DS342 - Data Analytics

Lecture 3

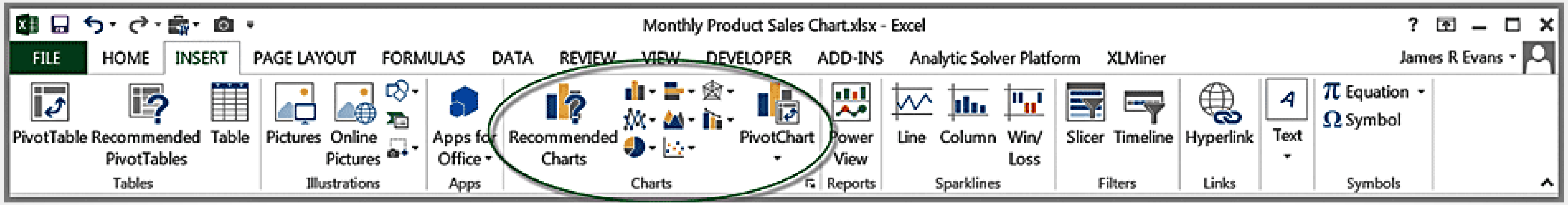
Describing the Distribution of a Single Variable

Part II

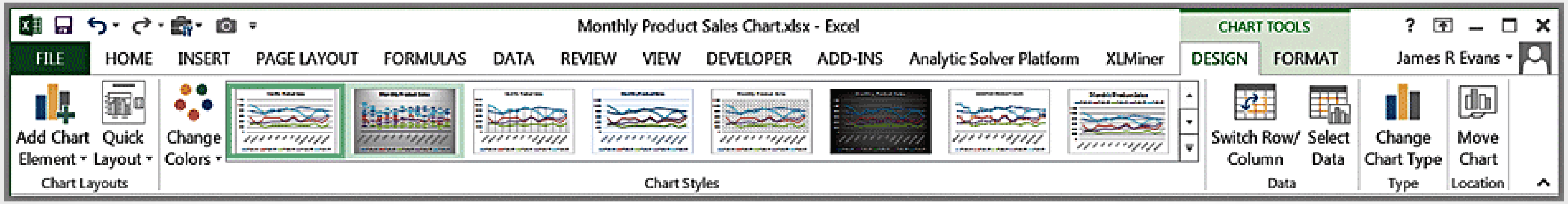
Dr. Marwa Sabry

Creating Charts in Microsoft Excel

- Select the *Insert* tab.
- Highlight the data.
- Click on chart type, then subtype.



- Use *Chart Tools* to customize.



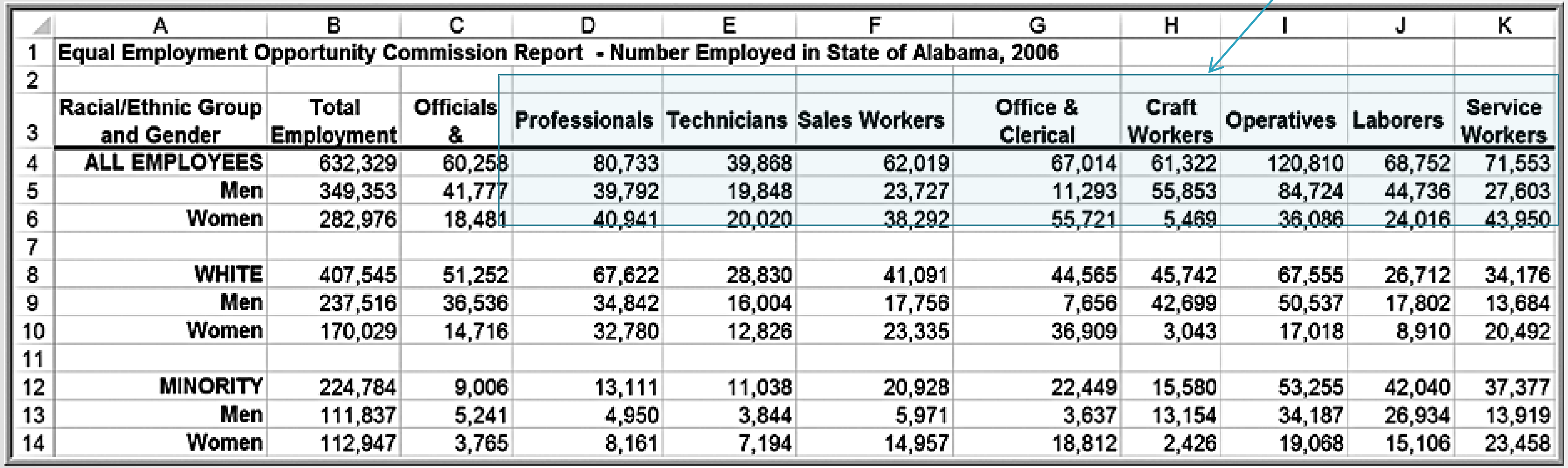
Column and Bar Charts

- Excel distinguishes between vertical and horizontal bar charts, calling the former *column charts* and the latter *bar charts*.
  - A clustered column chart compares values across categories using vertical rectangles;
  - a stacked column chart displays the contribution of each value to the total by stacking the rectangles;
  - a 100% stacked column chart compares the percentage that each value contributes to a total.
- Column and bar charts are useful for comparing categorical or ordinal data, for illustrating differences between sets of values, and for showing proportions or percentages of a whole.

Example 3.2: Creating a Column Chart

Highlight the range C3:K6, which includes the headings and data for each category. Click on the *Column Chart* button and then on the first chart type in the list (a clustered column chart).

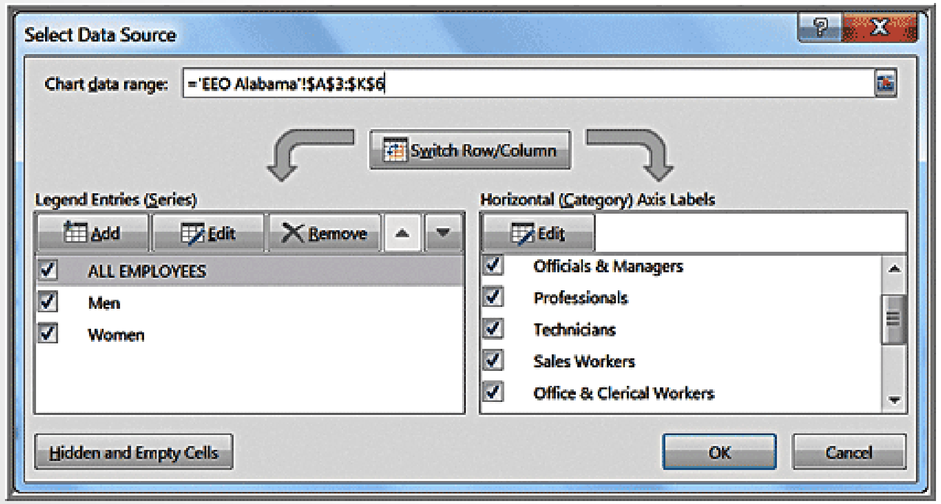
Highlighted Cells



	A	B	C	D	E	F	G	H	I	J	K
1	Equal Employment Opportunity Commission Report - Number Employed in State of Alabama, 2006										
2											
3	Racial/Ethnic Group and Gender	Total Employment	Officials &	Professionals	Technicians	Sales Workers	Office & Clerical	Craft Workers	Operatives	Laborers	Service Workers
4	ALL EMPLOYEES	632,329	60,258	80,733	39,868	62,019	67,014	61,322	120,810	68,752	71,553
5	Men	349,353	41,777	39,792	19,848	23,727	11,293	55,853	84,724	44,736	27,603
6	Women	282,976	18,481	40,941	20,020	38,292	55,721	5,469	36,086	24,016	43,950
7											
8	WHITE	407,545	51,252	67,622	28,830	41,091	44,565	45,742	67,555	26,712	34,176
9	Men	237,516	36,536	34,842	16,004	17,756	7,656	42,699	50,537	17,802	13,684
10	Women	170,029	14,716	32,780	12,826	23,335	36,909	3,043	17,018	8,910	20,492
11											
12	MINORITY	224,784	9,006	13,111	11,038	20,928	22,449	15,580	53,255	42,040	37,377
13	Men	111,837	5,241	4,950	3,844	5,971	3,637	13,154	34,187	26,934	13,919
14	Women	112,947	3,765	8,161	7,194	14,957	18,812	2,426	19,068	15,106	23,458

Example 3.2: Creating a Column Chart

To add a title, click on the first icon in the *Chart Layouts* group. Click on “Chart Title” in the chart and change it to “EEO Employment Report—Alabama.” The names of the data series can be changed by clicking on the *Select Data* button in the *Data* group of the *Design* tab. In the *Select Data Source* dialog (see below), click on “Series1” and then the *Edit* button. Enter the name of the data series, in this case “All Employees.” Change the names of the other data series to “Men” and “Women” in a similar fashion.



