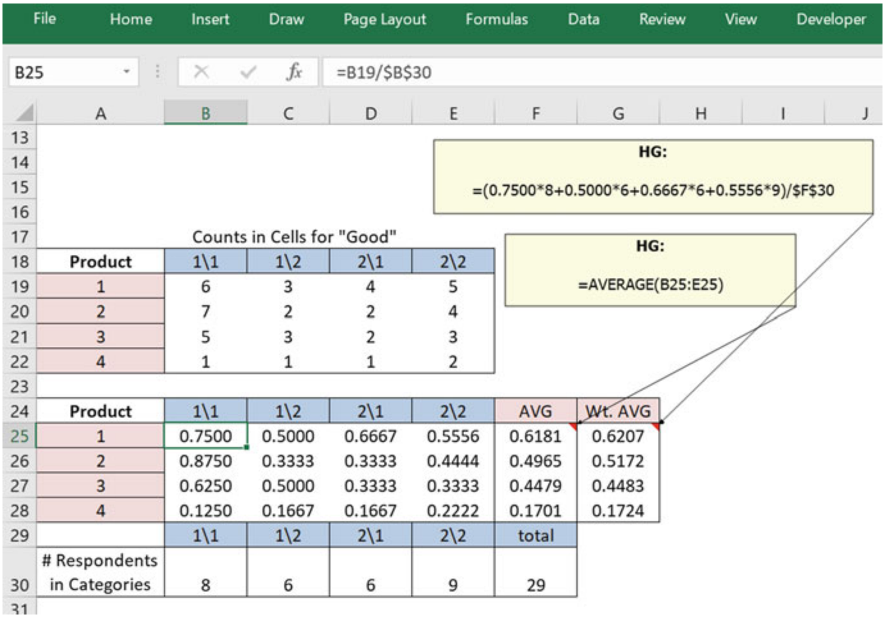
a straightforward method for TiendaMía.com to assess products. In Fig. 5.24 the calculation of averages is found in F25:F28–0.6181 for *Product 1*, 0.4965 for *Product 2*, etc. *Product 1* has the highest average favorable rating. But there are some questions that we might ask about the fairness of the calculated averages. Should there be an approximately similar number of respondents in each category for this approach to be fair? Stated differently, is it fair to count an individual category average equally to others when the number of respondents in that category is substantially less than other categories?

In TiendaMía.com's study, there *are* different numbers of respondents in the various categories. This can be significant for the calculation of averages. The difference in the numbers can be due to the particular sample that we selected. A random sample of this small size can lead to wide variation in the respondents selected. One way to deal with this problem is to consciously sample customers to reflect the proportion of category members that shop at TiendaMía.com. There are many techniques and methods for formally **sampling**[7] data that we cannot study here.

For the moment, let's assume that the random sample has selected a proportionally fair representation of respondents, and this is what TiendaMía.com desires. Thus, 28% (8÷29, 8 of 1\1 respondents out of 29) should be relatively close to the population of all 1\1's in TiendaMía.com's customer population. If we want to account for the difference in respondent category size in our analysis, then we will want to calculate a *weighted* average of favorable ratings, which reflects the relative size of the respondent categories. Note that the first average that we calculated is a special form of weighted average: one where all weights were assumed to be equal. In range G25:G28 of Fig. 5.24, we see the calculation of the weighted average. Each average is multiplied by the fraction of respondents that it represents of the total sample.[8] This approach provides a proportional emphasis on averages. If a particular

[7] Sampling theory is a rich science that should be carefully considered prior to initiating a study.

[8] $(0.7500 * 8 + 0.5000 * 6 + 0.6667 * 6 + 0.5556 * 9)/29 = 0.6207$.



