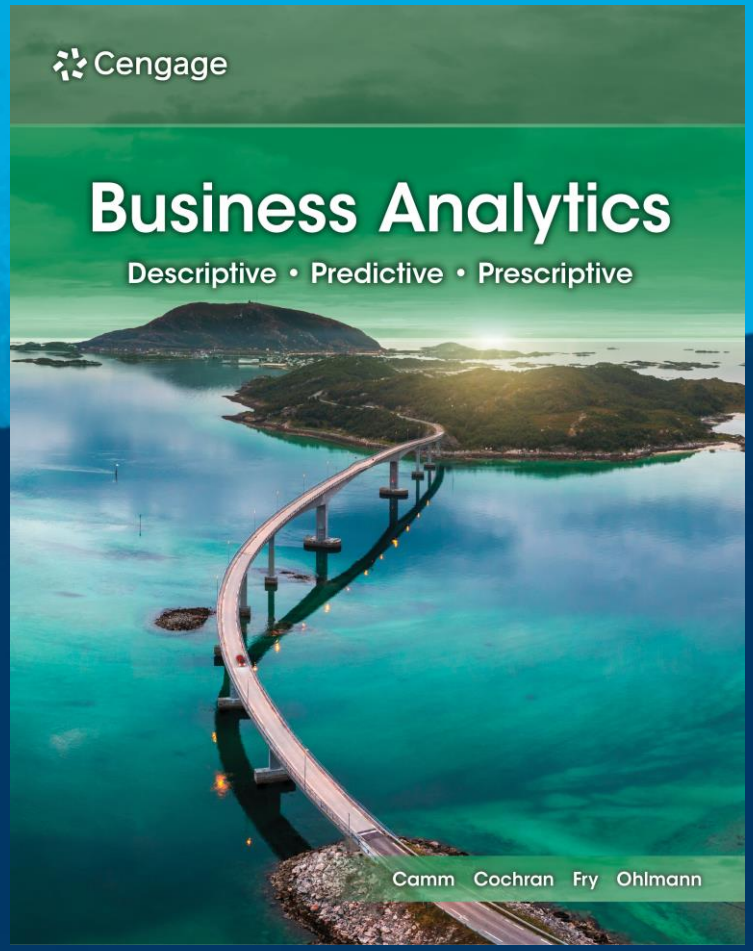


Business Analytics, 5e

Chapter 3 – Data Visualization

CENGAGE

Chapter Contents

Learning Objectives (1 of 2)

After completing this chapter, you will be able to:

LO 3-1 Apply concepts of preattentive attributes and data-ink ratio to improve table and chart designs.

LO 3-2 Create and use PivotTables and PivotCharts to explore and analyze data.

LO 3-3 Create and interpret scatter charts to examine the relationship between two variables.

LO 3-4 Create and interpret line charts, bar charts, and column charts.

LO 3-5 Create and interpret trendlines fit to data on scatter charts and line charts.

CENGAGE

Learning Objectives (2 of 2)

LO 3-6	Create and interpret sorted bar (column) charts, clustered bar (column) charts, and stacked bar (column) charts.
LO 3-7	Create and interpret bubble charts, scatter chart matrixes, and table lenses to visualize the relationships among multiple variables.
LO 3-8	Create and interpret heat maps, sparklines, treemaps, waterfall charts, stock charts, and parallel coordinates plots to visualize data.
LO 3-9	Create and interpret choropleth maps and cartograms for applications to geospatial data.
LO 3-10	Apply the principles of effective data dashboards to suggest designs for data dashboards.

3.1 Preattentive Attributes

Cognitive load is the effort needed to accurately and efficiently process information communicated by a data visualization.

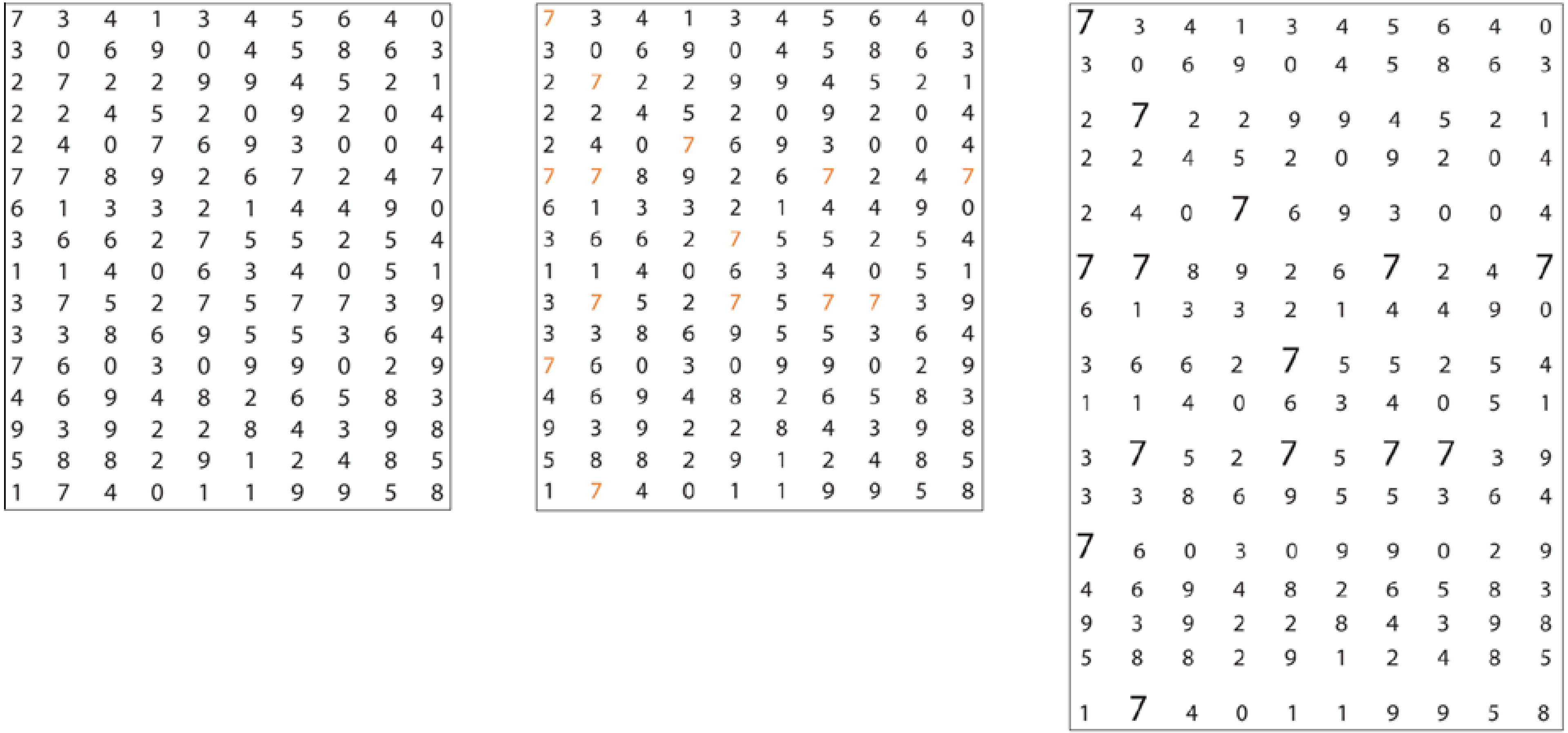
- A data visualization with a reduced cognitive load is easier for the viewer to interpret.

Preattentive attributes are features that can be used in a data visualization to reduce the cognitive load required to interpret it.

- Preattentive attributes related to visual perception include
  - color
  - size
  - shape
  - length

3.1 Application of Preattentive Attributes

The preattentive attributes of color and size aid with the counting of 7s

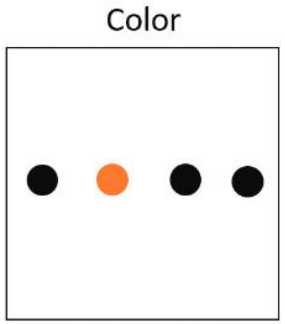


CENGAGE

3.1 The Preattentive Attributes of Color, Size & Length

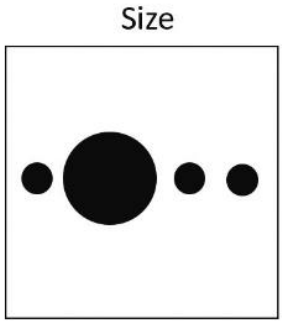
Color is extremely effective in differentiating data in a visualization.

- However, overuse of colors may become distracting.
- Color blindness may affect the ability to distinguish colors.



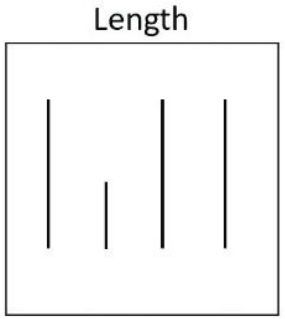
Size refers to the relative amount of space an object occupies in a visualization.

- Most people struggle to estimate 2D relative size differences.



Length refers to the horizontal, vertical, or diagonal distance of a line or bar/column.

- Length is useful for illustrating quantitative values because a longer line corresponds to a larger value.



CENGAGE

3.1 Data-Ink Ratio

Data-ink ratio is the proportion of ink used for data to the total amount of ink in a table or chart.

- The data-ink ratio is helpful for creating effective tables and charts for data visualization
- **Data-ink** is the ink used in a table or chart that is necessary to convey the meaning of the data to the audience
- **Non-data-ink** is the ink used in a table or chart that serves no useful purpose in conveying the data to the audience
- **Decluttering** is the process of increasing the data-ink ratio in a chart.

Next, we consider two examples of decluttering in tables and charts for a data sample of scarf sales.

3.1 Increase of Data-Ink Ratio in Tables

Example of low data-ink table

Scarf Sales			
Day	Sales (units)	Day	Sales (units)
1	150	11	170
2	170	12	160
3	140	13	290
4	150	14	200
5	180	15	210
6	180	16	110
7	210	17	90
8	230	18	140
9	140	19	150
10	200	20	230

Example of high data-ink table

Scarf Sales

Day	Sales (units)	Day	Sales (units)
1	150	11	170
2	170	12	160
3	140	13	290
4	150	14	200
5	180	15	210
6	180	16	110
7	210	17	90
8	230	18	140
9	140	19	150
10	200	20	230

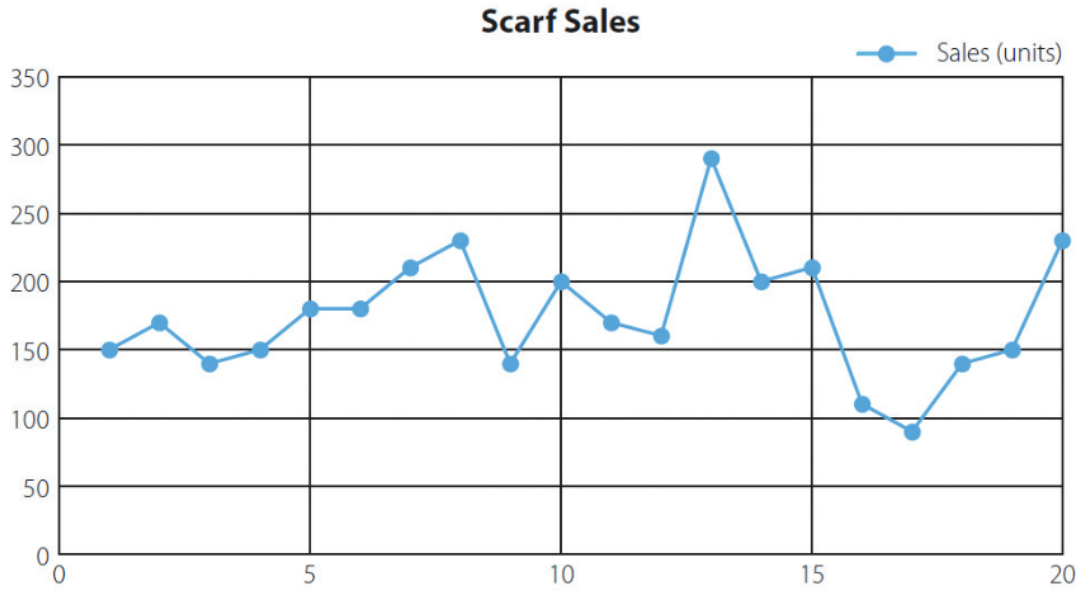
White space is the portion of a data visualization devoid of markings.

- White space use is equivalent to increasing the data-ink in a visualization.