Select Data Source

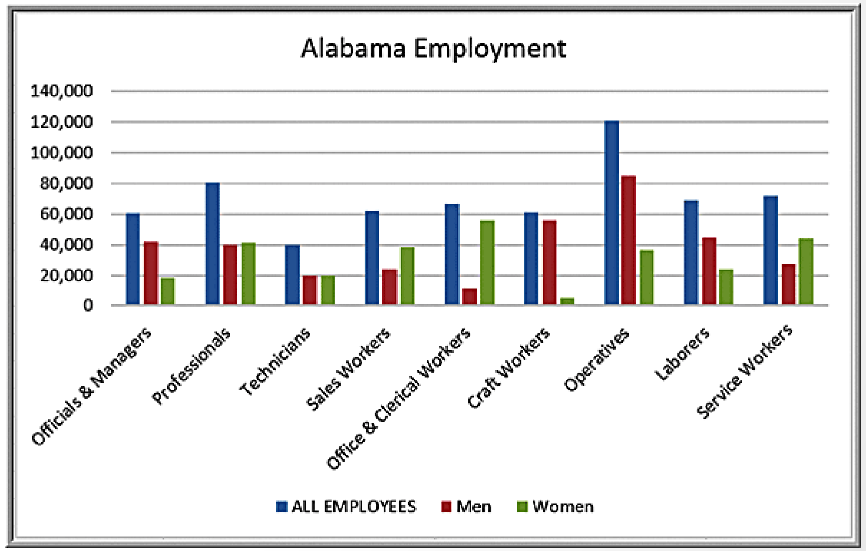
Chart data range: ='EEO Alabama'!$A$3:$K$6

Legend Entries (Series)

- ☑ ALL EMPLOYEES
- ☑ Men
- ☑ Women

Horizontal (Category) Axis Labels

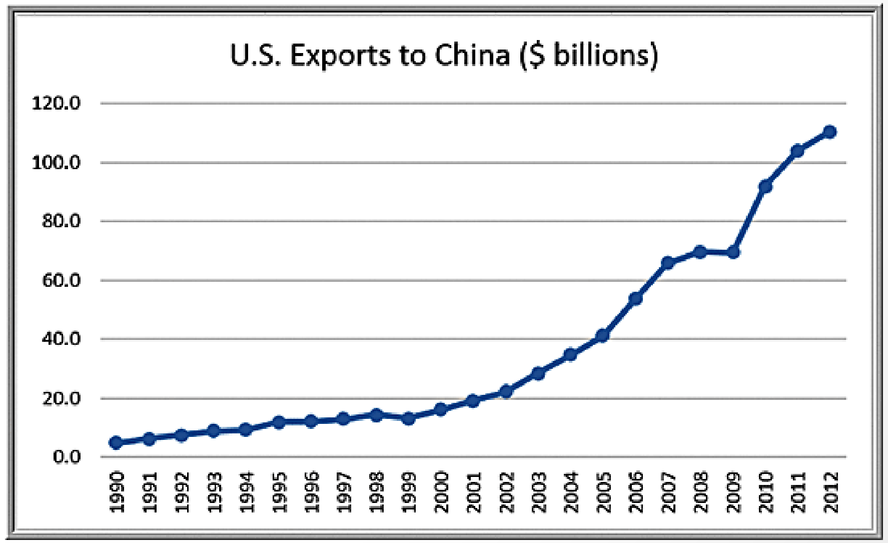
- ☑ Officials & Managers
- ☑ Professionals
- ☑ Technicians
- ☑ Sales Workers
- ☑ Office & Clerical Workers



Line Charts

- Line charts provide a useful means for displaying data over time.
  - You may plot multiple data series in line charts; however, they can be difficult to interpret if the magnitude of the data values differs greatly. In that case, it would be advisable to create separate charts for each data series.

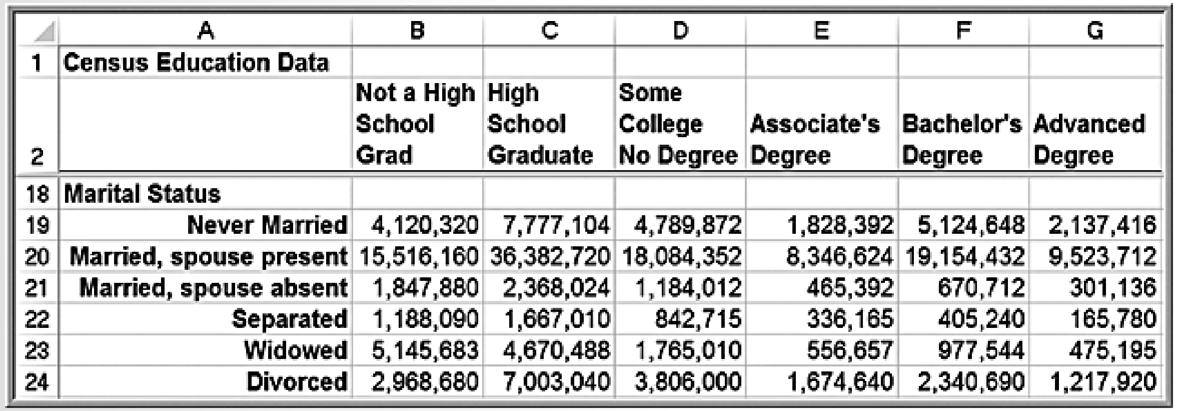
Example 3.3: A Line Chart for China Export Data



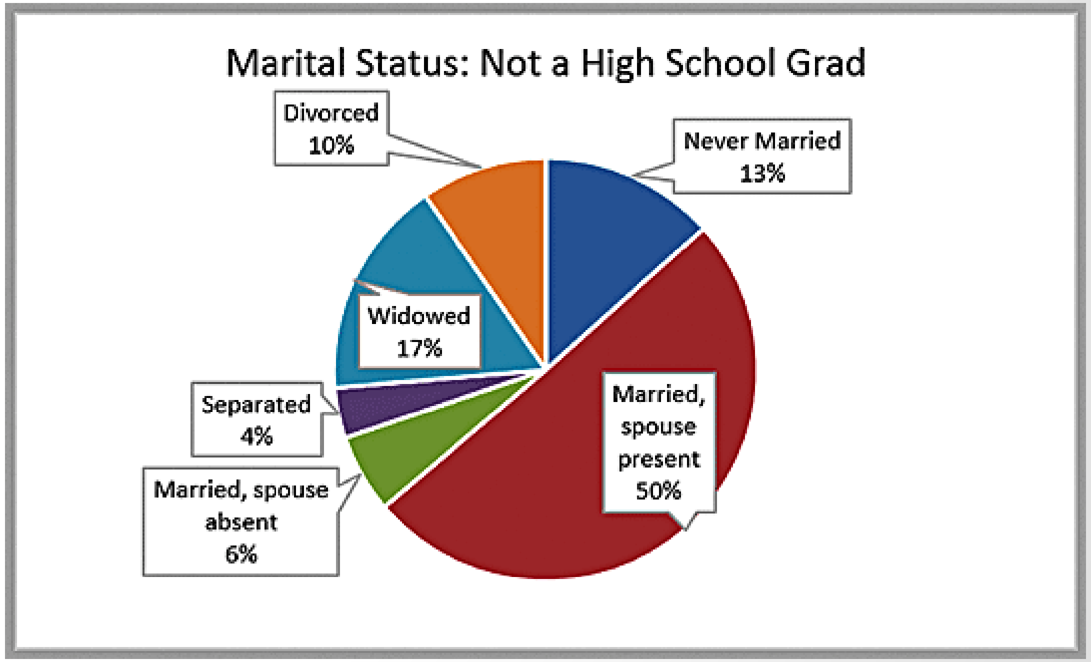
Pie Charts

- A pie chart displays this by partitioning a circle into pie-shaped areas showing the relative proportion.

