

PROBLEMS

Note: Student solutions for problems whose numbers appear within a colored box are available for purchase at www.cengagebrain.com.

Level A

1. The file **P02_01.xlsx** indicates the gender and nationality of the MBA incoming class in two successive years at the Kelley School of Business at Indiana University.
 - a. For each year separately, recode Nationality so that all nationalities with a count of 1 or 2 are listed as Other.
 - b. For each year, create a crosstabs of Gender versus the recoded Nationality and an associated column chart. Does there seem to be any relationship between Gender and the recoded Nationality? Is the pattern about the same in the two years?
2. The file **P02_03.xlsx** contains data from a survey of 399 people regarding a government environmental policy.
 - a. Create a crosstabs and an associated column chart for Gender versus Opinion. Express the counts as percentages so that for either gender, the percentages add to 100%. Discuss your findings. Specifically, do the two genders tend to differ in their opinions about the environmental policy?
 - b. Repeat part a with Age versus Opinion.
 - c. Recode Salary to be categorical with categories "Less than \$40K," "Between \$40K and \$70K," "Between \$70K and \$100K," and "Greater than \$100K" (where you can treat the breakpoints however you like). Then repeat part a with this new Salary variable versus Opinion.
3. The file **P02_02.xlsx** contains data about 211 movies released in 2006 and 2007.
 - a. Recode Distributor so that all distributors except for Paramount Pictures, Buena Vista, Fox Searchlight, Universal, Warner Bros., 20th Century Fox, and Sony Pictures are listed as Other. (Those in Other released fewer than 16 movies.) Similarly, recode Genre so that all genres except for Action, Adventure, Thriller/Suspense, Drama, and Comedy are listed as Other. (Again, those in Other are genres with fewer than 16 movies.)
 - b. Create a crosstabs and an associated column chart for these two recoded variables. Express the counts as percentages so that for any distributor, the percentages add to 100%. Discuss your findings.

4. Recall from Chapter 2 that the file **Supermarket Transactions.xlsx** contains over 14,000 transactions made by supermarket customers over a period of approximately two years. To understand which customers purchase which products, create a crosstabs and an associated column chart for each of the following. For each, express the counts as percentages so that for any value of the first variable listed, the percentages add to 100%. Do any patterns stand out?
 - a. Gender versus Product Department
 - b. Marital Status versus Product Department
 - c. Annual Income versus Product Department

Level B

5. Recall from Chapter 2 that HyTex Company is a direct marketer of technical products and that the file **Catalog Marketing.xlsx** contains recent data on 1000 HyTex customers. To understand these customers, first recode Salary and Amount Spent as indicated in part a, and then create each of the requested crosstabs and an associated column chart in parts b to e. Express each count as a percentage, so that for any value of the first variable listed, the percentages add up to 100%. Do any patterns stand out?
 - a. Find the first, second, and third quartiles of Salary, and then recode Salary as 1 to 4, depending on which quarter of the data each value falls into. For example, the first salary, \$16,400, is recoded as 1 because \$16,400 is less than the first quartile, \$29,975. Recode Amount Spent similarly, based on its quartiles. (Hint: The recoding can be done most easily with lookup tables.)
 - b. Age versus the recoded Amount Spent
 - c. Own Home versus the recoded Amount Spent
 - d. History versus the recoded Amount Spent
 - e. The recoded Salary versus the recoded Amount Spent
6. The smoking/drinking example in this section used the function COUNTIFS function to find the counts of the joint categories. Without using this function (or pivot tables), devise another way to get the counts. The raw data are in the file **Smoking Drinking.xlsx**. (Hint: One possibility is to concatenate the values in columns B and C into a new column D. But feel free to find the counts in any way you like.)