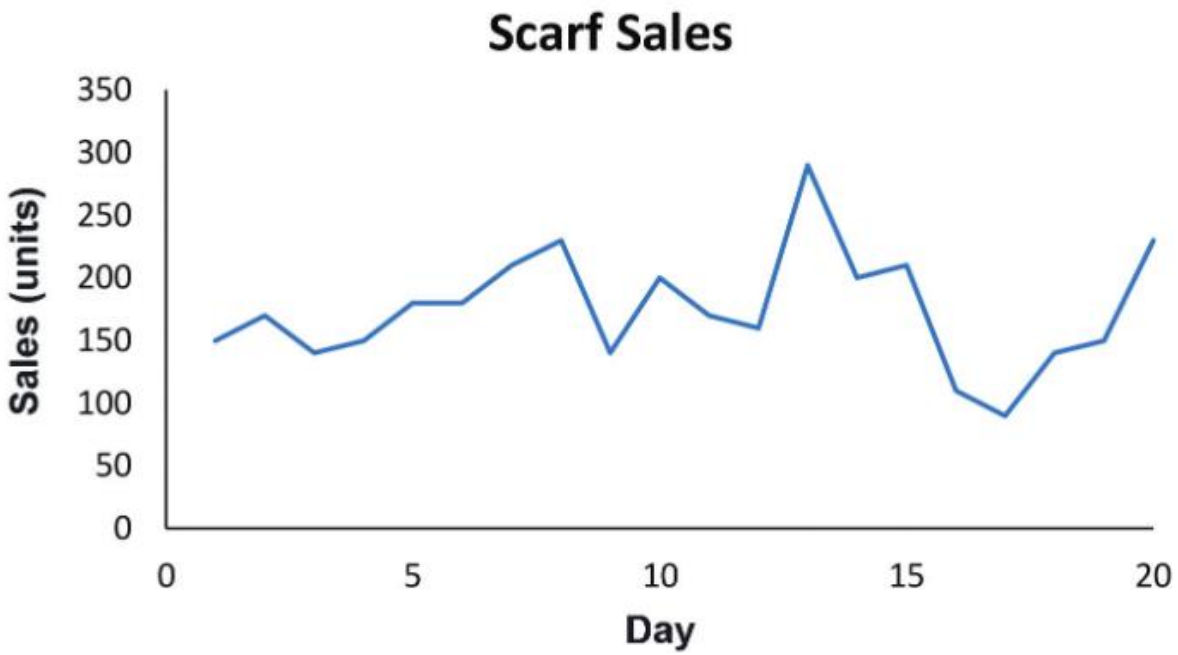
CENGAGE

3.1 Increase of Data-Ink Ratio in Charts

Example of low data-ink chart



Example of high data-ink chart



Before presenting Excel-generated tables and charts to others, it is worth the effort to modify/remove unnecessary default settings.

3.2 Use of Tables to Display Precise Values

Tables should be used when

- the reader needs to refer to specific numerical values,
- the reader needs to make precise comparisons between values, and
- the displayed values have different units or very different magnitudes.

Example of a table showing exact values of monthly costs and revenues for Gossamer Industries.

	Month						
	Jan	**Feb**	**Mar**	**Apr**	**May**	**June**	**Total**
Costs ($)	48,123	56,458	64,125	52,158	54,718	50,985	326,567
Revenues ($)	64,124	66,125	67,125	48,178	51,785	55,678	353,015

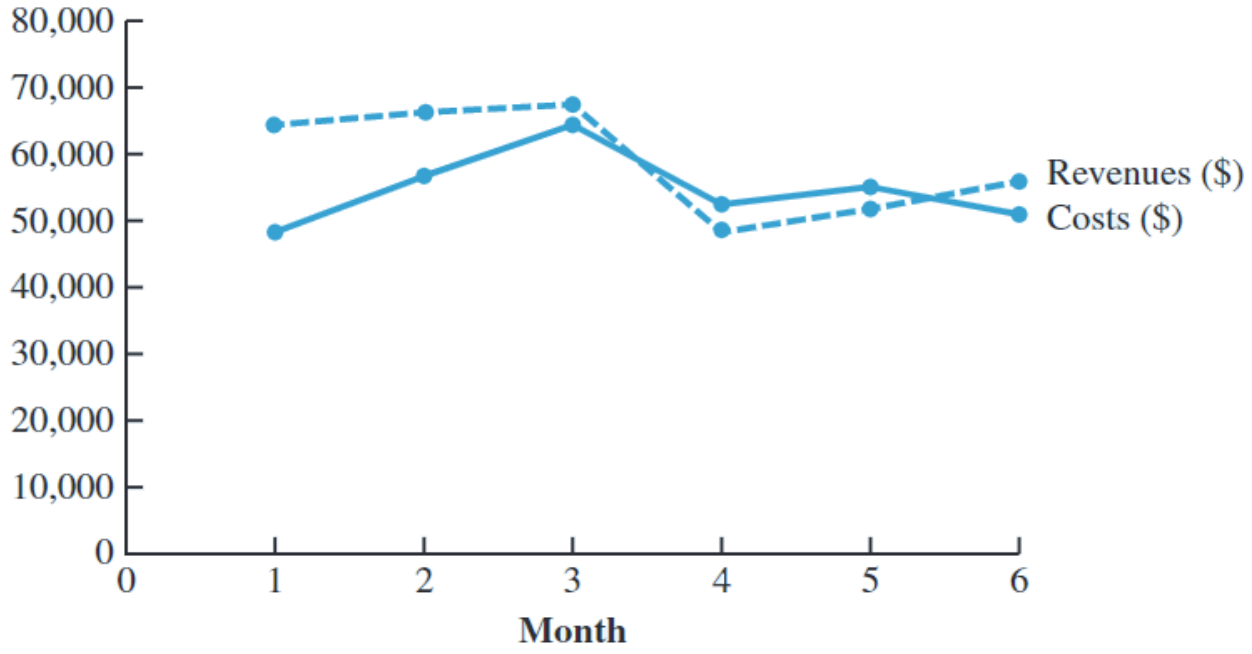
CENGAGE

3.2 Combined Use of Charts and Tables

A single chart may make it very difficult to determine the monthly revenues and costs.

Combining a chart with a table into a single figure allows the reader to

- easily see the monthly changes in revenues and costs, and
- being able to refer to the exact numerical values,
- while avoiding cluttering.



	Month						
	1	2	3	4	5	6	Total
Costs ($)	48,123	56,458	64,125	52,158	54,718	50,985	326,567
Revenues ($)	64,124	66,128	67,125	48,178	51,785	55,687	353,027

CENGAGE

3.2 Use of Tables To Display Levels of Magnitudes

Suppose that you wish to display data on revenues, costs, and head count for each month.

- Costs and revenues are measured in dollars, but head count is measured in number of employees.
- Because of the different levels of magnitudes, a table is preferred over a chart.

	Month						
	Jan	**Feb**	**Mar**	**Apr**	**May**	**June**	**Total**
Head count	8	9	10	9	9	9	
Costs ($)	48,123	56,458	64,125	52,158	54,718	50,985	326,567
Revenues ($)	64,124	66,125	67,125	48,178	51,785	55,678	353,015

CENGAGE

3.2 Table Design Principles

- Keep the data-ink ratio low.
- Use lines only to separate labels from data and calculated fields.
- Labels should be left-aligned.
- Values should be right-aligned.
- Center vertical labels.

Design A:

	Month						
	1	2	3	4	5	6	Total
Costs ($)	48,123	56,458	64,125	52,158	54,718	50,985	326,567
Revenues ($)	64,124	66,128	67,125	48,178	51,785	55,687	353,027
Profits ($)	16,001	9,670	3,000	(3,980)	(2,933)	4,702	26,460

Design B:

	Month						
	1	2	3	4	5	6	Total
Costs ($)	48,123	56,458	64,125	52,158	54,718	50,985	326,567
Revenues ($)	64,124	66,128	67,125	48,178	51,785	55,687	353,027
Profits ($)	16,001	9,670	3,000	(3,980)	(2,933)	4,702	26,460

Design C:

	Month						
	1	2	3	4	5	6	Total
Costs ($)	48,123	56,458	64,125	52,158	54,718	50,985	326,567
Revenues ($)	64,124	66,128	67,125	48,178	51,785	55,687	353,027
Profits ($)	16,001	9,670	3,000	(3,980)	(2,933)	4,702	26,460

Design D:

	Month						
	1	2	3	4	5	6	Total
Costs ($)	48,123	56,458	64,125	52,158	54,718	50,985	326,567
Revenues ($)	64,124	66,128	67,125	48,178	51,785	55,687	353,027
Profits ($)	16,001	9,670	3,000	(3,980)	(2,933)	4,702	26,460

CENGAGE

3.2 Crosstabulation

A **crosstabulation** is a useful approach to describing a tabular summary of data for two variables.

DATAfile: *restaurant*

Consider the following application based on data from a Zagat's Restaurant Review consisting of a sample of 300 restaurants in Los Angeles.

Variables and types:

Quality rating (categorical)	good, very good, or excellent
Meal Price (quantitative)	expressed in dollars
Wait time (quantitative)	expressed in minutes

3.2 Crosstabulation of Quality Rating and Meal Price

- The right and bottom margins of the crosstabulation give the frequencies of quality rating and meal price separately.
- The greatest number of restaurants in the sample (64) have a very good rating and a meal price in the $20–29 range.
- Only two restaurants have an excellent rating and $10–19 meal prices.

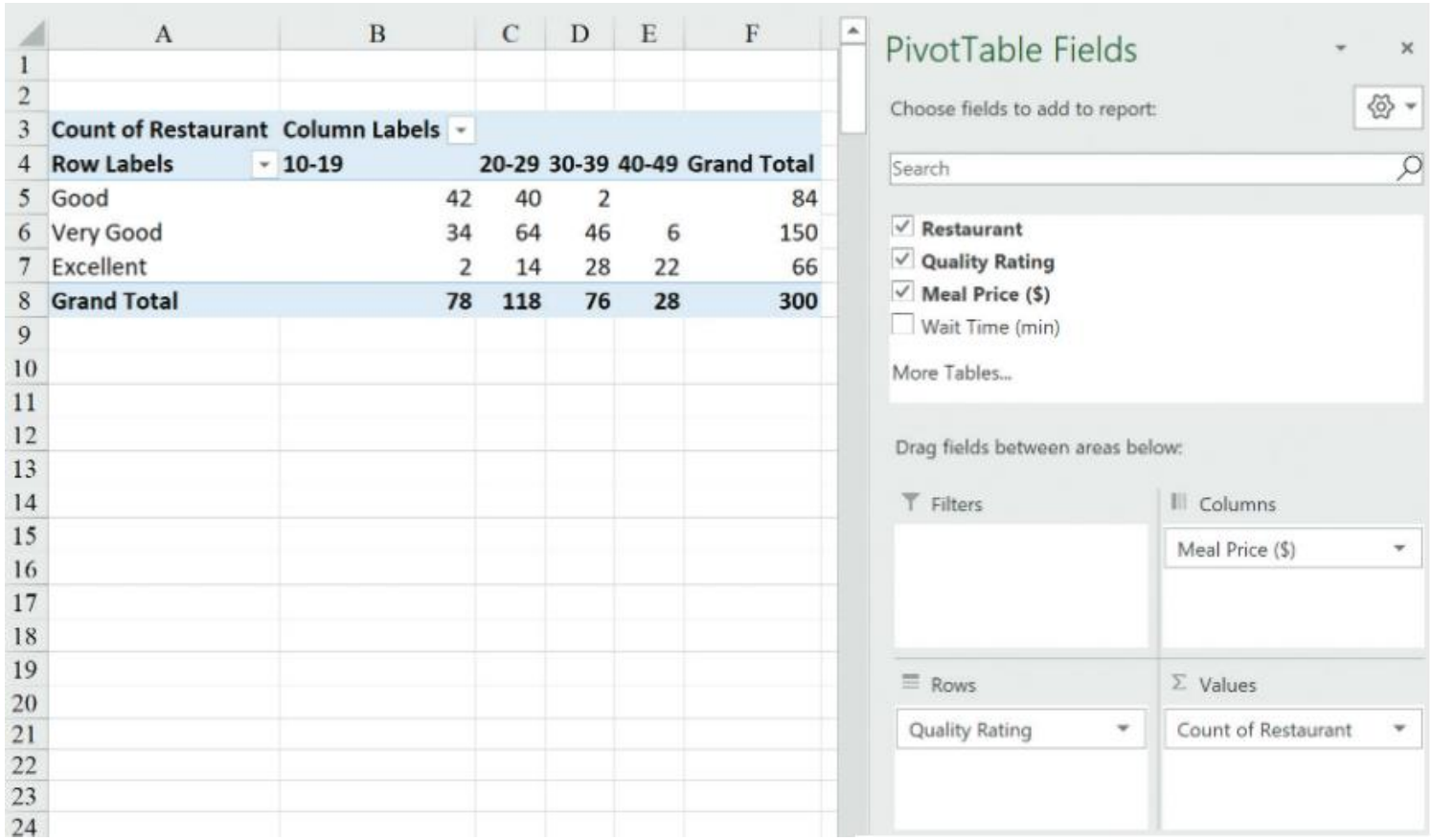
	Meal Price				
Quality Rating	**$10-19**	**$20-29**	**$30-39**	**$40-49**	**Total**
Good	42	40	2	0	84
Very Good	34	64	46	6	150
Excellent	2	14	28	22	300
Total	78	118	76	28	300

3.2 PivotTables in Excel

A crosstabulation is Excel is called a **PivotTable**.

The image to the right shows a PivotTable reproducing the cross-tabulation from the previous example.

To create this PivotTable in Excel, follow the steps shown in the notes.