Example 3.4: A Pie Chart for Census Data

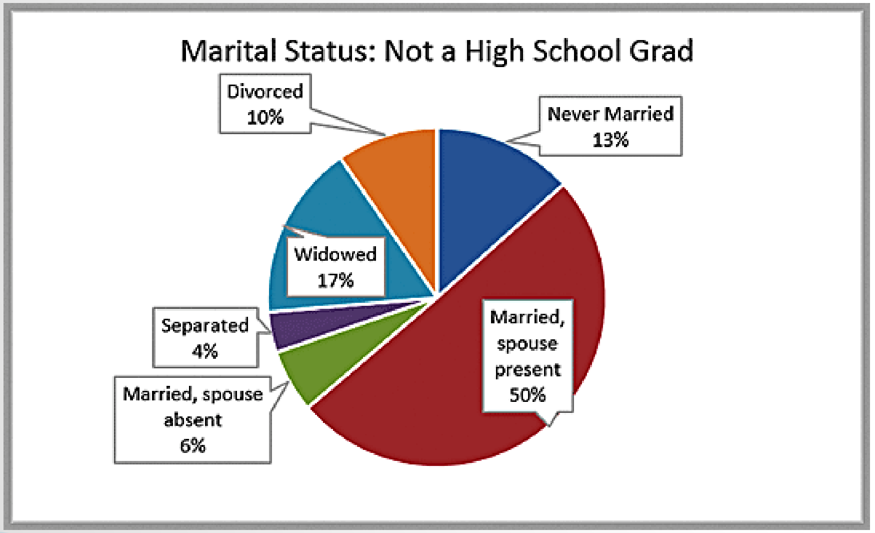


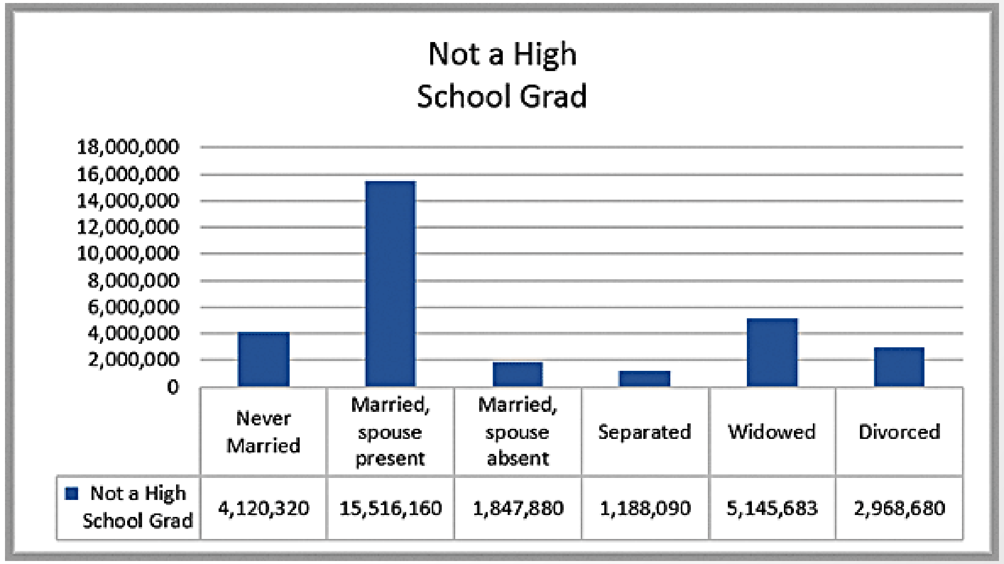
	A	B	C	D	E	F	G
1	Census Education Data						
2		Not a High School Grad	High School Graduate	Some College No Degree	Associate's Degree	Bachelor's Degree	Advanced Degree
18	Marital Status						
19	Never Married	4,120,320	7,777,104	4,789,872	1,828,392	5,124,648	2,137,416
20	Married, spouse present	15,516,160	36,382,720	18,084,352	8,346,624	19,154,432	9,523,712
21	Married, spouse absent	1,847,880	2,368,024	1,184,012	465,392	670,712	301,136
22	Separated	1,188,090	1,667,010	842,715	336,165	405,240	165,780
23	Widowed	5,145,683	4,670,488	1,765,010	556,657	977,544	475,195
24	Divorced	2,968,680	7,003,040	3,806,000	1,674,640	2,340,690	1,217,920



Pie Charts

- Data visualization professionals don't recommend using pie charts. In a pie chart, it is difficult to compare the relative sizes of areas; however, the bars in the column chart can easily be compared to determine relative ratios of the data.
  - If you do use pie charts, restrict them to small numbers of categories, always ensure that the numbers add to 100%, and use labels to display the group names and actual percentages. Avoid three-dimensional (3-D) pie charts—especially those that are rotated—and keep them simple.

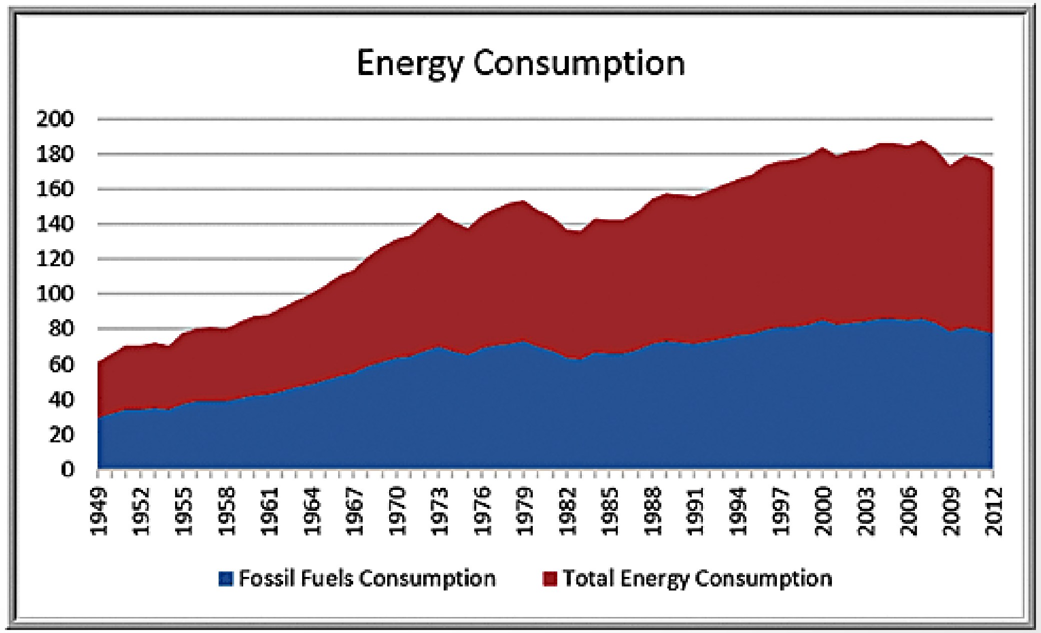




Area Charts

- An area chart combines the features of a pie chart with those of line charts.
  - Area charts present more information than pie or line charts alone but may clutter the observer’s mind with too many details if too many data series are used; thus, they should be used with care.

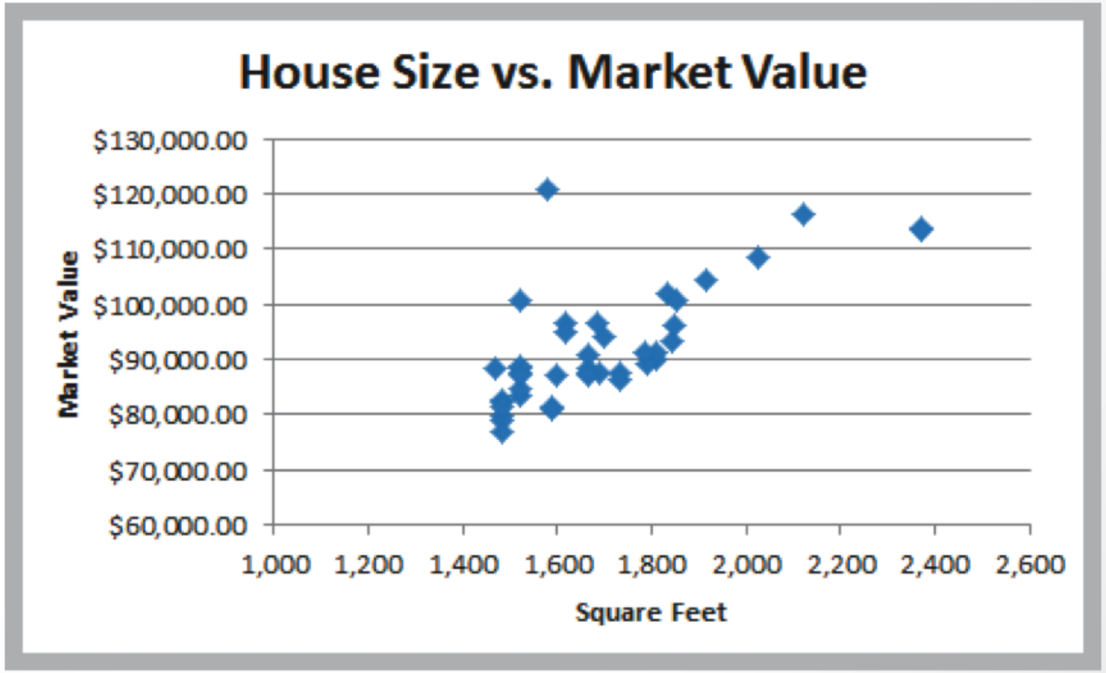
Example 3.5: An Area Chart for Energy Consumption