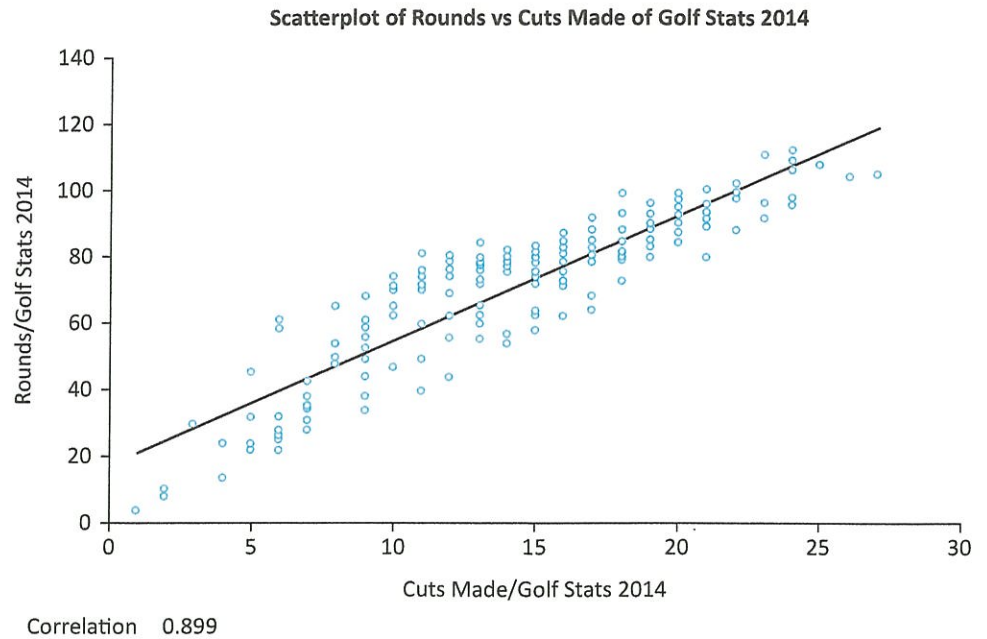
PROBLEMS

Level A

7. Recall that the file **Baseball Salaries Extra.xlsx** contains data on 868 Major League Baseball (MLB) players during the 2015 season. Find the mean, median, standard deviation, and first and third quartiles of Salary, broken down by each of the following categories. Comment on your findings.
 - a. Team
 - b. Division
 - c. Whether they played for the Yankees
 - d. Whether they were in the playoffs
8. The file **P02_07.xlsx** includes data on 204 employees at the (fictional) company Beta Technologies. Find the mean, median, and standard deviation of Annual Salary, broken down by each of the following categories. Comment on your findings.
 - a. Gender
 - b. A recoded version of Education, with new values 1 for Education less than 4, 2 for Education equal to 4, and 3 for Education greater than 4
 - c. A recoded version of Age, with people aged less than 34 listed as Young, those aged at least 34 and less than 50 listed as Middle-aged, and those aged at least 50 listed as Older
9. The file **Golf Stats.xlsx** contains data on the 200 top golfers each year from 2003 to 2014. (This data set is used in an example in the next section.) Create a recoded Age variable, with values “Twenties,” “Thirties,” and “Forties,” based on their ages in the 2014 sheet. Then calculate the mean, median, and standard deviation of the following 2014 variables, broken down by the recoded Age. Comment on whether it appears that golfers peak in their thirties.
 - a. Earnings
 - b. Yards/Drive and Driving Accuracy
 - c. Greens in Regulation
 - d. Putting Average (Golfers want this to be small.)
10. Recall from Chapter 2 that HyTex Company is a direct marketer of technical products and that the file **Catalog Marketing.xlsx** contains recent data on 1000 HyTex customers. Find the mean, median, and standard deviation of Amount Spent, broken down by the following variables. Then create side-by-side box plots of Amount Spent, broken down by the same variables. Comment on how the box plots complement the summary measures.
 - a. Age
 - b. Gender
 - c. Close
 - d. Region
 - e. Year of first purchase. (Hint: For this one, use Excel’s YEAR function to create a Year column.)
- f. The combination of Married and Own Home. (For this one, create a code variable, with values from 1 to 4, for the four combinations of Married and Own Home. Alternatively, create a text variable with values such as “Not married, Owns home.”)
11. The file **P02_35.xlsx** contains data from a survey of 500 randomly selected households.
 - a. Create a new column Has Second Income with values “Yes” and “No” depending on whether the household has a reported second income.
 - b. Find the mean, median, and standard deviation of First Income, broken down by the variable you created in part a. Is there any indication that first income tends to be any larger or smaller, or has more or less variation, depending on whether there is a second income?
 - c. Repeat part b for each of the Monthly Payment and Debt variables.
12. The file **P02_02.xlsx** contains data about 211 movies released in 2006 and 2007.
 - a. Recode Genre so that all genres except for Action, Adventure, Thriller/Suspense, Drama, and Comedy are listed as Other. (Those in Other are genres with fewer than 16 movies.)
 - b. Find the mean, median, and standard deviation of Total US Gross, broken down by the recoded Genre variable. Also, create side-by-side box plots of Total US Gross, again broken down by the recoded Genre variable. Comment on what the results say about the popularity of different genres.
13. *The Wall Street Journal CEO Compensation Study* analyzed chief executive officer (CEO) pay from many U.S. companies with fiscal year 2008 revenue of at least \$5 billion that filed their proxy statements between October 2008 and March 2009. The data are in the file **P02_30.xlsx**.
 - a. Create a new variable Total 2008, the sum of Salary 2008 and Bonus 2008. (Actually, this is not “total” compensation because it omits the very lucrative compensation from stock options.) Also, recode Company Type so that the Technology and Telecommunications are collapsed into a Tech/Telecomm category.
 - b. Find the mean, median, and standard deviation of Total 2008, broken down by the recoded Company Type. Also, create side-by-side box plots of Total 2008, again broken down by the recoded Company Type. What do the results tell you about differences in level or variability across company types?