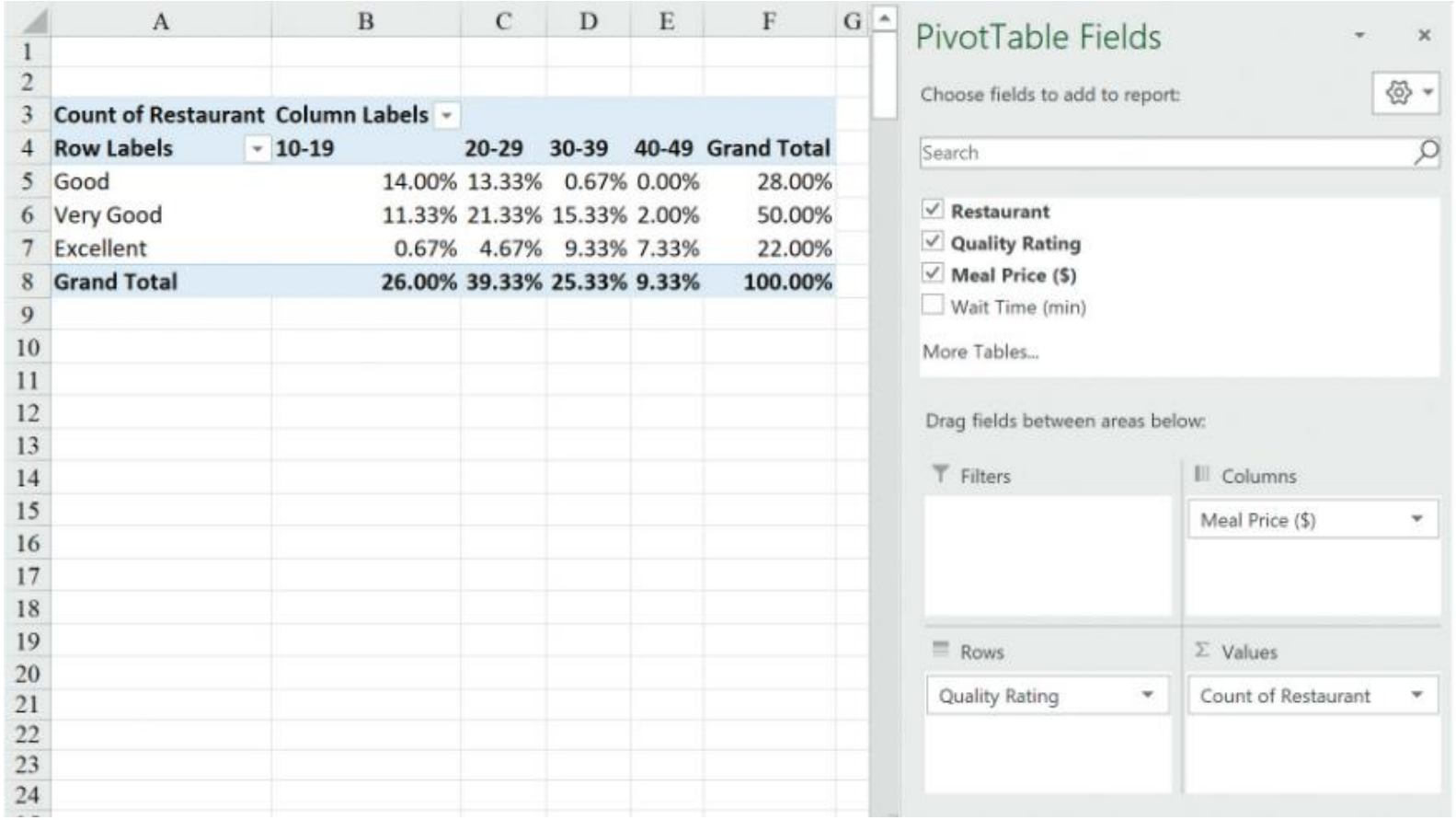
Count of Restaurant	Column Labels				
Row Labels	10-19	20-29	30-39	40-49	Grand Total
Good	42	40	2		84
Very Good	34	64	46	6	150
Excellent	2	14	28	22	66
Grand Total	78	118	76	28	300

3.2 Percent Frequency Distribution as a PivotTable

The PivotTable indicates that

- 50% of restaurants are very good, and
- 26% have meal prices in the $10-$19 range.

See notes for Excel instructions.



Count of Restaurant	Column Labels				
Row Labels	10-19	20-29	30-39	40-49	Grand Total
Good	14.00%	13.33%	0.67%	0.00%	28.00%
Very Good	11.33%	21.33%	15.33%	2.00%	50.00%
Excellent	0.67%	4.67%	9.33%	7.33%	22.00%
Grand Total	26.00%	39.33%	25.33%	9.33%	100.00%

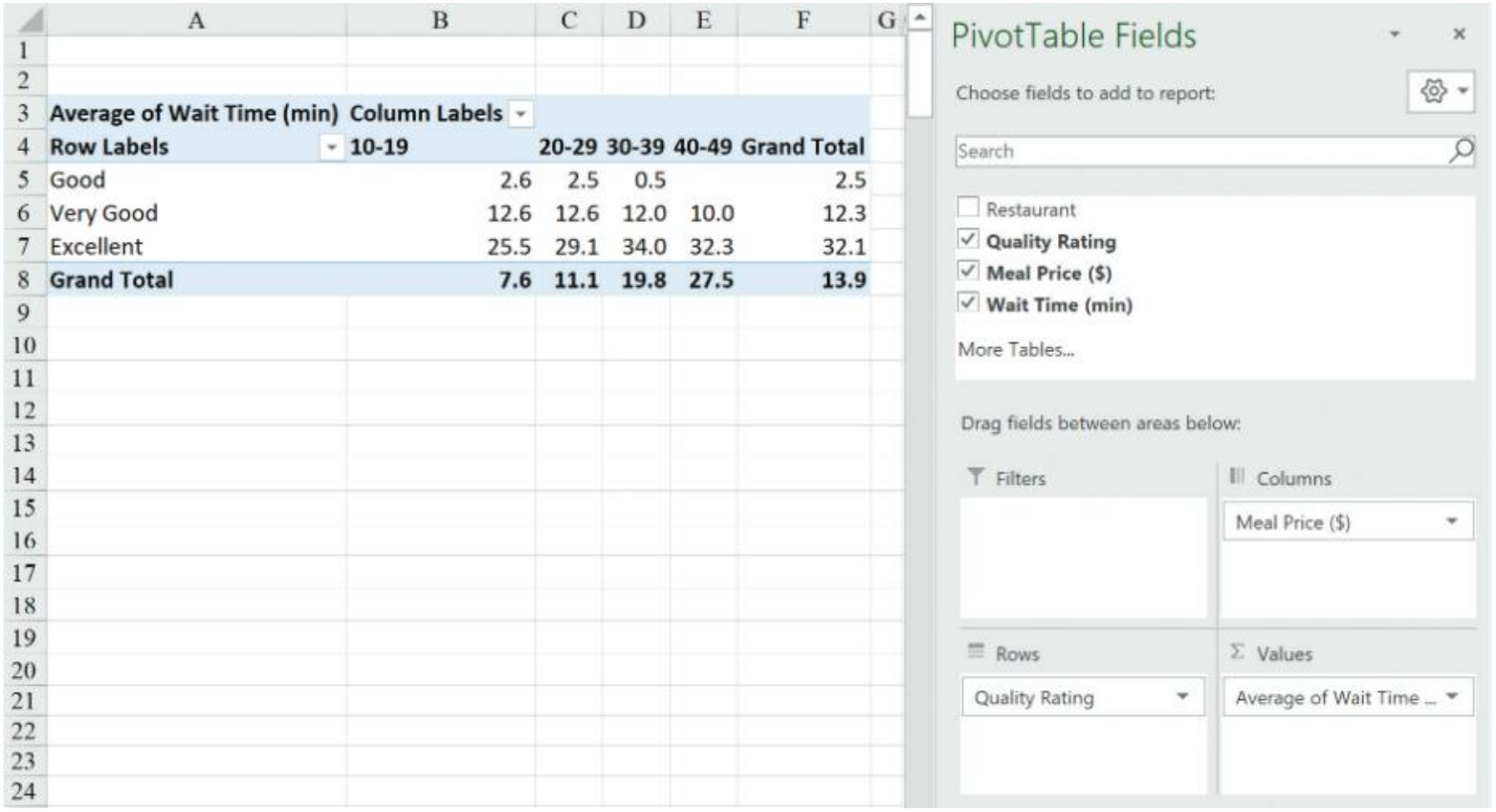
3.2 PivotTable Report with Average Wait Time Added

The PivotTable indicates that the wait time increases with

- quality rating and
- meal prices in excellent restaurants.

While it decreases with increasing meal prices in good restaurants.

See notes for Excel instructions.

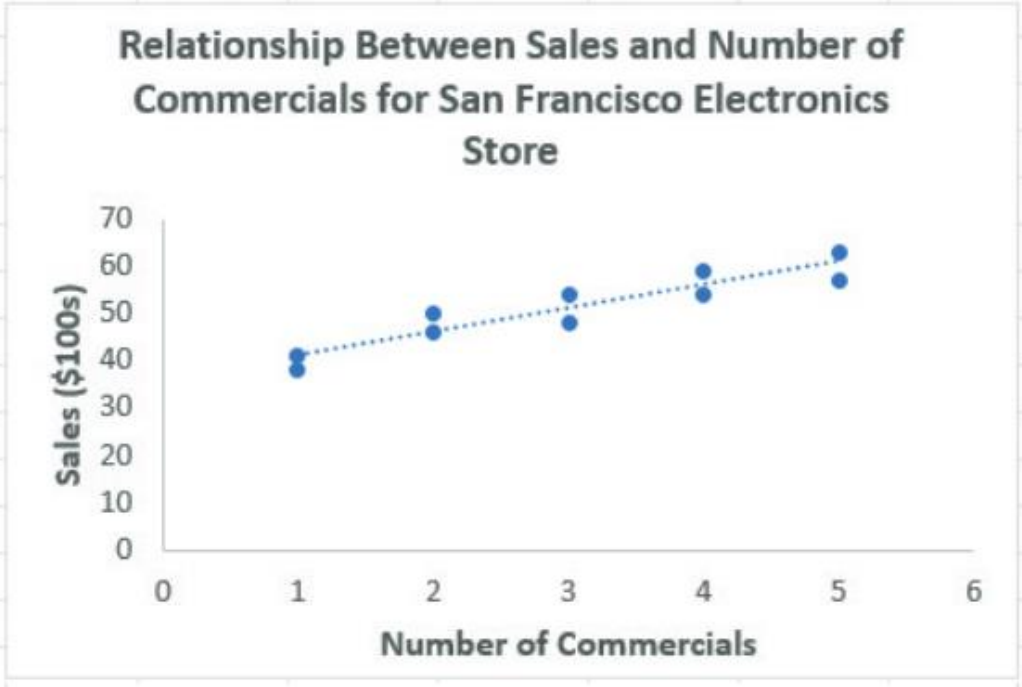


Average of Wait Time (min)	Column Labels				
Row Labels	10-19	20-29	30-39	40-49	Grand Total
Good	2.6	2.5	0.5		2.5
Very Good	12.6	12.6	12.0	10.0	12.3
Excellent	25.5	29.1	34.0	32.3	32.1
Grand Total	**7.6**	**11.1**	**19.8**	**27.5**	**13.9**

3.3 Scatter Charts

A **scatter chart** is a graphical presentation of the relationship between two quantitative variables.

Week	Number of Commercials	Sales ($100s)
1	2	50
2	5	57
3	1	41
4	3	54
5	4	54
6	1	38
7	5	63
8	3	48
9	4	59
10	2	46



See notes on how to create the scatter chart in Excel for the *electronics* data.

3.3 Recommended Charts in Excel

The steps below demonstrate the Recommended Charts tool in Excel for the San Francisco electronics store data.