=B19/$B$30

HG:

=(0.7500*8+0.5000*6+0.6667*6+0.5556*9)/$F$30

HG:

=AVERAGE(B25:E25)

Counts in Cells for "Good"

Product	1\1	1\2	2\1	2\2
1	6	3	4	5
2	7	2	2	4
3	5	3	2	3
4	1	1	1	2

Product	1\1	1\2	2\1	2\2	AVG	Wt. AVG
1	0.7500	0.5000	0.6667	0.5556	0.6181	0.6207
2	0.8750	0.3333	0.3333	0.4444	0.4965	0.5172
3	0.6250	0.5000	0.3333	0.3333	0.4479	0.4483
4	0.1250	0.1667	0.1667	0.2222	0.1701	0.1724
	1\1	1\2	2\1	2\2	total	
# Respondents in Categories	8	6	6	9	29	

Fig. 5.24 Calculation of average (AVG) and weighted average (Wt-AVG)

average is composed of many respondents, then it will receive a higher weight; if an average is composed of fewer respondents, then it will receive a lower weight.

So, what do our respondent weighted averages (G25:G28) reveal about Products compared to the equally weighted averages (F25:F28)? The results are approximately the same for *Products 1*and *4*. The exceptions are *Product 2* with a somewhat stronger showing, moving from 0.4965 to 0.5172, and Product 3 with a drop in score from 0.3073 to 0.2931. Still, there is no change in the ranking of the products; it remains P-1, P-2, P-3, and P-4.

What has led to the increase in the *Product 2* score? Categories 1\1 and 2\2 are the highest favorable ratings for *Product 2*; they also happen to be the largest weighted categories (8/29 = 0.276 and 9/29 = 0.310). Larger weights applied to the highest scores will of course yield a higher weighted average. If TiendaMía.com wants to focus attention on these market segments, then a weighted average may be appropriate. **Market segmentation** is a very important element in their marketing strategy.

There may be other ways to weight the favorable ratings. For example, there may be categories that are more important than others due to their higher spending per transaction or more frequent transactions at the site. So, as you can see, many weighting schemes are possible.

5.6 Summary

Cross-tabulation analysis through the use *PivotTables* and *PivotCharts* is a simple and effective way to analyze qualitative data, but to insure fair and accurate analysis, the data must be carefully examined and prepared. Rarely is a data set of significant size without errors. Although most errors are usually accidental, there may be some that are intentional. Excel provides many logical cell functions to determine if data have been accurately captured and fit the data specifications that an analyst has imposed.

PivotTables and *PivotCharts* allow the analyst to view the interaction of several variables in a data set. To do so, it is often necessary to convert data elements in surveys into values that permit easier manipulation—e.g. we converted *Income* and *Age* into categorical data. This does not suggest that we have made an error in how we collected data; on the contrary, it is often advantageous to collect data in its *purest* form (e.g. 23 years of age) versus providing a category value (e.g. the 19–24 years of age category). This allows future detailed uses of the data that may not be anticipated.

In the next chapter, we will begin to apply more sophisticated statistical techniques to qualitative data. These techniques will permit us to study the interaction between variables, and allow us to *quantify* how confident we are that the conclusions we reach. In other words, is my sample representative of the population of interest? Among the techniques we will introduce are Analysis of Variance (ANOVA), tests of hypothesis with t-tests and z-tests, and chi-square tests. These are powerful statistical techniques that can be used to study the effect of *independent* variables on *dependent* variables, and determine similarity or dissimilarity in data samples. When used in conjunction with the techniques we have learned in this chapter, we have the capability of uncovering the complex data interactions that are essential to successful decision-making.

Key Terms

Data Errors	Collectively Exhaustive
EXACT (text1, text2)	Data area
Error checking	Count
TRUE/FALSE	Filters, column, row, and values
OR, AND, NOT, TRUE, FALSE	Sum, count, average, min, max
MOD (number, divisor)	COUNTIF (range, criteria)
Cross-tabulation	Grand totals
PivotTable/PivotChart	Sampling
Data scrubbing	Market segmentation
Mutually exclusive	