Scatter Charts

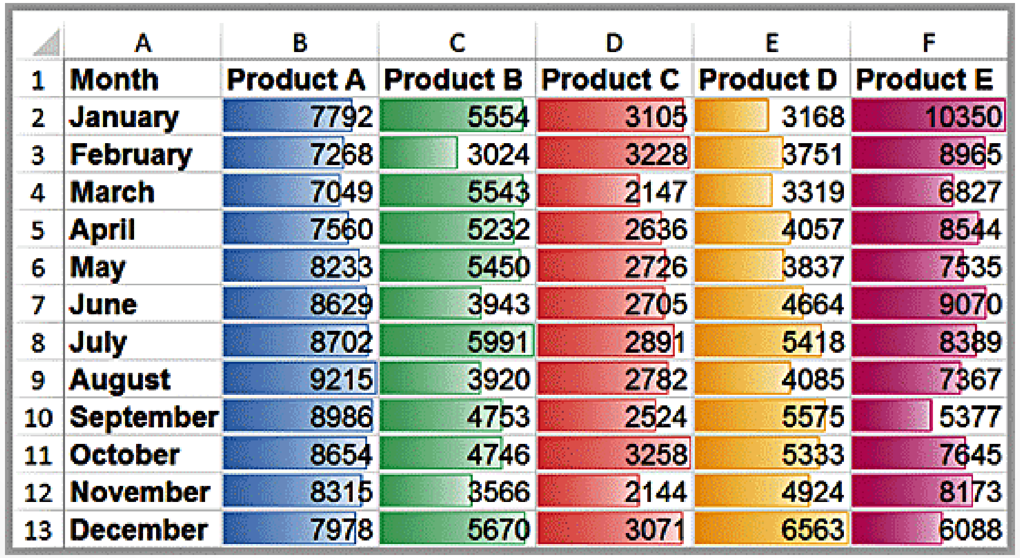
- Scatter charts show the relationship between two variables. To construct a scatter chart, we need observations that consist of *pairs* of variables.

Example 3.6: A Scatter Chart for Real Estate Data



Example 3.8: Data Visualization through Conditional Formatting

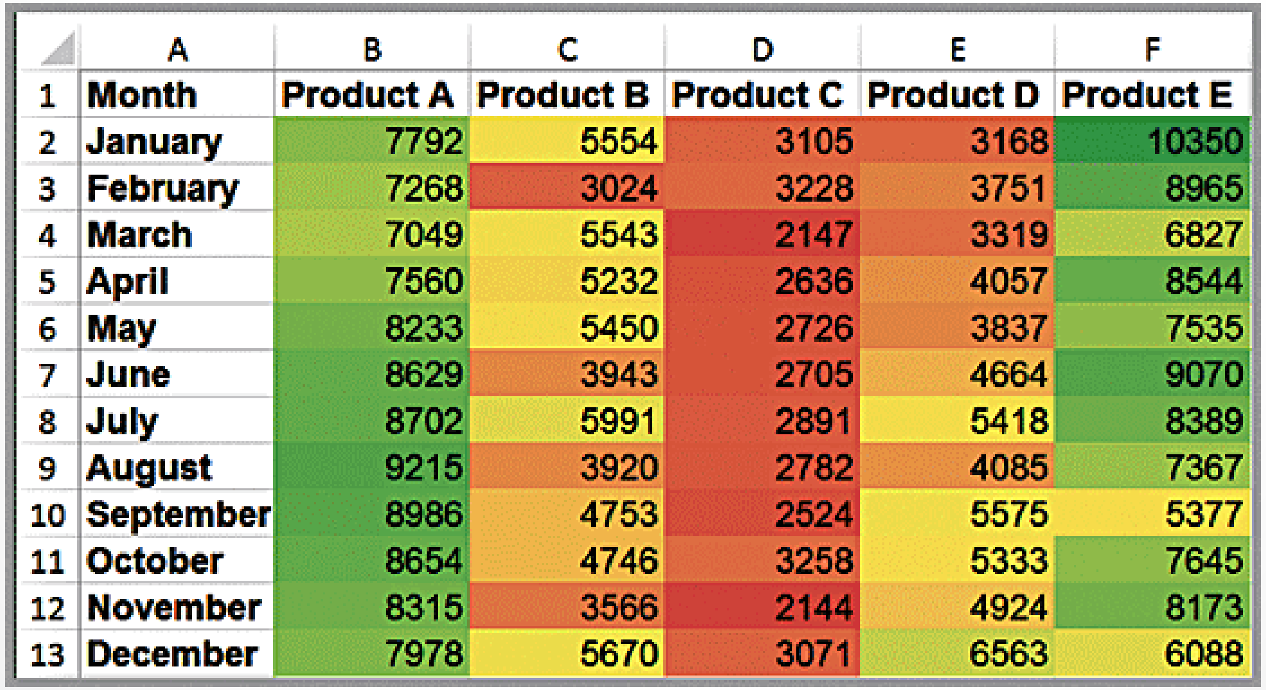
- **Data bars** display colored bars that are scaled to the magnitude of the data values (similar to a bar chart) but placed directly within the cells of a range.
  - Highlight the data in each column, click the *Conditional Formatting* button in the *Styles* group within the *Home* tab, select *Data Bars*, and choose the fill option and color.



	A	B	C	D	E	F
1	Month	Product A	Product B	Product C	Product D	Product E
2	January	7792	5554	3105	3168	10350
3	February	7268	3024	3228	3751	8965
4	March	7049	5543	2147	3319	6827
5	April	7560	5232	2636	4057	8544
6	May	8233	5450	2726	3837	7535
7	June	8629	3943	2705	4664	9070
8	July	8702	5991	2891	5418	8389
9	August	9215	3920	2782	4085	7367
10	September	8986	4753	2524	5575	5377
11	October	8654	4746	3258	5333	7645
12	November	8315	3566	2144	4924	8173
13	December	7978	5670	3071	6563	6088

Example 3.8: Data Visualization through Conditional Formatting

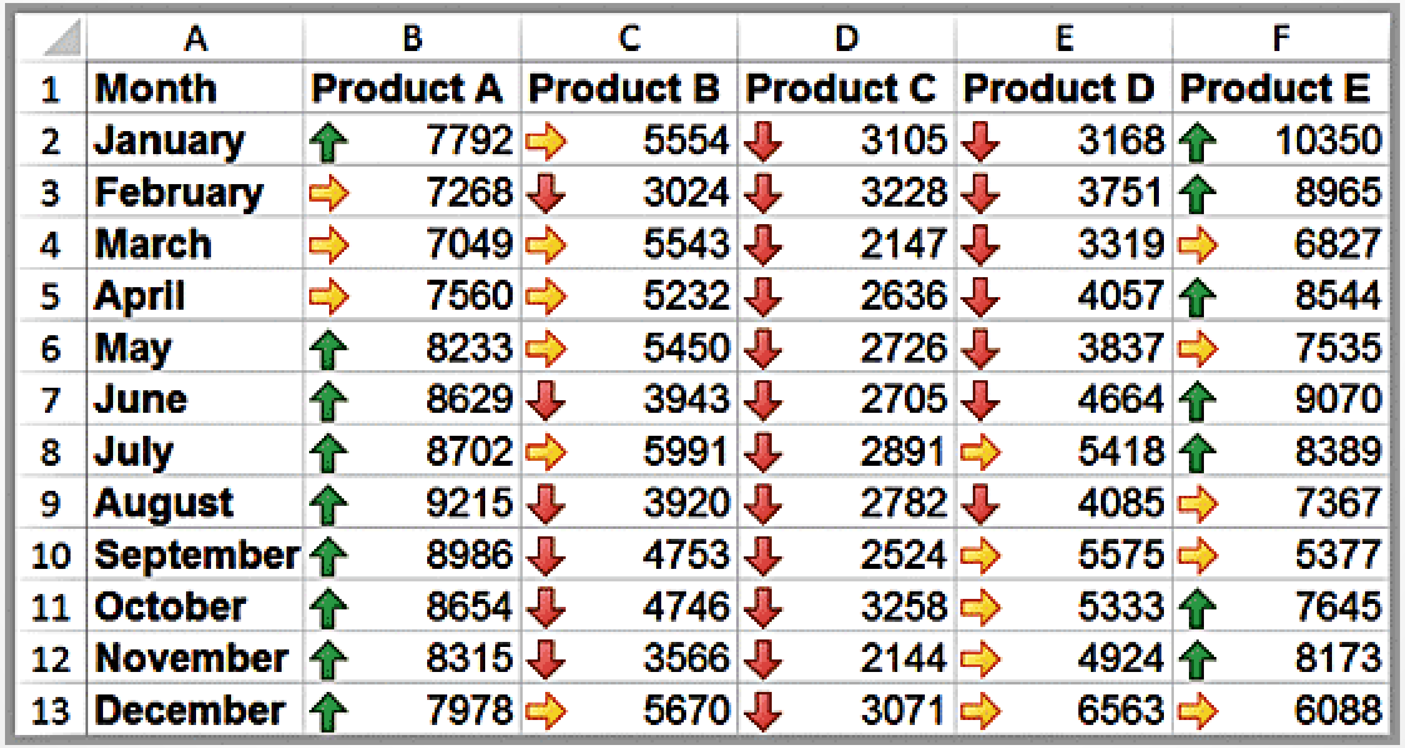
- **Color scales** shade cells based on their numerical value using a color palette.
  - Color-coding of quantitative data is commonly called a **heatmap**.