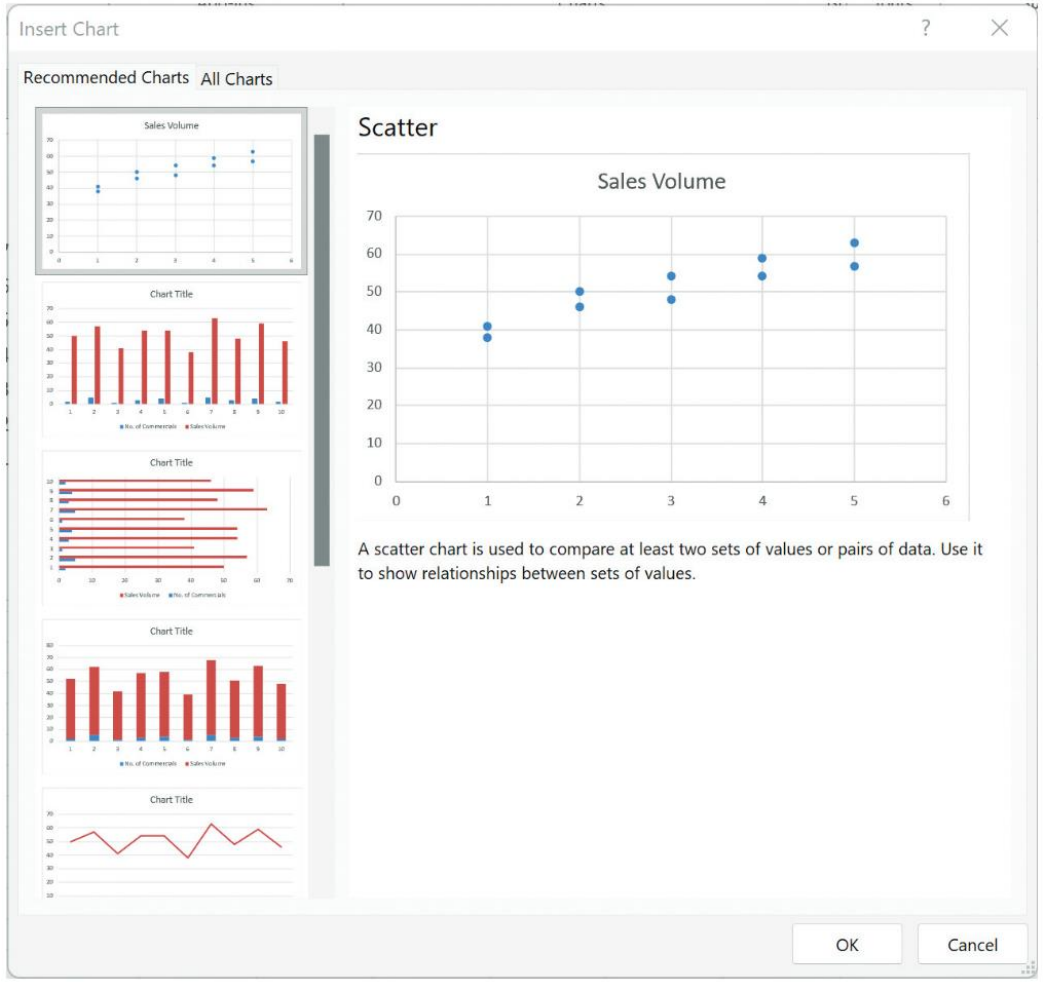
Step 1. Select cells B1:C11

Step 2: Click the **Insert** tab in the Ribbon

Step 3: Click the **Recommended Charts** button in the **Charts** group

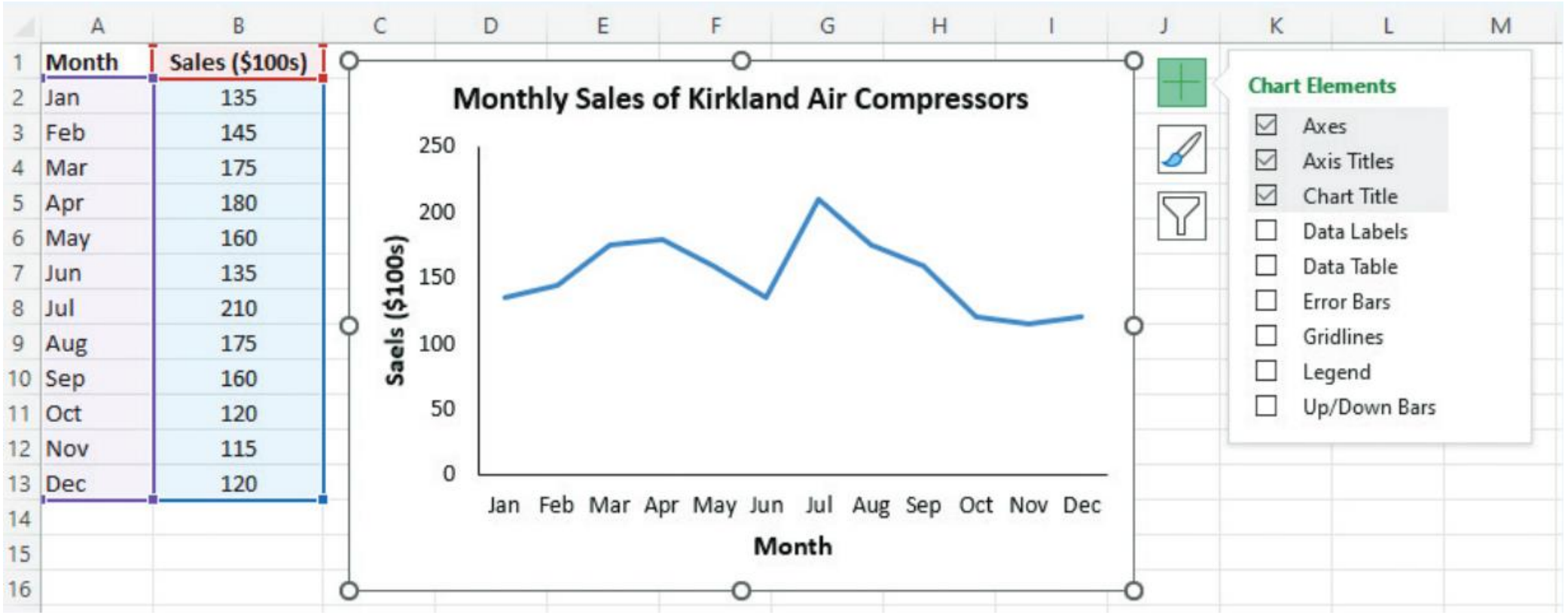
Step 4: When the **Insert Chart** dialog box appears, select the **Scatter** option

Click **OK**



3.3 Line Charts

A **line chart** is a useful representation of time series data (DATAfile: *kirkland*.)



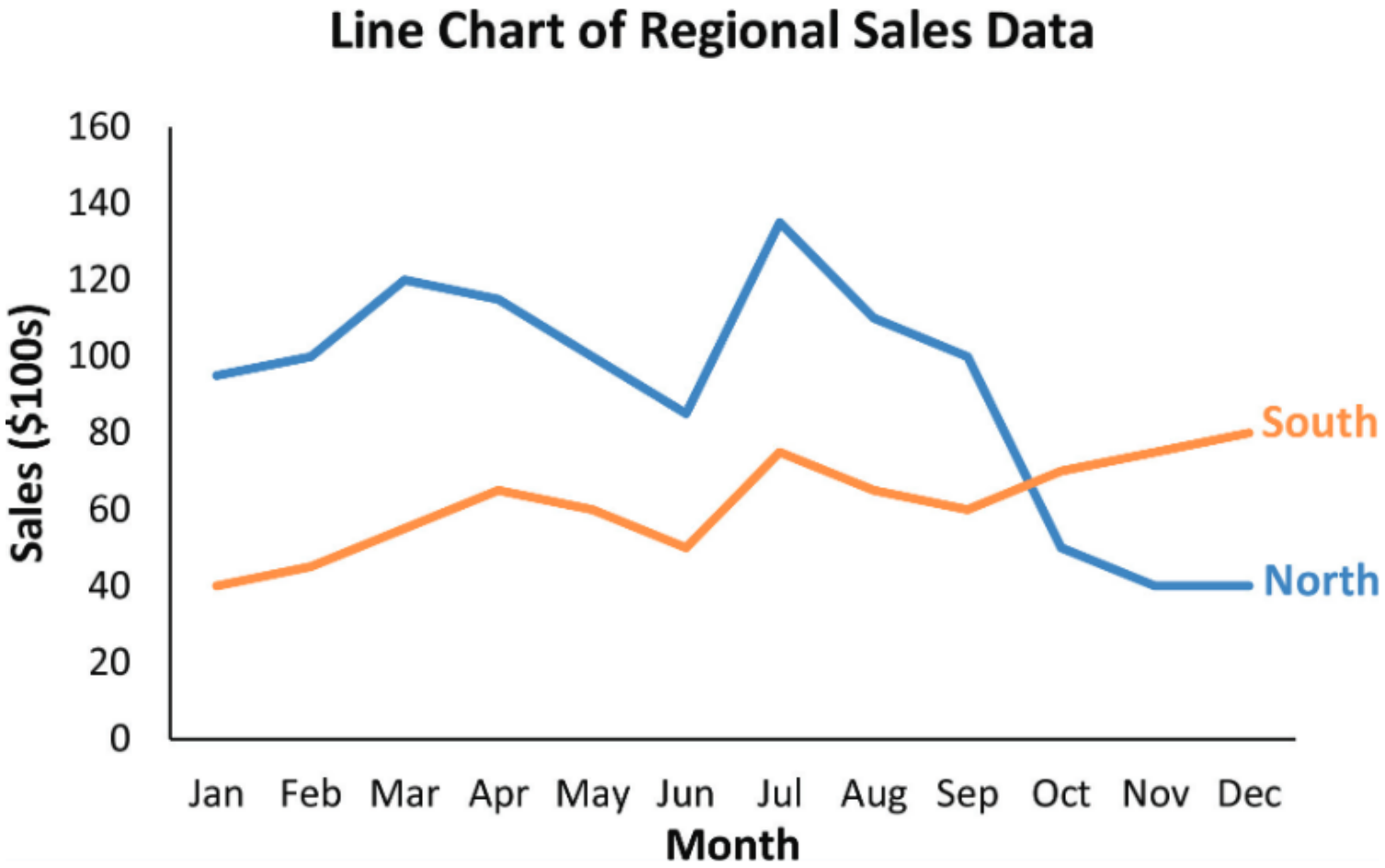
See notes on how to create the line chart in Excel for the *kirkland* data.

3.3 Multiple Line Charts

To create multiple line charts in Excel, select the A2:C14 range in the DATAfile: *kirklandregional*.

Follow the same steps used to create a single line chart.

The preattentive attribute of color is used to differentiate between South and North.

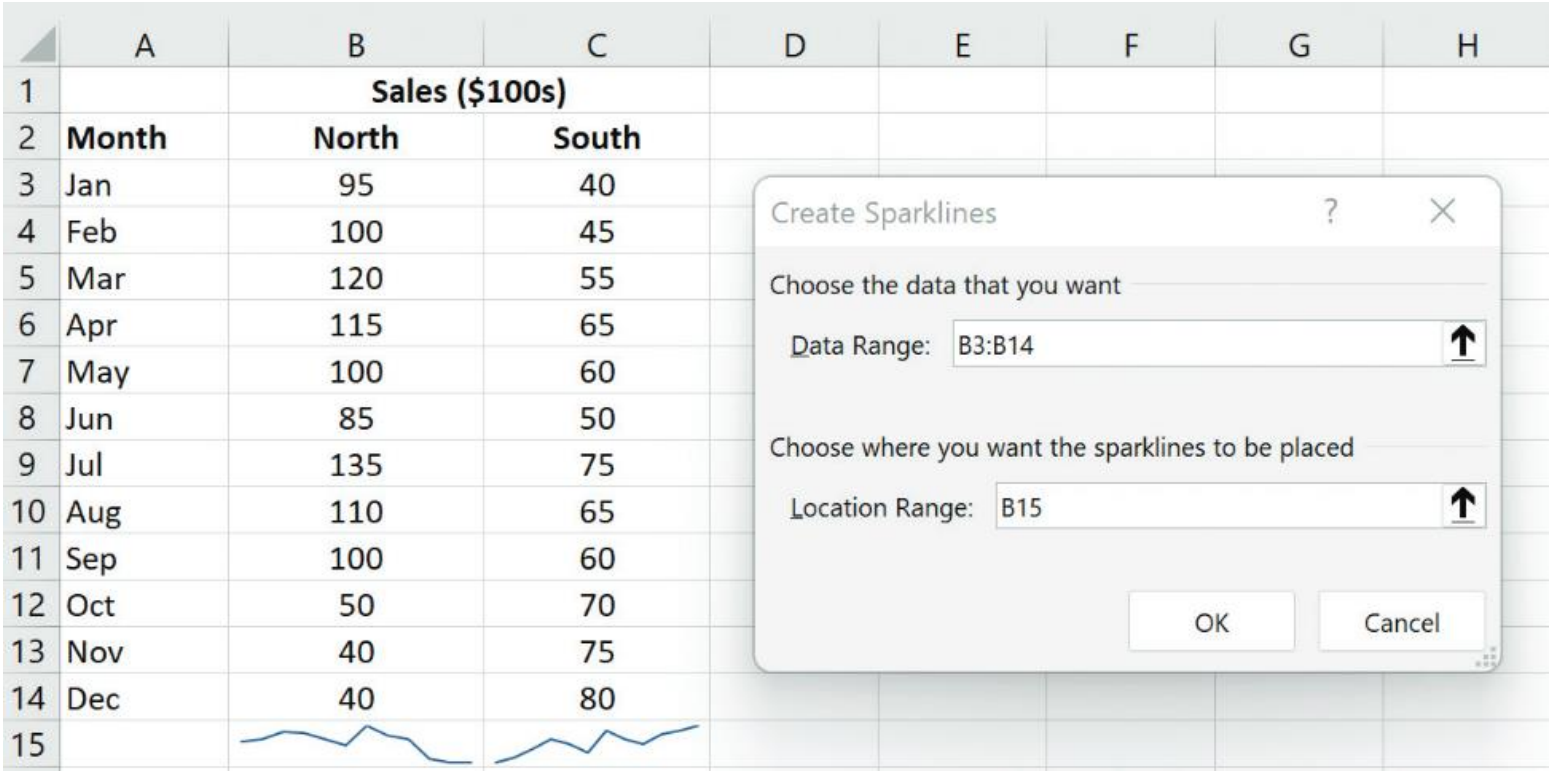


3.3 Sparklines

A **sparkline** is a minimalist type of line chart that can be placed directly into a cell in Excel.

Sparklines take up little space and can be used to provide information on overall trends for time series data.

See notes to create sparklines for Kirkline Industries regional sales data.



3.3 Bar Charts

Bar charts use horizontal bars to display the magnitude of the quantitative variable.