Figure 3.26

Scatterplot of
Cuts Made Versus
Rounds



PROBLEMS

Level A

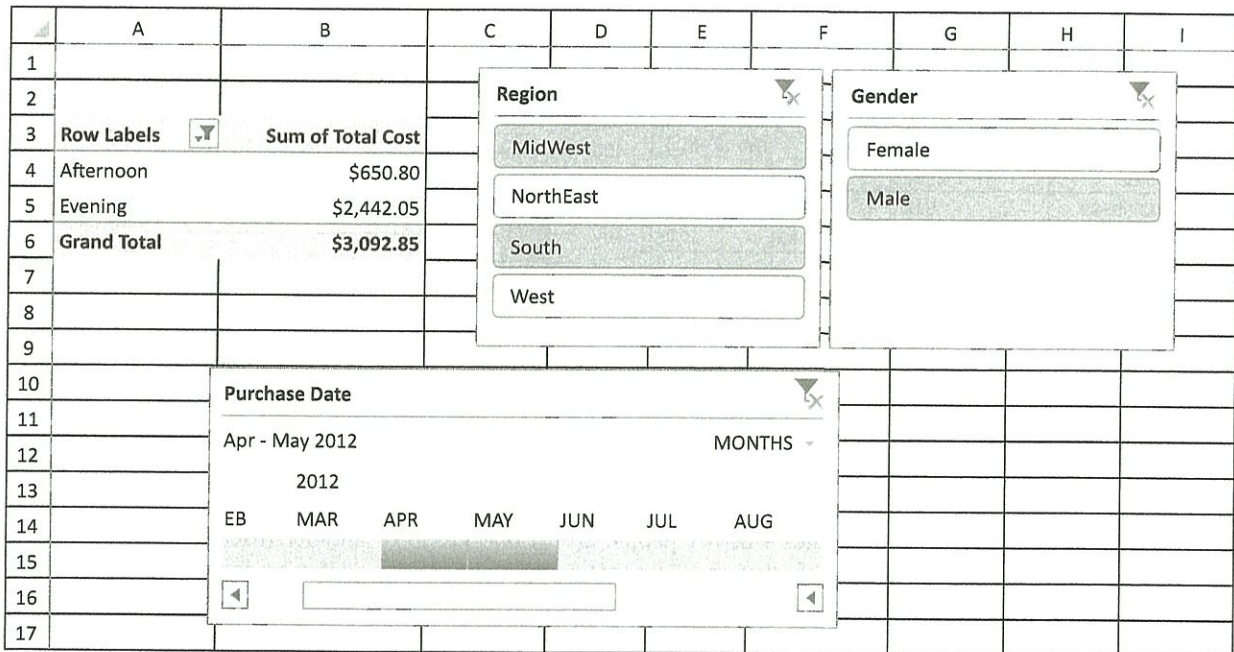
- 21.** The file **P02_07.xlsx** includes data on 204 employees at the (fictional) company Beta Technologies.
- Create a table of correlations between the variables Age, Prior Experience, Beta Experience, Education, and Annual Salary. Which of the first four of these variables is most highly correlated (in a positive direction) with Annual Salary?
 - Create scatterplots of Annual Salary (Y axis) versus each of Age, Prior Experience, Beta Experience, and Education.
 - For the variable from part **a** most highly correlated with Annual Salary, create a (linear) trend line in its scatterplot with the corresponding equation shown in the chart. What does this equation imply about the relationship between the two variables? Be specific.
- 22.** The file **P03_22.xlsx** lists financial data on movies released from 1980 to 2011 with budgets of at least \$20 million.
- Reduce the size of this data set by deleting all movies with a budget of more than \$100 million. Also, delete all movies where US Gross and/or Worldwide Gross is listed as Unknown.
 - For the remaining movies, create a table of correlations between the variables Budget, US Gross, and Worldwide Gross. Comment on the results. Are there any surprises?
 - For the movies remaining after part **a**, create a scatterplot of Worldwide Gross (Y axis) versus US Gross and another scatterplot of US Gross (Y axis) versus Budget. Briefly explain any patterns you see in these scatterplots. Do they seem to be consistent with the corresponding correlations?
- 23.** The file **P02_10.xlsx** contains midterm and final exam scores for 96 students in a corporate finance course.
- Do the students' scores for the two exams tend to go together, so that those who do poorly on the midterm tend to do poorly on the final, and those who do well on the midterm tend to do well on the final? Create a scatterplot, along with a correlation, to answer this question.
 - Superimpose a (linear) trend line on the scatterplot, along with the equation of the line. Based on this equation, what would you expect a student with a 75 on the midterm to score on the final exam?
- 24.** Recall that the file **Golf Stats.xlsx** contains data on the 200 top golfers each year from 2003 to 2014. The