	A	B	C	D	E	F
1	**Month**	**Product A**	**Product B**	**Product C**	**Product D**	**Product E**
2	**January**	7792	5554	3105	3168	10350
3	**February**	7268	3024	3228	3751	8965
4	**March**	7049	5543	2147	3319	6827
5	**April**	7560	5232	2636	4057	8544
6	**May**	8233	5450	2726	3837	7535
7	**June**	8629	3943	2705	4664	9070
8	**July**	8702	5991	2891	5418	8389
9	**August**	9215	3920	2782	4085	7367
10	**September**	8986	4753	2524	5575	5377
11	**October**	8654	4746	3258	5333	7645
12	**November**	8315	3566	2144	4924	8173
13	**December**	7978	5670	3071	6563	6088

Example 3.8: Data Visualization through Conditional Formatting

- **Icon sets** provide similar information using various symbols such as arrows or stoplight colors.



	A	B	C	D	E	F
1	Month	Product A	Product B	Product C	Product D	Product E
2	January	↑ 7792	→ 5554	↓ 3105	↓ 3168	↑ 10350
3	February	→ 7268	↓ 3024	↓ 3228	↓ 3751	↑ 8965
4	March	→ 7049	→ 5543	↓ 2147	↓ 3319	→ 6827
5	April	→ 7560	→ 5232	↓ 2636	↓ 4057	↑ 8544
6	May	↑ 8233	→ 5450	↓ 2726	↓ 3837	→ 7535
7	June	↑ 8629	↓ 3943	↓ 2705	↓ 4664	↑ 9070
8	July	↑ 8702	→ 5991	↓ 2891	→ 5418	↑ 8389
9	August	↑ 9215	↓ 3920	↓ 2782	↓ 4085	→ 7367
10	September	↑ 8986	↓ 4753	↓ 2524	→ 5575	→ 5377
11	October	↑ 8654	↓ 4746	↓ 3258	→ 5333	↑ 7645
12	November	↑ 8315	↓ 3566	↓ 2144	→ 4924	↑ 8173
13	December	↑ 7978	→ 5670	↓ 3071	→ 6563	→ 6088

Charts for Numerical Variables

- There are many graphical ways to indicate the distribution of a numerical variable.
  - For cross-sectional variables:
    - Histograms
    - Boxplots
  - For time series variables:
    - Time series graphs

Histograms

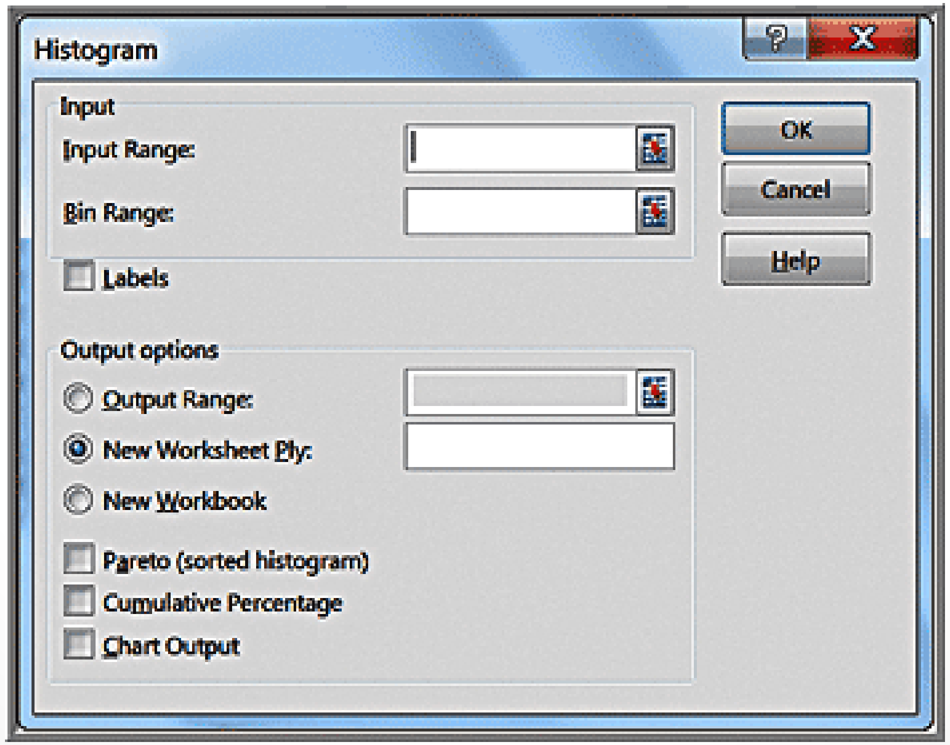
- A **histogram** is the most common type of chart for showing the distribution of a numerical variable.
  - It is based on binning the variable—that is, dividing it up into discrete categories.
  - It is a column chart of the counts in the various categories (with no gaps between the vertical bars).
- A histogram is great for showing the shape of a distribution—whether the distribution is symmetric or skewed in one direction.

Excel *Histogram* Tool

- A graphical depiction of a frequency distribution for numerical data in the form of a column chart is called a **histogram**.
- Frequency distributions and histograms can be created using the *Analysis Toolpak* in Excel.
  - Click the *Data Analysis* tools button in the *Analysis* group under the *Data* tab in the Excel menu bar and select *Histogram* from the list.

Histogram Dialog

- Specify the *Input Range* corresponding to the data. If you include the column header, then also check the *Labels* box so Excel knows that the range contains a label. The *Bin Range* defines the groups (Excel calls these "bins") used for the frequency distribution.



Histogram

Input

Input Range:

Bin Range:

□ Labels

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

□ Pareto (sorted histogram)

□ Cumulative Percentage

□ Chart Output

Using Bin Ranges

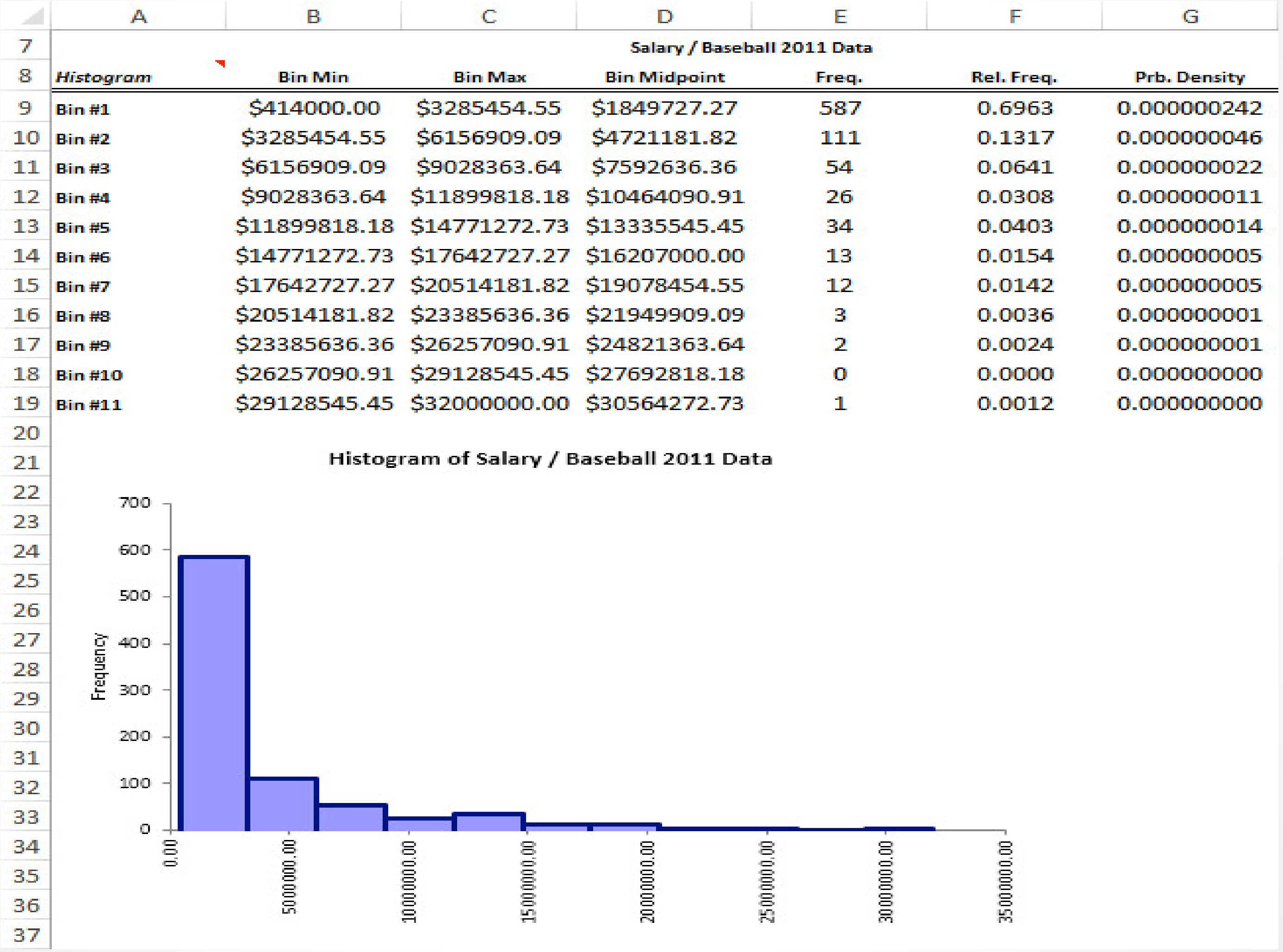
- If you do not specify a *Bin Range*, Excel will automatically determine bin values for the frequency distribution and histogram, which often results in a rather poor choice.