See the notes for instructions on creating a sorted bar chart with data labels for the DATAfile: *accountsmanaged.*

Manager	Accounts Managed
Williams	6
Edwards	11
Jones	15
Smith	21
Davis	24
Francois	28
Lopez	29
Gentry	37

3.3 Column Charts

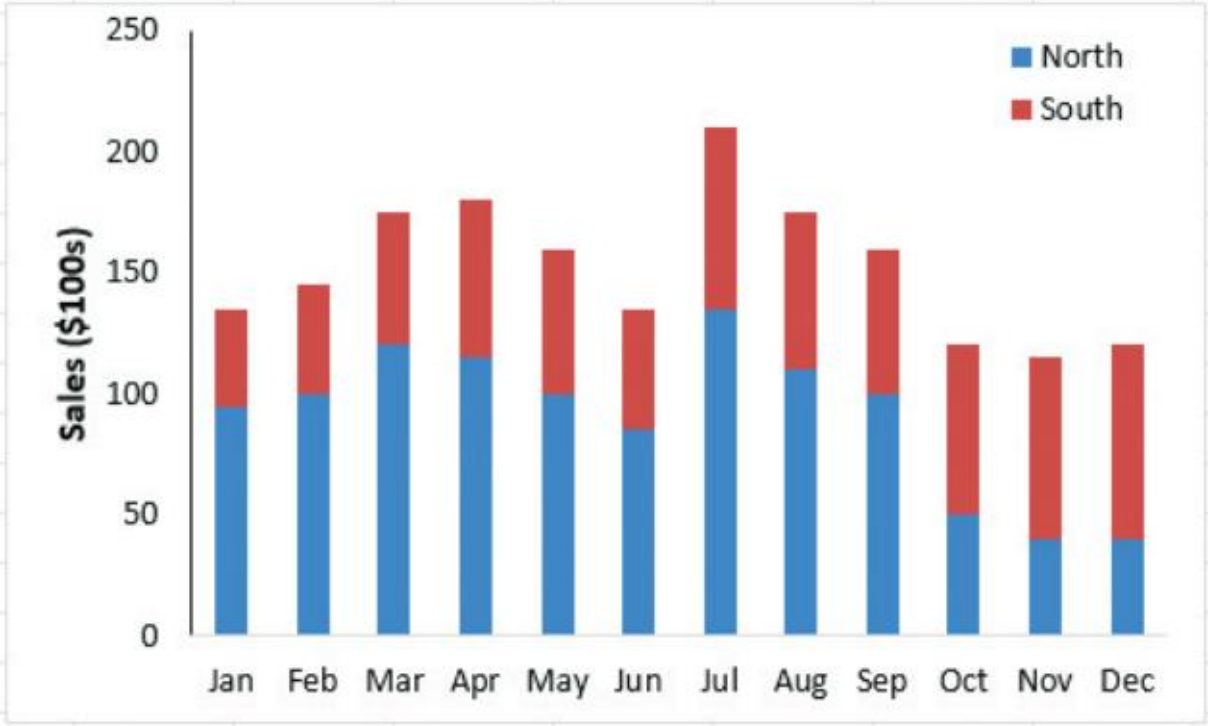
Column charts use vertical bars to display the magnitude of the quantitative variable.

- When visualizing more than one category of data, a **stacked-column chart** or a **clustered-column chart** can be used.
- When horizontal instead of vertical bars are used, we have a stacked-bar chart or a clustered-bar chart, respectively
- Clustered charts are often superior to stacked charts, but they can become cluttered when using more than a few variables per category.

The examples of clustered- and stacked-column charts in the next two slides use the *kirklandregional* data.

3.3 Stacked-Column Chart

	A	B	C
1		Sales ($100s)	
2	Month	North	South
3	Jan	95	40
4	Feb	100	45
5	Mar	120	55
6	Apr	115	65
7	May	100	60
8	Jun	85	50
9	Jul	135	75
10	Aug	110	65
11	Sep	100	60
12	Oct	50	70
13	Nov	40	75
14	Dec	40	80
15			
16			

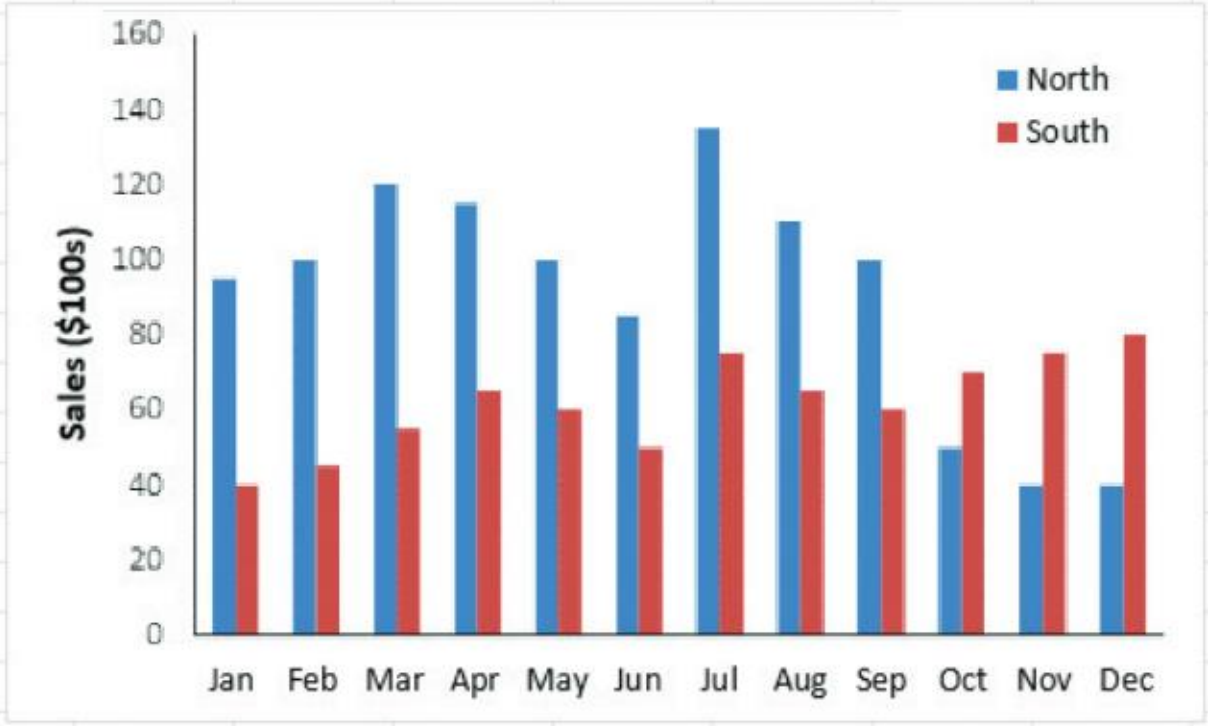


See notes for instructions on how to build a stacked-column chart in Excel.

CENGAGE

3.3 Clustered-Column Chart

	A	B	C
1		Sales ($100s)	
2	**Month**	**North**	**South**
3	Jan	95	40
4	Feb	100	45
5	Mar	120	55
6	Apr	115	65
7	May	100	60
8	Jun	85	50
9	Jul	135	75
10	Aug	110	65
11	Sep	100	60
12	Oct	50	70
13	Nov	40	75
14	Dec	40	80
15			
16			



See notes for instructions on how to build a clustered-column chart in Excel.

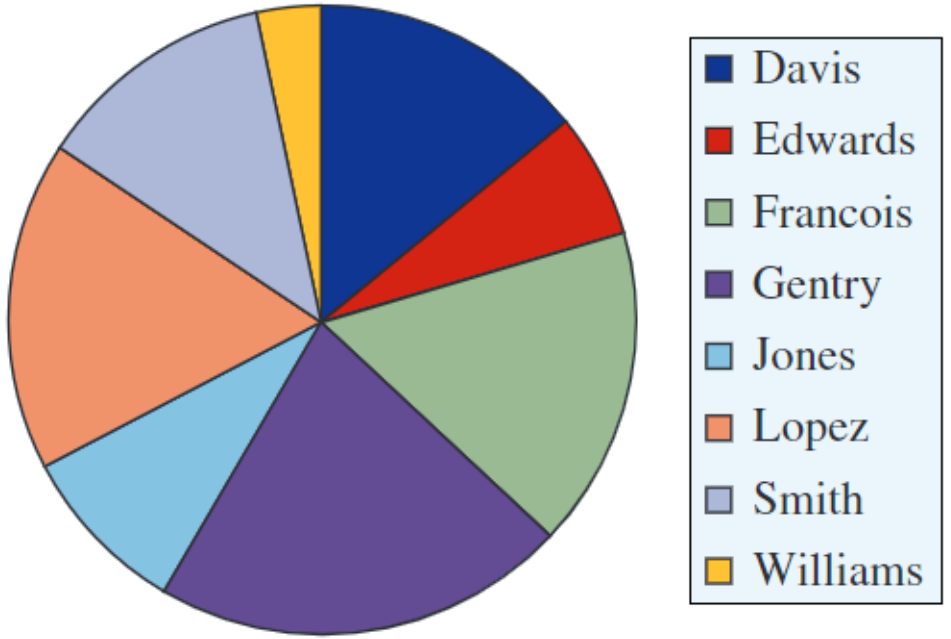
CENGAGE

3.3 A Note on Pie Charts and 3-D Charts

Pie charts are a common form of chart used to compare categorical data.

- However, their inefficient use of the preattentive attributes of color and size make it a poor choice for data visualization.
- Color: A color-blind person may struggle to recognize the color coding for the different account managers.
- Size: it is difficult to compare the relative sizes of the pie wedges.

The same argument can also be made against using three-dimensional (3-D) charts.



3.3 Bubble Charts

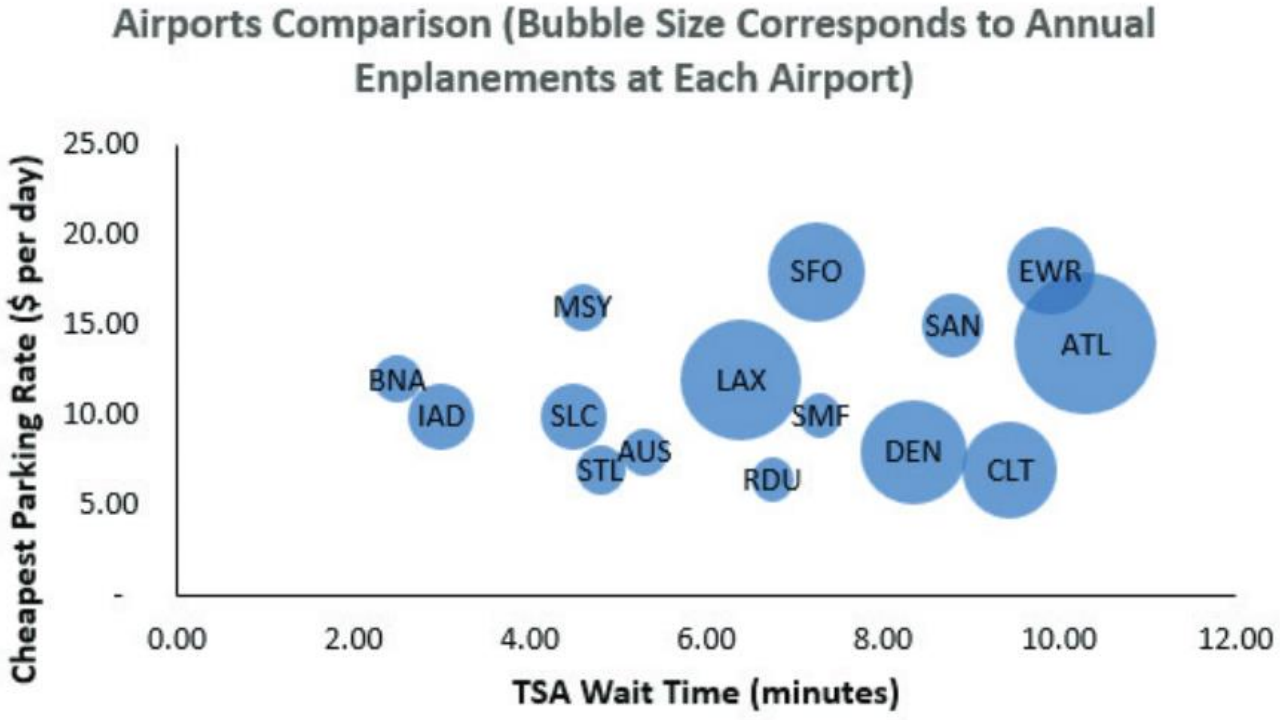
A **bubble chart** is a graphical means of visualizing three variables in a two-dimensional graph.

DATAfile: *airportdata*

For each airport, we have the following quantitative variables:

- Average TSA wait time (min.)
- Cheapest parking rate ($/hr.)
- Boarding passengers/year (millions)

See notes for Excel instructions.