Electmart sales data, you can choose Region as a slicer. You then see a list on the screen with a button for each possible value: Midwest, Northeast, South, and West. You can then click any combination of these buttons to filter on the chosen regions. Note that a slicer variable does *not* have to be part of the pivot table. For example, if you are showing sum of Total Cost, and Region is not part of the pivot table, a Region slicer will still filter sum of Total Cost for the regions selected. On the other hand, if Region is already in the Rows area, say, you can filter on it through the slicer. In this case, selecting regions from the slicer is equivalent to filtering on the same regions in the Rows area. Basically, the slicers have been added as a convenience to users. They make filtering easier and more transparent.

A similar Timeline feature was added in Excel 2013. A timeline is like a slicer, but it is specifically for filtering on a date variable.

As an example, Figure 3.64 shows a pivot table accompanied by two slicers and a timeline. The Rows field is Time, which has been filtered in the usual way (through the dropdown list in the Rows area) to show only Afternoon and Evening. The two slicers appear next to the pivot table. The Region slicer has been filtered on Midwest and South, and the Gender slicer has been filtered on Male. Note that the Date field for the timeline has been renamed to Purchase Date so that the timeline works properly. Here you can see that the sum of Total Cost for all sales in the evening by males in the Midwest and South during the months of April and May is \$2442.05.

Figure 3.64 Pivot Table with Slicers and a Timeline



PROBLEMS

Level A

- 33.** Solve problem 1 with pivot tables and create corresponding pivot charts. Express the counts as percentages of row totals. What do these percentages indicate

about this particular data set? Then repeat, expressing the counts as percentages of column totals.

- 34.** Solve problem 2 with pivot tables and create corresponding pivot charts. Express the counts as

percentages of row totals. What do these percentages indicate about this particular data set? Then repeat, expressing the counts as percentages of column totals.