- If you have discrete values, set up a column of these values in your spreadsheet for the bin range and specify this range in the *Bin Range* field.

Example 2.3 (Continued): Baseball Salaries 2011.xlsx (slide 1 of 2)

- **Objective**: To see the shape of the salary distribution through a histogram.
- **Solution**: It is possible to create a histogram with Excel tools only—but it is a tedious process.
  - The resulting table of counts is usually called a **frequency table**.
  - The counts are called **frequencies**.
- It is much easier to create a histogram with Data Analysis add-in.

Example 2.3 (Continued): Baseball Salaries 2011.xlsx (slide 2 of 2)



Salary / Baseball 2011 Data

Histogram	Bin Min	Bin Max	Bin Midpoint	Freq.	Rel. Freq.	Prb. Density
Bin #1	$414000.00	$3285454.55	$1849727.27	587	0.6963	0.000000242
Bin #2	$3285454.55	$6156909.09	$4721181.82	111	0.1317	0.000000046
Bin #3	$6156909.09	$9028363.64	$7592636.36	54	0.0641	0.000000022
Bin #4	$9028363.64	$11899818.18	$10464090.91	26	0.0308	0.000000011
Bin #5	$11899818.18	$14771272.73	$13335545.45	34	0.0403	0.000000014
Bin #6	$14771272.73	$17642727.27	$16207000.00	13	0.0154	0.000000005
Bin #7	$17642727.27	$20514181.82	$19078454.55	12	0.0142	0.000000005
Bin #8	$20514181.82	$23385636.36	$21949909.09	3	0.0036	0.000000001
Bin #9	$23385636.36	$26257090.91	$24821363.64	2	0.0024	0.000000001
Bin #10	$26257090.91	$29128545.45	$27692818.18	0	0.0000	0.000000000
Bin #11	$29128545.45	$32000000.00	$30564272.73	1	0.0012	0.000000000

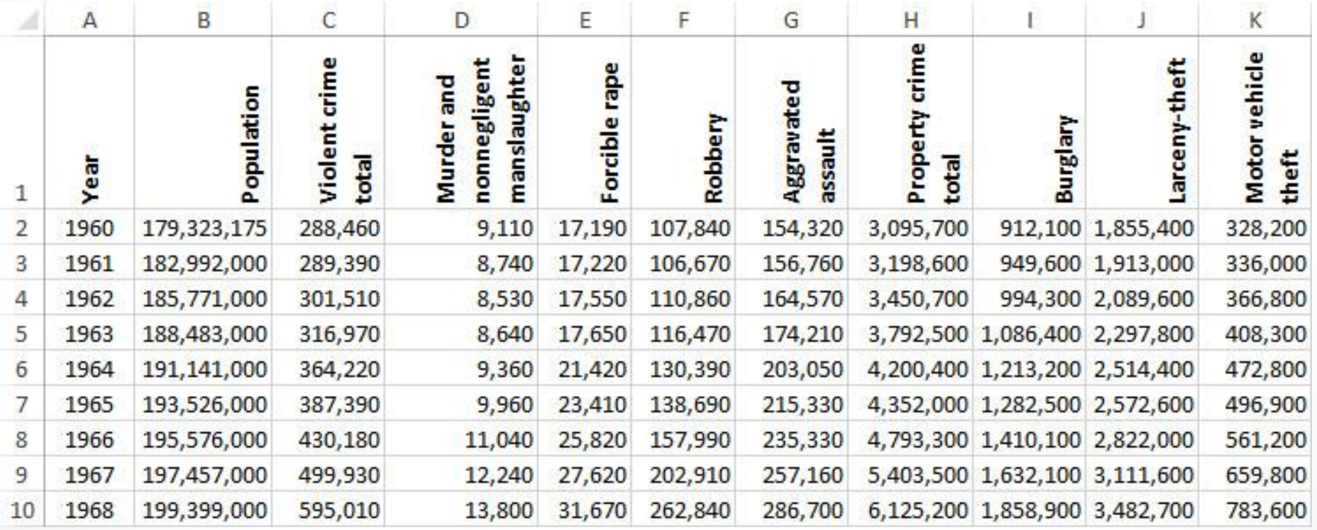
Histogram of Salary / Baseball 2011 Data

Time Series Data

- Our main interest in time series variables is how they change over time, and this information is lost in traditional summary measures and in histograms or box plots.
- For time series data, a **time series graph** is used. This is a graph of the values of one or more time series, using time on the horizontal axis.
  - This is always the place to start a time series analysis.

Example 2.5:
Crime in US.xlsx (slide 1 of 3)

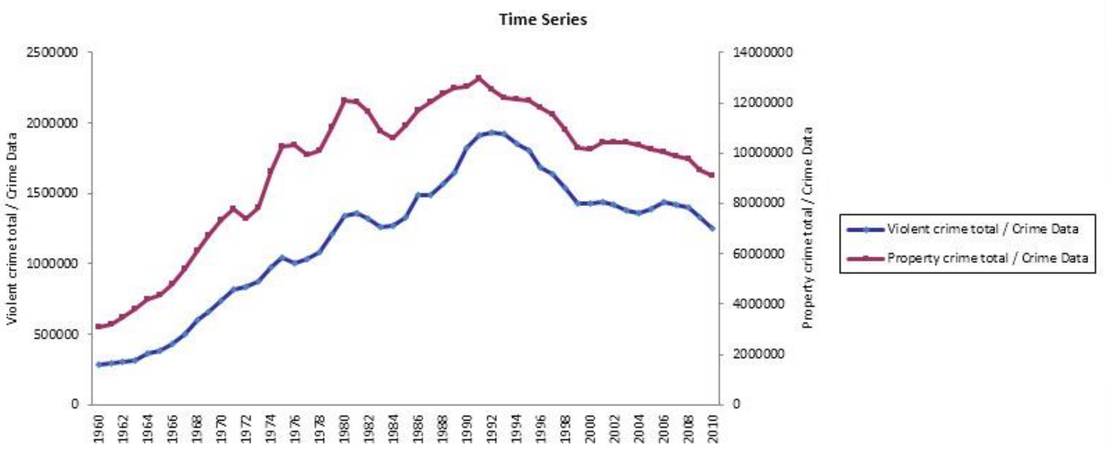
- **Objective**: To see how time series graphs help to detect trends in crime data.
- **Solution**: Data set contains annual data on violent and property crimes for the years 1960 to 2010.