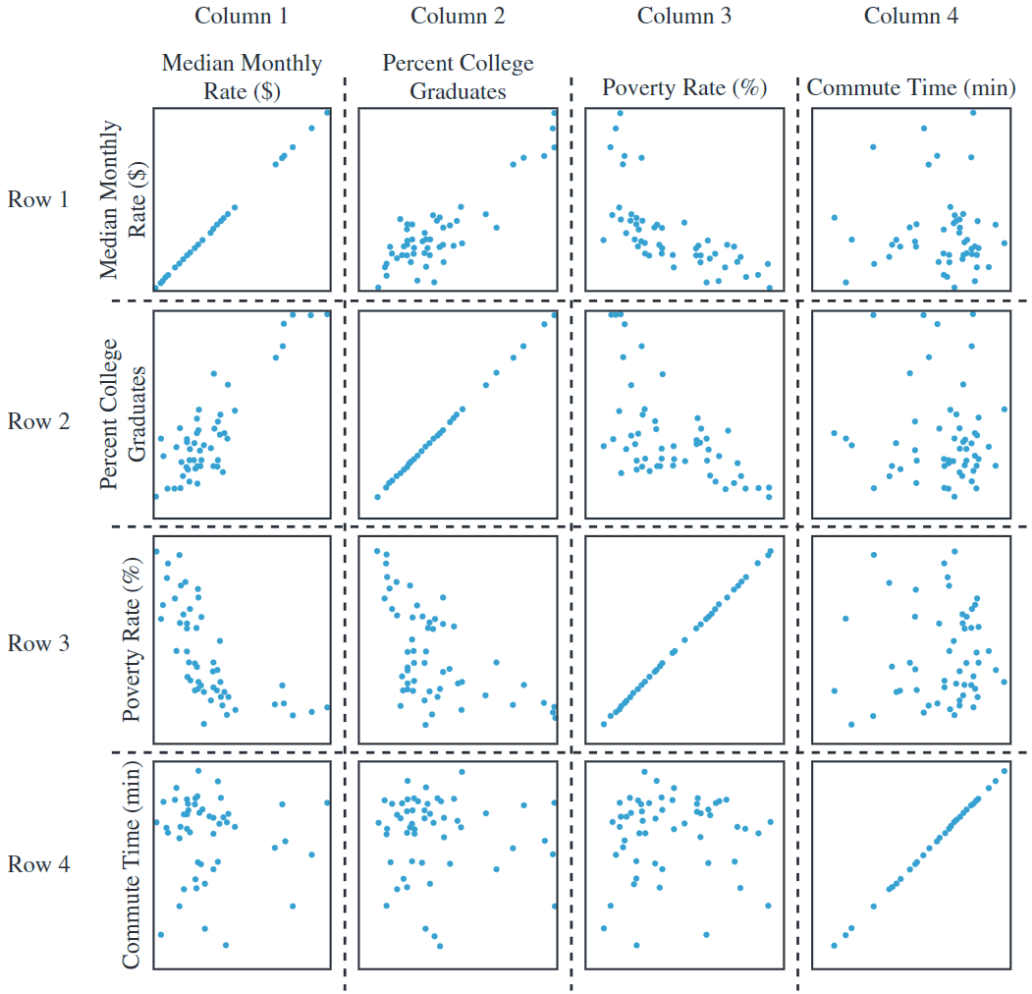
3.3 Scatter-Chart Matrix

A **scatter-chart matrix** allows the reader to compare the relationships among multiple variables.

The DATAfile *nyc* contains data for each of New York City's 55 sub-boroughs on

- median monthly rent,
- % college graduates,
- poverty rate, and
- mean commute time to work.

Instructions for several software packages are shown in the appendixes.



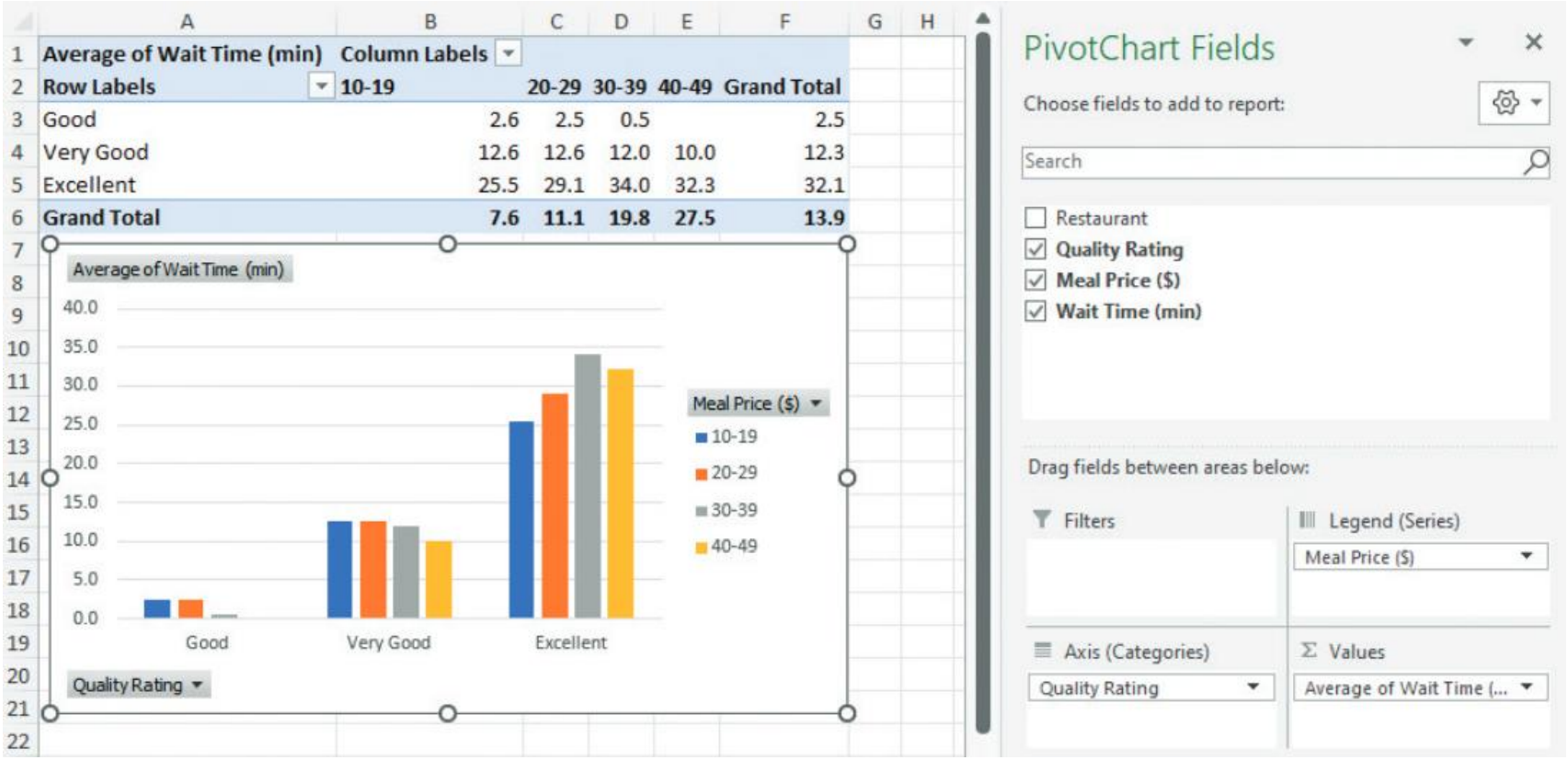
3.3 Table Lens

A **table lens** is also used to visualize relationships between pairs of variables. Shown below is a portion of the table lens for 11 of the 55 NYC sub-boroughs. See notes for instructions on how to create a table lens using Excel.

	A	B	C	D	E
1	**Area (Sub-Borough)**	**Median Monthly Rent ($)**	**Percent College Graduates (%)**	**Poverty Rate (%)**	**Commute Time (min)**
2	Greenwich Village/Financial District	1965	78.3	7.9	44.0
3	Stuyvesant Town/Turtle Bay	1849	78.0	6.9	36.7
4	Upper East Side	1706	78.2	5.9	29.1
5	Upper West Side	1635	74.6	8.8	38.7
6	Chelsea/Clinton/Midtown	1624	66.1	12.7	43.7
7	Park Slope/Carroll Gardens	1566	61.8	8.4	37.5
8	Bayside/Little Neck	1243	41.3	7.6	40.6
9	Rego Park/Forest Hills	1192	51.2	10.4	42.5
10	Queens Village	1184	30.2	6.2	41.3
11	Flushing/Whitestone	1170	32.5	11.4	23.4
12	South Ozone Park/Howard Beach	1148	17.3	8.1	42.0

3.3 PivotCharts

PivotCharts paired with PivotTables allow to summarize and analyze data with a crosstabulation and a chart. See notes for Excel instructions.



3.4 Heat Maps

A **heat map** is a 2-D graphical representation of data that uses different shades of color to indicate magnitude. See notes for Excel instructions.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	SPARKLINES
2	St. Louis	-2%	-1%	-1%	0%	2%	4%	3%	5%	6%	7%	8%	8%	
3	Phoenix	5%	4%	4%	2%	2%	-2%	-5%	-8%	-6%	-5%	-7%	-8%	
4	Albany	-5%	-6%	-4%	-5%	-2%	-5%	-5%	-3%	-1%	-2%	-1%	-2%	
5	Austin	16%	15%	15%	16%	18%	17%	14%	15%	16%	19%	18%	16%	
6	Cincinnati	-9%	-6%	-7%	-3%	3%	6%	8%	11%	10%	11%	13%	11%	
7	San Francisco	2%	4%	5%	8%	4%	2%	4%	3%	1%	-1%	1%	2%	
8	Seattle	7%	7%	8%	7%	5%	4%	2%	0%	-2%	-4%	-6%	-5%	
9	Chicago	5%	3%	2%	6%	8%	7%	8%	5%	8%	10%	9%	8%	
10	Atlanta	12%	14%	13%	17%	12%	11%	8%	7%	7%	8%	5%	3%	
11	Miami	2%	3%	0%	1%	-1%	-4%	-6%	-8%	-11%	-13%	-11%	-10%	
12	Minneapolis	-6%	-6%	-8%	-5%	-6%	-5%	-5%	-7%	-5%	-2%	-1%	-2%	
13	Denver	5%	4%	1%	1%	2%	3%	1%	-1%	0%	1%	2%	3%	
14	Salt Lake City	7%	7%	7%	13%	12%	8%	5%	9%	10%	9%	7%	6%	
15	Raleigh	4%	2%	0%	5%	4%	3%	5%	5%	9%	11%	8%	6%	
16	Boston	-5%	-5%	-3%	4%	-5%	-4%	-3%	-1%	1%	2%	3%	5%	
17	Pittsburgh	-6%	-6%	-4%	-5%	-3%	-3%	-1%	-2%	-2%	-1%	-2%	-1%	

CENGAGE

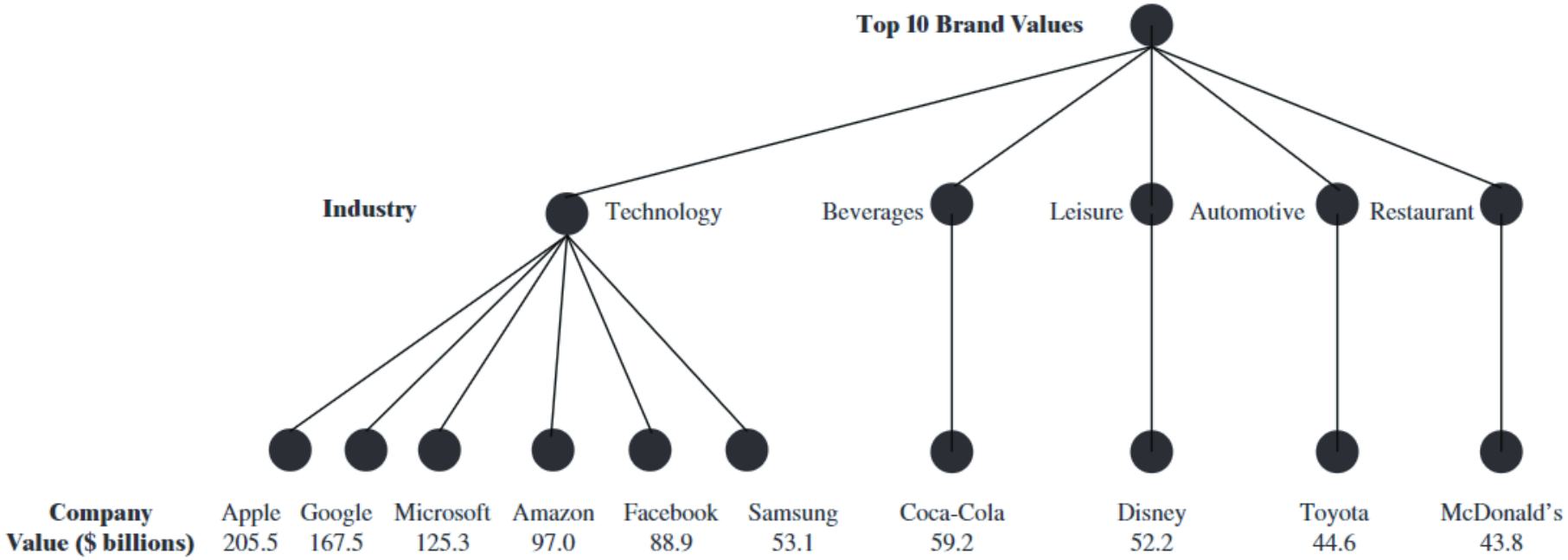
3.4 Hierarchical Tree Structures

Hierarchical data are categorical data further decomposed into subcategories.

Hierarchical data can be represented with a tree-like structure, where the branches of the tree lead to categories and subcategories.