Problems and Exercises

1. Data errors are of little consequence in data analysis—T or F.
2. What does the term *data scrubbing* mean?
3. Write a *logical if* function for a cell (A1) that tests if a cell contains a value larger than or equal to 15 or less than 15. Return phrases that say "15 or more" or "less than 15."
4. For Fig. 5.1, write a *logical IF* function in the cells H2:I4 that calculates the difference between *Original Data Entry* and *Secondary Data Entry* for each cell of the corresponding cells. If the difference is not 0, then return the phrase "Not Equal", otherwise return "Equal."
5. Use a *logical IF* function in cell A1 to test a value in B1. Examine the contents of B1and return "In" if the values are between and include the range 2 and 9. If the value is outside this range, return "Not In."
6. Use a *logical IF* function in cell A1 to test values in B1 and C1. If the contents of B1 *and* C1 are 12 and 23, respectively, return "Values are 12 and 23", otherwise return "Values are not 12 and 23."
7. Use a *logical IF* function in cell A1 to test if a value in B1 is an integer. Use the *Mod* function to make the determination. Return either "Integer" or "Not Integer."
8. What type of analysis does cross-tabulation allow a data analyst to perform?
9. What types of data (categorical, ordinal, etc.) will *PivotTables* and *PivotCharts* permit you to cross-tabulate?
10. Create a *PivotTable* from the data in Fig. 5.2 (minus Case 13) that performs a cross-tabulation analysis for the following configuration: *Region* on the Row field; *Income* in the Values field; *Product 2* on the Column field; and *Age* on the Filter field.

    (a) What are the *counts* in the Values field?
    (b) What are the *averages* in the Values field?
    (c) What are the *maximums* in the Value field?
    (d) What Region has the maximum *count*?
    (e) For all regions is there a clear preference, good or bad, for Product 2?
    (f) What is the highest *average* income for a region/preference combination?
    (g) What combination of region and preference has the highest *variation* in Income?
11. Create a cell for counting values in range A1:A25 if the cell contents are equal to the text "New."
12. Create a *PivotChart* from the *PivotTable* analysis in 10c above.
13. The *Income* data in Table 5.6 is organized into two categories. Re-categorize the *Income* data into 3 categories—0–24,000; 24,001–75,000; 75,001 and above? How will the Fig. 5.23 change?
14. Perform the conversion of the data in Table 5.7 into a column chart that presents the same data graphically?

15. Create a weighted average based on the sum of incomes for the various categories. Hint—The weight should be related to the proportion of the category sum of income to the total of all income.
16. Your boss believes that the analysis in Fig. 5.23 is interesting, but she would like to see the Age category replaced with Region. Perform the new analysis and display the results similarly to those of Fig. 5.23.
17. *Advanced Problem*—A clinic that specializes in alcohol abuse has collected some data on their current clients. Their data for clients includes the number of years of abuse have experienced, age, years of schooling, number of parents in the household as a child, and the perceived chances for recovery by a panel of experts at the clinic. Determine the following using *PivotTables* and *PivotCharts*:

    (a) Is there a general relationship between age and the number of years of abuse?

    (b) For the following age categories, what proportion of their lives have clients abused alcohol:

    (i) 0–24
    (ii) 25–35
    (iii) 36–49
    (iv) 49–over.

    (c) What factor is the most reliable predictor of perceived chances for recovery? Which is the least?

    (d) What is the co-relationship between number of parents in the household and years of schooling?

    (e) What is the average age of the clients with bad prospects?

    (f) What is the average number of years of schooling for clients with one parent in the household as a child?

    (g) What is the average number of parents for all clients that have poor prospects for recovery?

Case	Yrs. abuse	Age	Years school	Number of parents	Prospects
1	6	26	12	1	G
2	9	41	12	1	B
3	11	49	11	2	B
4	5	20	8	2	G
5	6	29	9	1	B
6	8	34	13	2	B
7	12	54	16	2	G
8	7	33	16	1	G
9	9	37	14	1	G
10	7	31	10	2	B
11	6	26	7	2	B
12	7	30	12	1	G
13	8	37	12	2	B
14	12	48	7	2	B

(continued)

Case	Yrs. abuse	Age	Years school	Number of parents	Prospects
15	9	40	12	1	B
16	6	28	12	1	G
17	8	36	12	2	G
18	9	37	11	2	B
19	4	19	10	1	B
20	6	29	14	2	G
21	6	28	17	1	G
22	6	24	12	1	B
23	8	38	10	1	B
24	9	41	8	2	B
25	10	44	9	2	G
26	5	21	12	1	B
27	6	26	10	0	B
28	9	38	12	2	G
29	8	38	10	1	B
30	9	37	13	2	G