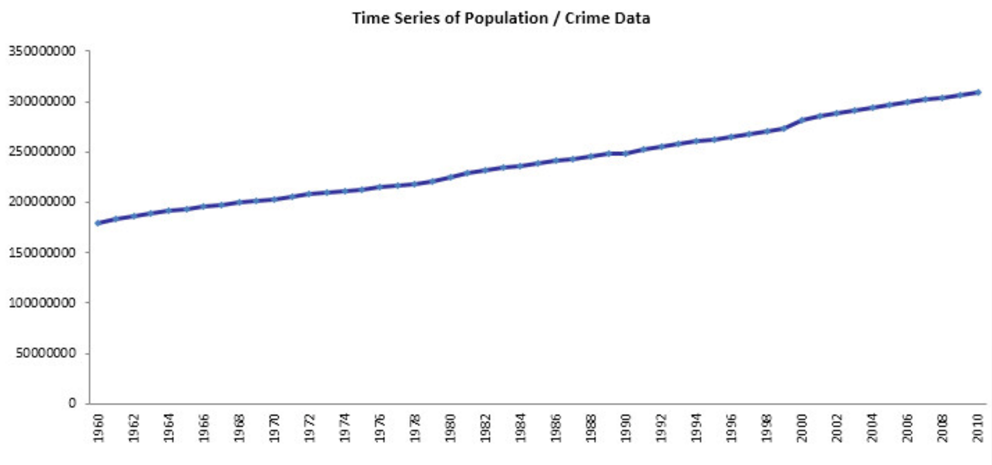
	A	B	C	D	E	F	G	H	I	J	K
1	Year	Population	Violent crime total	Murder and nonnegligent manslaughter	Forcible rape	Robbery	Aggravated assault	Property crime total	Burglary	Larceny-theft	Motor vehicle theft
2	1960	179,323,175	288,460	9,110	17,190	107,840	154,320	3,095,700	912,100	1,855,400	328,200
3	1961	182,992,000	289,390	8,740	17,220	106,670	156,760	3,198,600	949,600	1,913,000	336,000
4	1962	185,771,000	301,510	8,530	17,550	110,860	164,570	3,450,700	994,300	2,089,600	366,800
5	1963	188,483,000	316,970	8,640	17,650	116,470	174,210	3,792,500	1,086,400	2,297,800	408,300
6	1964	191,141,000	364,220	9,360	21,420	130,390	203,050	4,200,400	1,213,200	2,514,400	472,800
7	1965	193,526,000	387,390	9,960	23,410	138,690	215,330	4,352,000	1,282,500	2,572,600	496,900
8	1966	195,576,000	430,180	11,040	25,820	157,990	235,330	4,793,300	1,410,100	2,822,000	561,200
9	1967	197,457,000	499,930	12,240	27,620	202,910	257,160	5,403,500	1,632,100	3,111,600	659,800
10	1968	199,399,000	595,010	13,800	31,670	262,840	286,700	6,125,200	1,858,900	3,482,700	783,600

Example 2.5: Crime in US.xlsx (slide 2 of 3)

Total Violent and Property Crimes

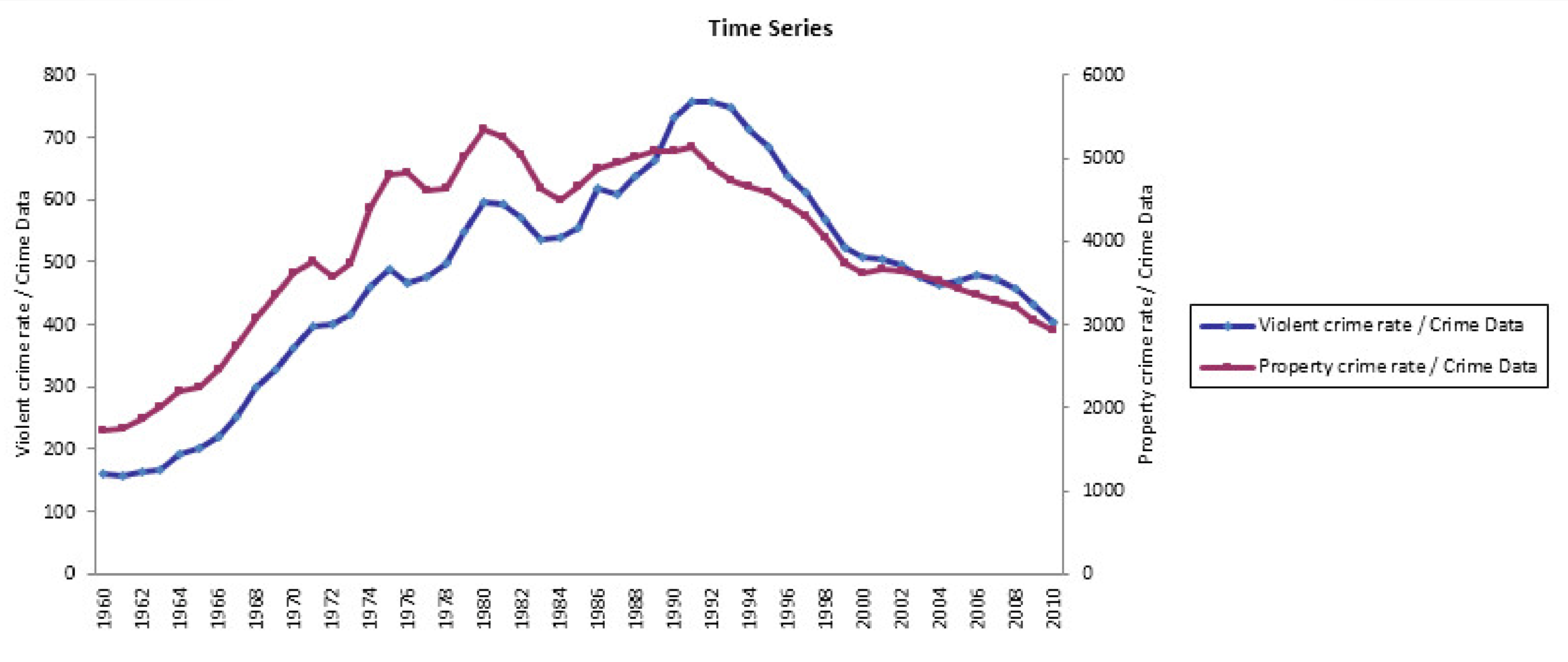


Population Totals



Example 2.5: Crime in US.xlsx (slide 3 of 3)

Violent and Property Crime Rates

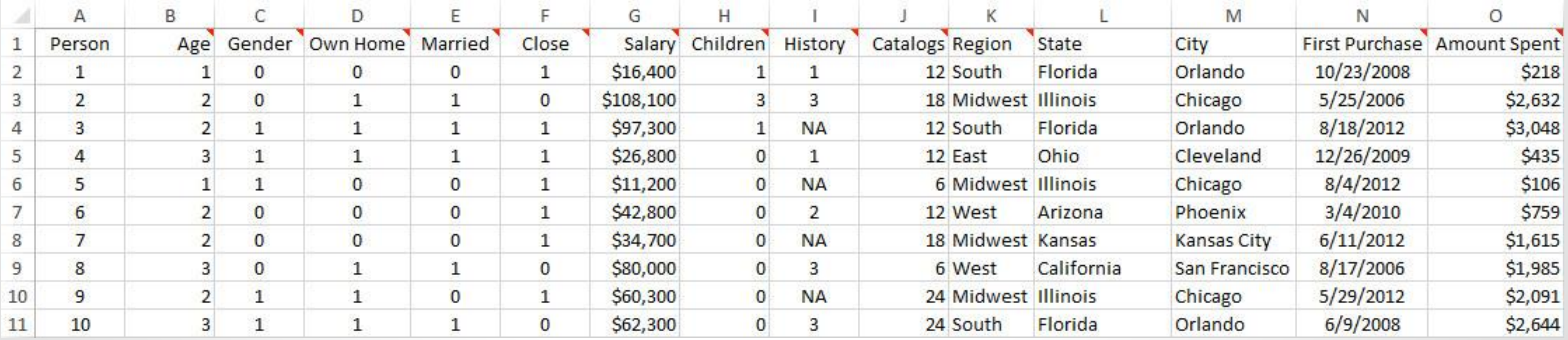


Excel Tables for Filtering, Sorting, and Summarizing

- Tables are a tool introduced in Excel 2007.
- You now have the ability to designate a rectangular data set as a table and then employ a number of powerful tools for analyzing tables.
- These tools include:
  - Filtering
  - Sorting
  - Summarizing

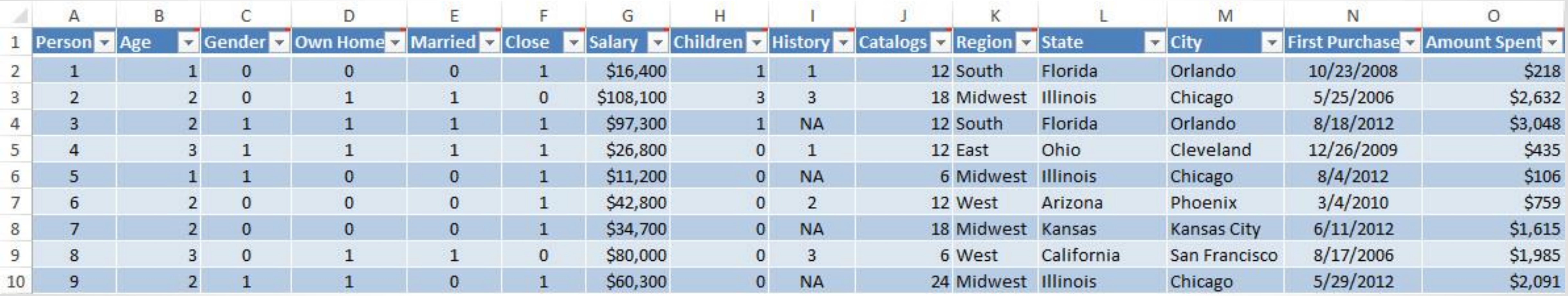
Example 2.7: Catalog Marketing.xlsx (slide 1 of 2)