DATAfile:

brandvalues



3.4 Treemaps

A **treemap** is a chart that uses

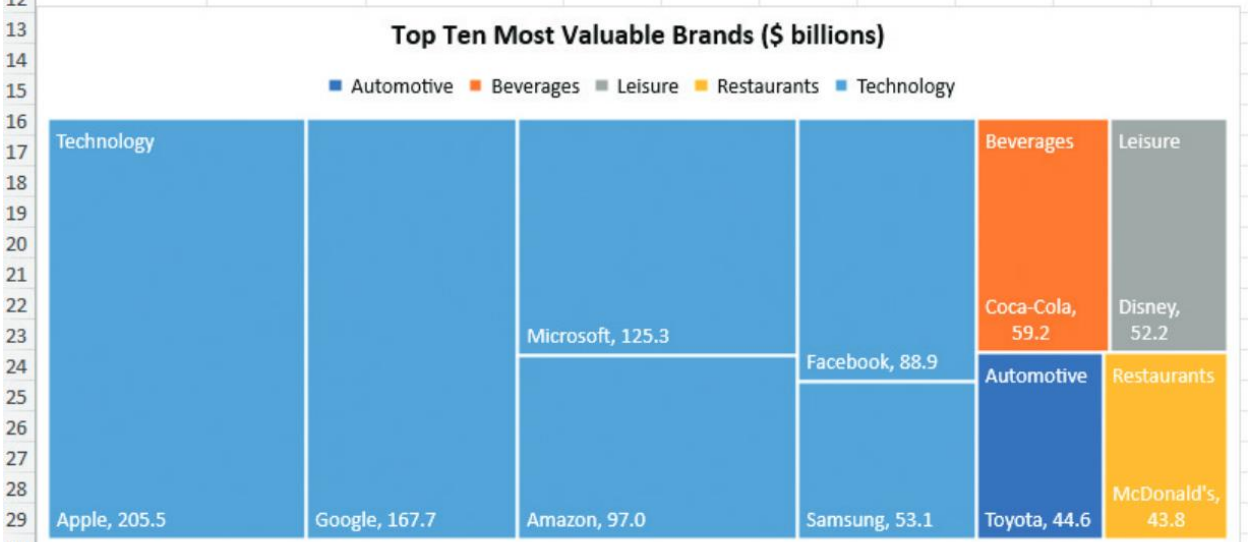
- size,
- color, and
- arrangement

of rectangles to display a quantitative variable for different categories and subcategories.

Shown to the right is the treemap for the *brandvalues* data.

See notes for instructions on how to build a treemap using Excel.

	A	B	C
1	**Industry**	**Brand**	**Value ($ billions)**
2	Automotive	Toyota	44.6
3	Beverages	Coca-Cola	59.2
4	Leisure	Disney	52.2
5	Restaurants	McDonald's	43.8
6	Technology	Apple	205.5
7	Technology	Google	167.7
8	Technology	Microsoft	125.3
9	Technology	Amazon	97.0
10	Technology	Facebook	88.9
11	Technology	Samsung	53.1



3.4 Waterfall Charts

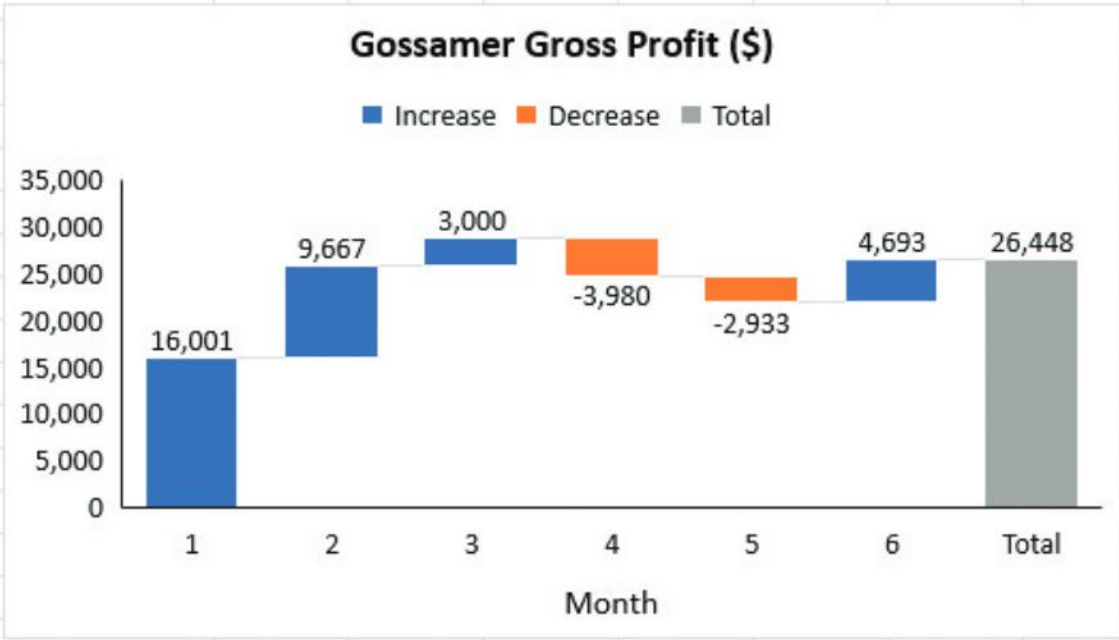
A **waterfall chart** displays the cumulative effect of changes on a variable of interest.

- The magnitude of each change is represented by a column anchored at the cumulative height of the changes in the preceding categories.

DATAfile: *gossamer_profit*

See notes for instructions on how to create a waterfall chart.

		Month						
		1	2	3	4	5	6	Total
Costs ($)		48,123	56,458	64,125	52,158	54,718	50,985	326,567
Revenues ($)		64,124	66,125	67,125	48,178	51,785	55,678	353,015
Gross Profit ($)		16,001	9,667	3,000	-3,980	-2,933	4,693	26,448



CENGAGE

3.4 Stock Charts

A **stock chart** is a graphical display of stock prices over time.

Excel provides four different types of stock charts.

In this slide, we consider the **high-low-close stock chart** for five trading days in April of the company Verizon Wireless.

DATAfile: *verizon*

See notes for Excel instructions on how to build a high-low-close stock chart.

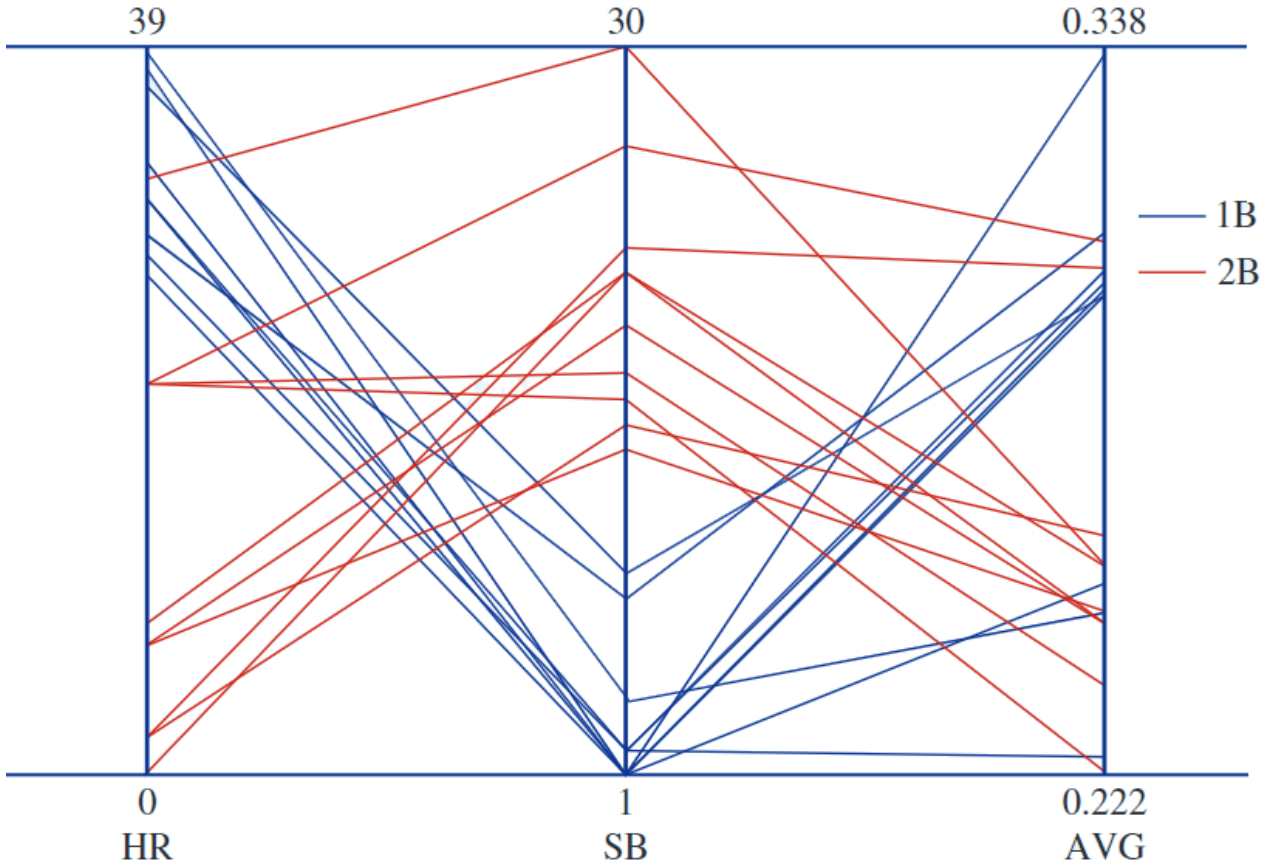
	A	B	C	D
1	**Date**	**High**	**Low**	**Close**
2	20-Apr	$58.91	$57.96	$58.13
3	21-Apr	$58.04	$56.72	$56.82
4	22-Apr	$58.57	$57.23	$57.99
5	23-Apr	$58.66	$57.47	$57.59
6	24-Apr	$57.99	$56.83	$57.93

CENGAGE

3.4 Parallel-Coordinates Plot

A **parallel-coordinates plot** is a helpful chart for examining data with more than two variables.

- Each observation is represented by a line connecting each vertical axis.
- The height of the line represents the value taken by that observation for the variable on each vertical axis.

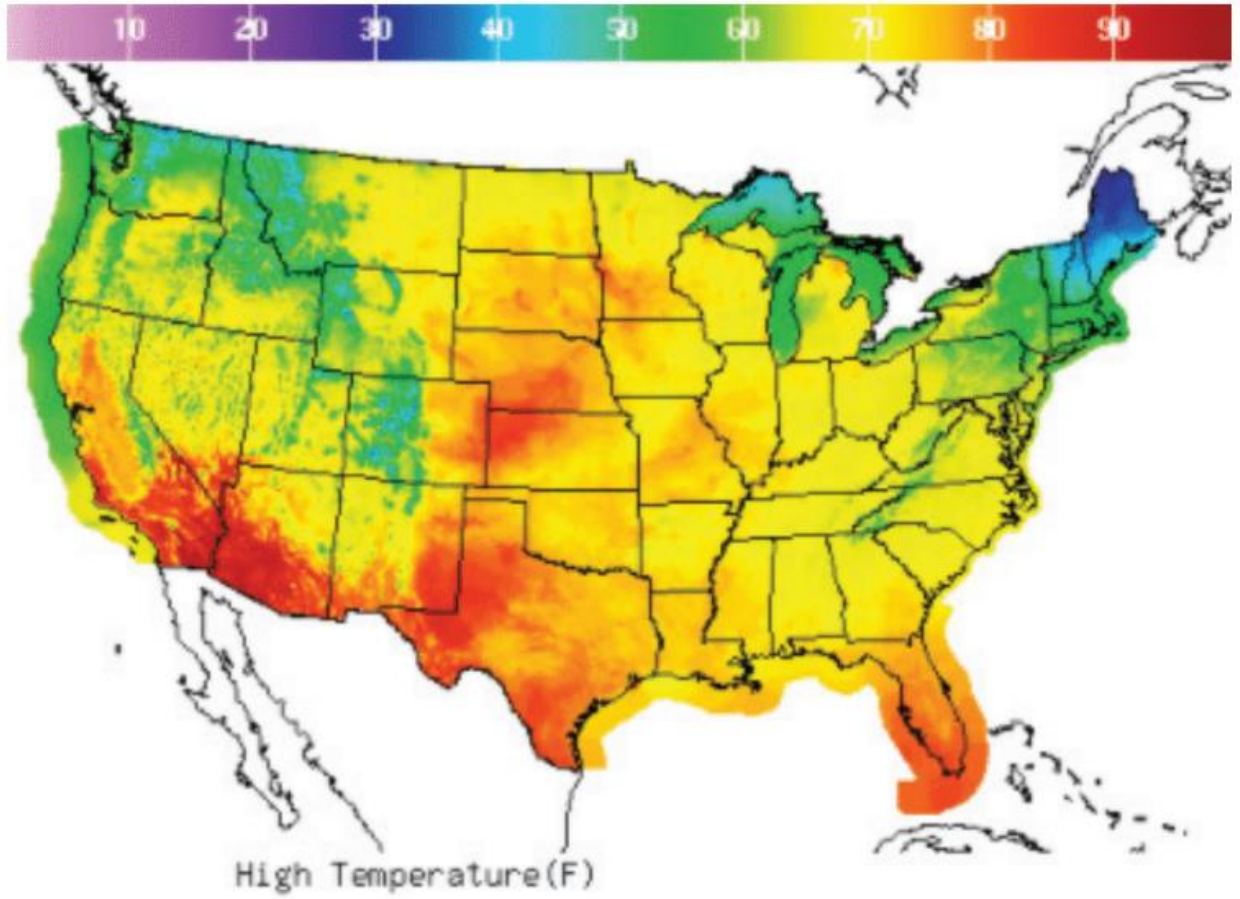


3.5 A Choropleth Map Example

A **choropleth map** is a geographic visualization that uses shades of color, different colors, or symbols to indicate quantitative or categorical geospatial data values.