35. Solve problem 3 with pivot tables and create corresponding pivot charts. Express the counts as percentages of row totals. What do these percentages indicate about this particular data set? Then repeat, expressing the counts as percentages of column totals.
36. Solve problem 4 with pivot tables and create corresponding pivot charts. Express the counts as percentage of row totals. What do these percentages indicate about this particular data set? Then repeat, expressing the counts as percentages of column totals.
37. Solve problem 7 with pivot tables and create corresponding pivot charts. However, find only means and standard deviations, not medians or quartiles. (This is one drawback of pivot tables. Medians, quartiles, and percentiles are not in the list of summary measures.)
38. Solve problem 8 with pivot tables and create corresponding pivot charts. However, find only means and standard deviations, not medians. (This is one drawback of pivot tables. Medians are not among their summary measures.)
39. Solve problem 9 with pivot tables and create corresponding pivot charts. However, find only means and standard deviations, not medians. (This is one drawback of pivot tables. Medians are not among their summary measures.)
40. The file [P03_40.xlsx](#) contains monthly data on the number of vehicles crossing the border from Mexico into four southwestern states.
 - a. Restructure this data set on a new sheet so that there are three long columns: Month, State, and Crossings. Essentially, you should stack the original columns B through E on top of one another to get the Crossings column, and you should also indicate which state each row corresponds to in the State column. The Month column should have four replicas of the original Month column.
 - b. Create a pivot table and corresponding pivot table chart based on the restructured data. It should break down the average of Crossings by Year and State. Comment on any patterns you see in the chart.
41. *The Wall Street Journal CEO Compensation Study* analyzed CEO pay from many U.S. companies with fiscal year 2008 revenue of at least \$5 billion that filed their proxy statements between October 2008 and March 2009. The data are in the file [P02_30.xlsx](#).
 - a. Create a pivot table and a corresponding pivot chart that simultaneously shows average of Salary 2008 and average of Bonus 2008, broken down by Company Type. Comment on any striking results in the chart.
 - b. In the Data sheet, create a new column, Total 2008, which is the sum of Salary 2008 and Bonus 2008.
- Then create two pivot tables and corresponding pivot charts on a single sheet. The first should show the counts of CEOs broken down by Company Type, and the second should simultaneously show the average of Total 2008, the minimum of Total 2008, and the maximum of Total 2008, all broken down by Company Type. Comment on any striking results in these charts.
42. One pivot table element we didn't explain is a **calculated item**. This is usually a new category for some categorical variable that is created from existing categories. It is easiest to learn from an example. Open the file [ElecMart Sales.xlsx](#) from this section, create a pivot table, and put Day in the Rows area. Proceed as follows to create two new categories, Weekday and Weekend.
 - a. Select any day and select Calculated Item from the Formulas dropdown list on the PivotTable Tools Options ribbon. This will open a dialog box. Enter Weekend in the Name box and enter the formula `=Sat+Sun` in the formula box. (You can double-click the items in the Items list to help build this formula.) When you click OK, you will see Weekend in the pivot table.
 - b. Do it yourself. Create another calculated item, Weekday, for Mon through Fri.
 - c. Filter out all of the individual days from the row area, so that only Weekday and Weekend remain, and then find the sum of Total Cost for these two new categories. How can you check whether these sums are what you think they should be? (Notes about calculated items: First, if you have Weekend, Weekday, and some individual days showing in the Rows area, the sum of Total Cost will double-count these individual days, so be careful about this. Second, be aware that if you create a calculated item from some variable such as Day, you are no longer allowed to drag that variable to the Filters area.)
43. Building on the previous problem, another pivot table element we didn't explain is a **calculated field**. This is usually a new numerical variable built from numerical variables that can be summarized in the Values area. It acts somewhat like a new column in the spreadsheet data, but there is an important difference. Again, it is easiest to learn from an example. Open the file [ElecMart Sales.xlsx](#) and follow the instructions below.
 - a. Create a new column in the data, CostPerItem, which is Total Cost divided by Items Ordered. Then create a pivot table and find the average of CostPerItem, broken down by Region. You should find averages such as \$50.41 for the Midwest. Explain exactly how this value was calculated. Would such an average be of much interest to a manager at ElecMart? Why or why not?
 - b. Select any average in the pivot table and then select Calculated Field from the Formulas dropdown list