- **Objective**: To illustrate Excel tables for analyzing the HyTex data.
- **Solution**: Data set contains data on 1000 customers of HyTex, a fictional direct marketing company.
- Designate the data set as a table by selecting any cell in the data set and clicking the Table button on the Insert ribbon.
- Use the dropdown arrows next to the variable names to filter in many different ways.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Person	Age	Gender	Own Home	Married	Close	Salary	Children	History	Catalogs	Region	State	City	First Purchase	Amount Spent
2	1	1	0	0	0	1	$16,400	1	1	12	South	Florida	Orlando	10/23/2008	$218
3	2	2	0	1	1	0	$108,100	3	3	18	Midwest	Illinois	Chicago	5/25/2006	$2,632
4	3	2	1	1	1	1	$97,300	1	NA	12	South	Florida	Orlando	8/18/2012	$3,048
5	4	3	1	1	1	1	$26,800	0	1	12	East	Ohio	Cleveland	12/26/2009	$435
6	5	1	1	0	0	1	$11,200	0	NA	6	Midwest	Illinois	Chicago	8/4/2012	$106
7	6	2	0	0	0	1	$42,800	0	2	12	West	Arizona	Phoenix	3/4/2010	$759
8	7	2	0	0	0	1	$34,700	0	NA	18	Midwest	Kansas	Kansas City	6/11/2012	$1,615
9	8	3	0	1	1	0	$80,000	0	3	6	West	California	San Francisco	8/17/2006	$1,985
10	9	2	1	1	0	1	$60,300	0	NA	24	Midwest	Illinois	Chicago	5/29/2012	$2,091
11	10	3	1	1	1	0	$62,300	0	3	24	South	Florida	Orlando	6/9/2008	$2,644

Example 2.7: Catalog Marketing.xlsx (slide 2 of 2)



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Person	Age	Gender	Own Home	Married	Close	Salary	Children	History	Catalogs	Region	State	City	First Purchase	Amount Spent
2	1	1	0	0	0	1	$16,400	1	1	12	South	Florida	Orlando	10/23/2008	$218
3	2	2	0	1	1	0	$108,100	3	3	18	Midwest	Illinois	Chicago	5/25/2006	$2,632
4	3	2	1	1	1	1	$97,300	1	NA	12	South	Florida	Orlando	8/18/2012	$3,048
5	4	3	1	1	1	1	$26,800	0	1	12	East	Ohio	Cleveland	12/26/2009	$435
6	5	1	1	0	0	1	$11,200	0	NA	6	Midwest	Illinois	Chicago	8/4/2012	$106
7	6	2	0	0	0	1	$42,800	0	2	12	West	Arizona	Phoenix	3/4/2010	$759
8	7	2	0	0	0	1	$34,700	0	NA	18	Midwest	Kansas	Kansas City	6/11/2012	$1,615
9	8	3	0	1	1	0	$80,000	0	3	6	West	California	San Francisco	8/17/2006	$1,985
10	9	2	1	1	0	1	$60,300	0	NA	24	Midwest	Illinois	Chicago	5/29/2012	$2,091

Filtering

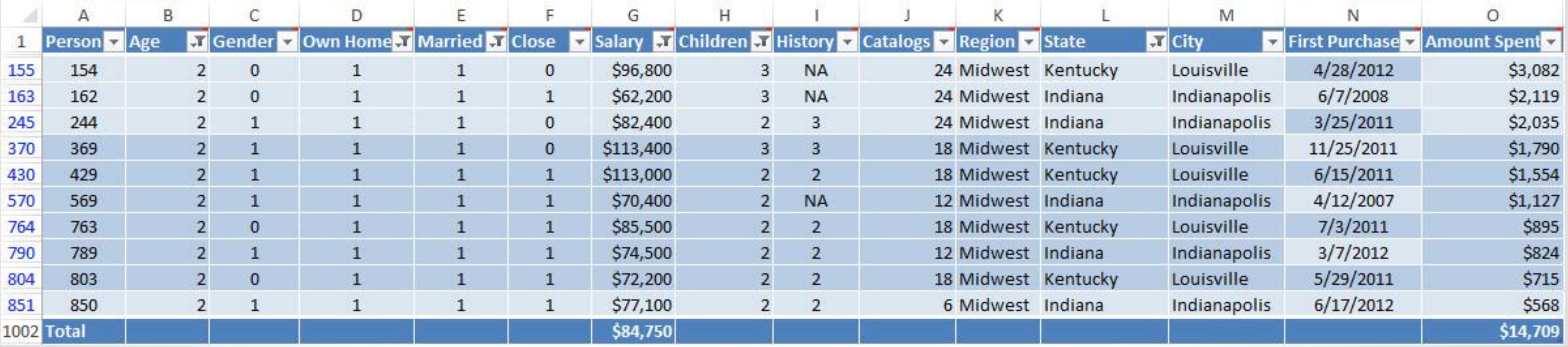
- Finding records that match particular criteria is called *filtering*.
- One way to filter is to create an Excel table, which automatically provides dropdown arrows next to the field names that allow you to filter.
- There are also three ways to filter on any rectangular data set with variable names:
  1. Use the Filter button from the Sort & Filter dropdown list on the Home ribbon.
  2. Use the Filter button from the Sort & Filter group on the Data ribbon.
  3. Right-click any cell in the data set and select Filter. You get several options, the most popular of which is Filter by Selected Cell's Value.

Example 2.7 (Continued): Catalog Marketing.xlsx (slide 1 of 2)

- **Objective**: To investigate the types of filters that can be applied to the HyTex data.
- **Solution**: There is almost no limit to the filters you can apply, but here are a few possibilities:
  - Filter on one or more values in a field.
  - Filter on more than one field.
  - Filter on a continuous numerical field.
  - *Top 10* and *Above/Below Average* filters.
  - Filter on a text field.
  - Filter on a date field.
  - Filter on color or icon.
  - Use a custom filter.

Example 2.7 (Continued): Catalog Marketing.xlsx (slide 2 of 2)

Results from a Typical Filter



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Person	Age	Gender	Own Home	Married	Close	Salary	Children	History	Catalogs	Region	State	City	First Purchase	Amount Spent
155	154	2	0	1	1	0	$96,800	3	NA	24	Midwest	Kentucky	Louisville	4/28/2012	$3,082
163	162	2	0	1	1	1	$62,200	3	NA	24	Midwest	Indiana	Indianapolis	6/7/2008	$2,119
245	244	2	1	1	1	0	$82,400	2	3	24	Midwest	Indiana	Indianapolis	3/25/2011	$2,035
370	369	2	1	1	1	0	$113,400	3	3	18	Midwest	Kentucky	Louisville	11/25/2011	$1,790
430	429	2	1	1	1	1	$113,000	2	2	18	Midwest	Kentucky	Louisville	6/15/2011	$1,554
570	569	2	1	1	1	1	$70,400	2	NA	12	Midwest	Indiana	Indianapolis	4/12/2007	$1,127
764	763	2	0	1	1	1	$85,500	2	2	18	Midwest	Kentucky	Louisville	7/3/2011	$895
790	789	2	1	1	1	1	$74,500	2	2	12	Midwest	Indiana	Indianapolis	3/7/2012	$824
804	803	2	0	1	1	1	$72,200	2	2	18	Midwest	Kentucky	Louisville	5/29/2011	$715
851	850	2	1	1	1	1	$77,100	2	2	6	Midwest	Indiana	Indianapolis	6/17/2012	$568
1002	Total						$84,750								$14,709

Properties of Designated Tables

- A number of table styles are available for making the table attractive. You can experiment with these, including the various table styles and table style options. Note the dropdown list in the Table Styles group. It gives you many more styles than the ones originally visible.

- In the Tools group, you can click Convert to Range. This undesignates the range as a table (and the dropdown arrows disappear).

- A particularly useful option is the Total Row in the Table Style Options group. If you check this, a new row is appended to the bottom of the table. It creates a sum formula in the rightmost column. This sum includes ***only*** the nonhidden rows.

- Excel tables expand automatically as new rows are added to the bottom or new columns are added to the right.