The example shows a weather map using color to depict the daily high temperature (°F.)

- Warmer colors: higher °F
- Cooler colors: lower °F



CENGAGE

3.5 State-Level Choropleth Map of United States

DATAfile: *income_state*

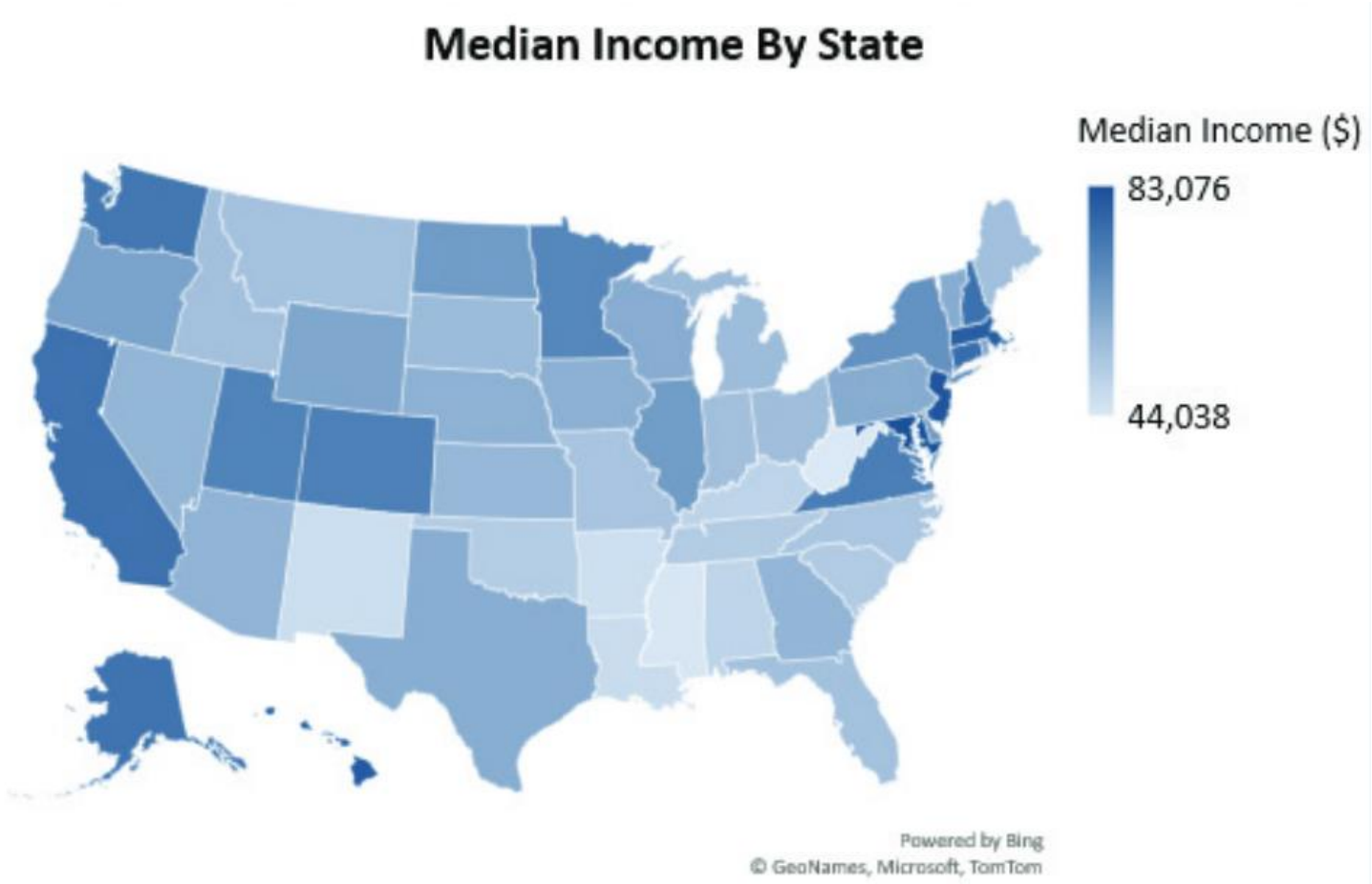
To create a map of the median income by state in the United States, follow these steps:

Step 1. Select cells A1:B51

Step 2. Click the **Insert** tab on the Ribbon

Step 3. In the **Charts** group, click the **Maps** button

Step 4. Select **Filled Map**



3.5 County-Level Choropleth Map of United States

DATAfile: *income_county*

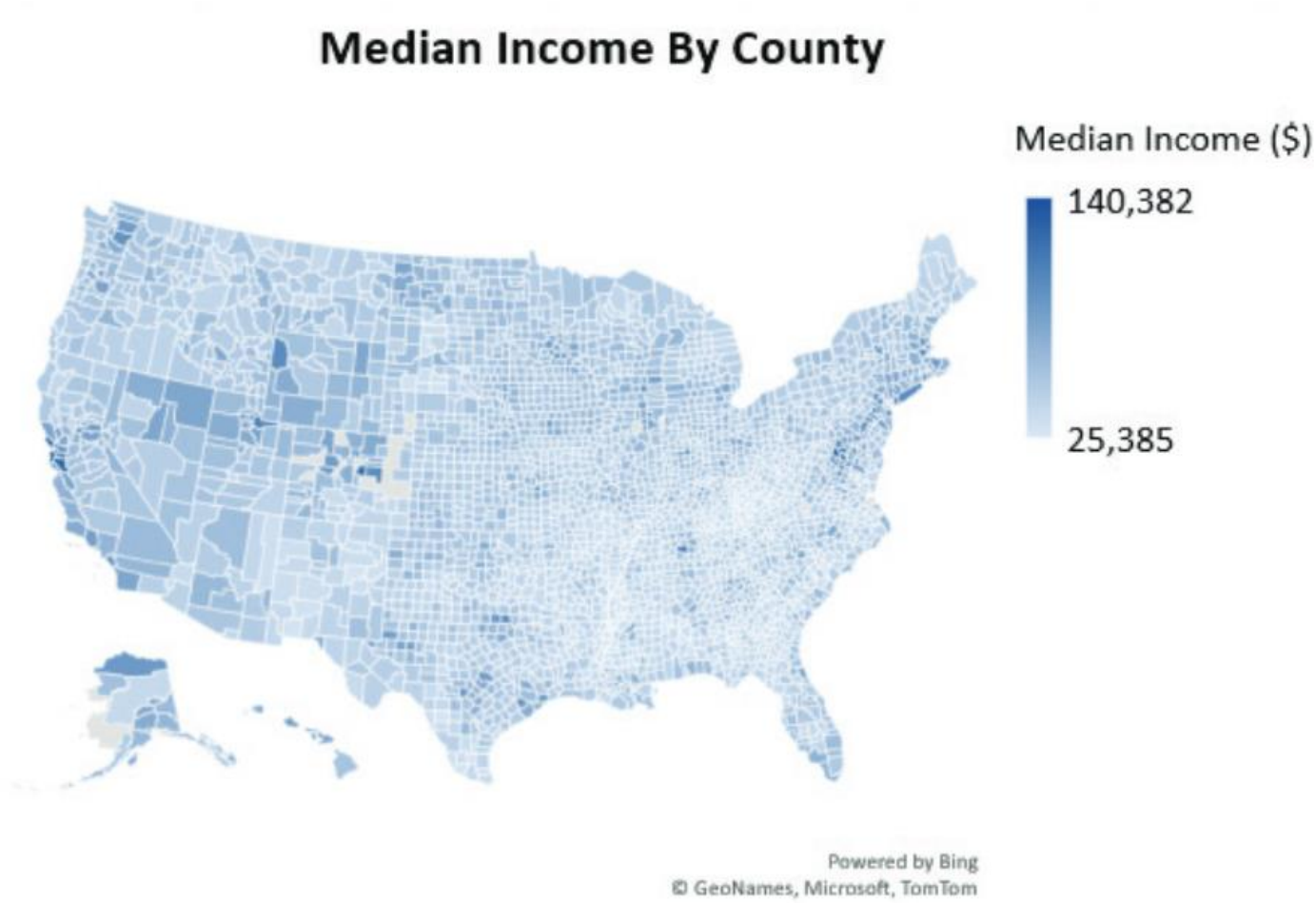
To create a map of the median income by county in the United States, follow these steps:

Step 1. Select cells A1:C3148

Step 2. Click the **Insert** tab on the Ribbon

Step 3. In the **Charts** group, click the **Maps** button

Step 4. Select **Filled Map**



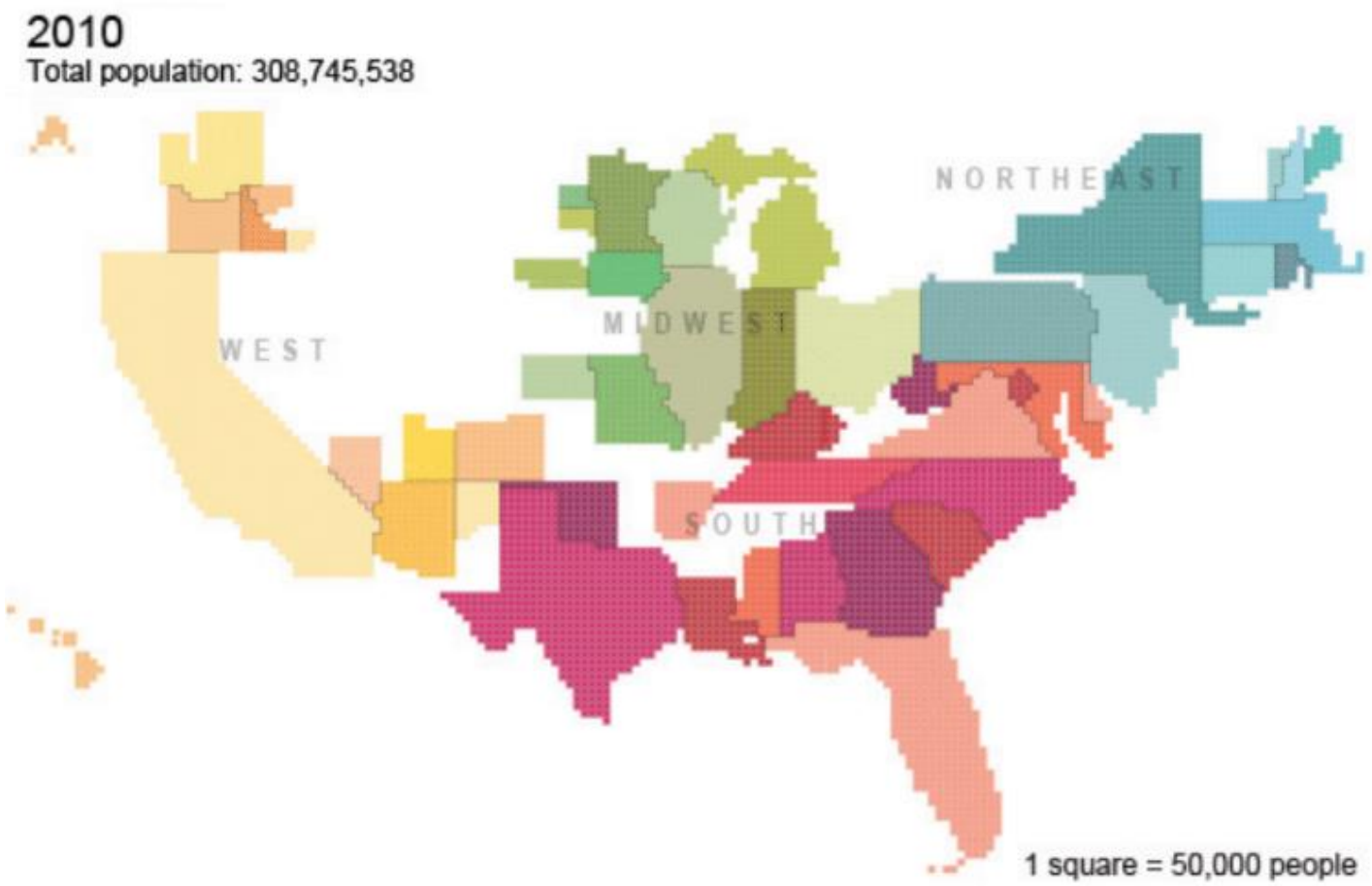
CENGAGE

3.5 State-Population Cartogram of the United States

A **cartogram** is a map-like diagram that uses geographic positioning to represent map regions that do not necessarily correspond to land area.

In the cartogram of the United States to the right, the area for each state is based on its population.

A weakness of the cartogram is that the distortions brought by relative sizing may render its geography meaningless.



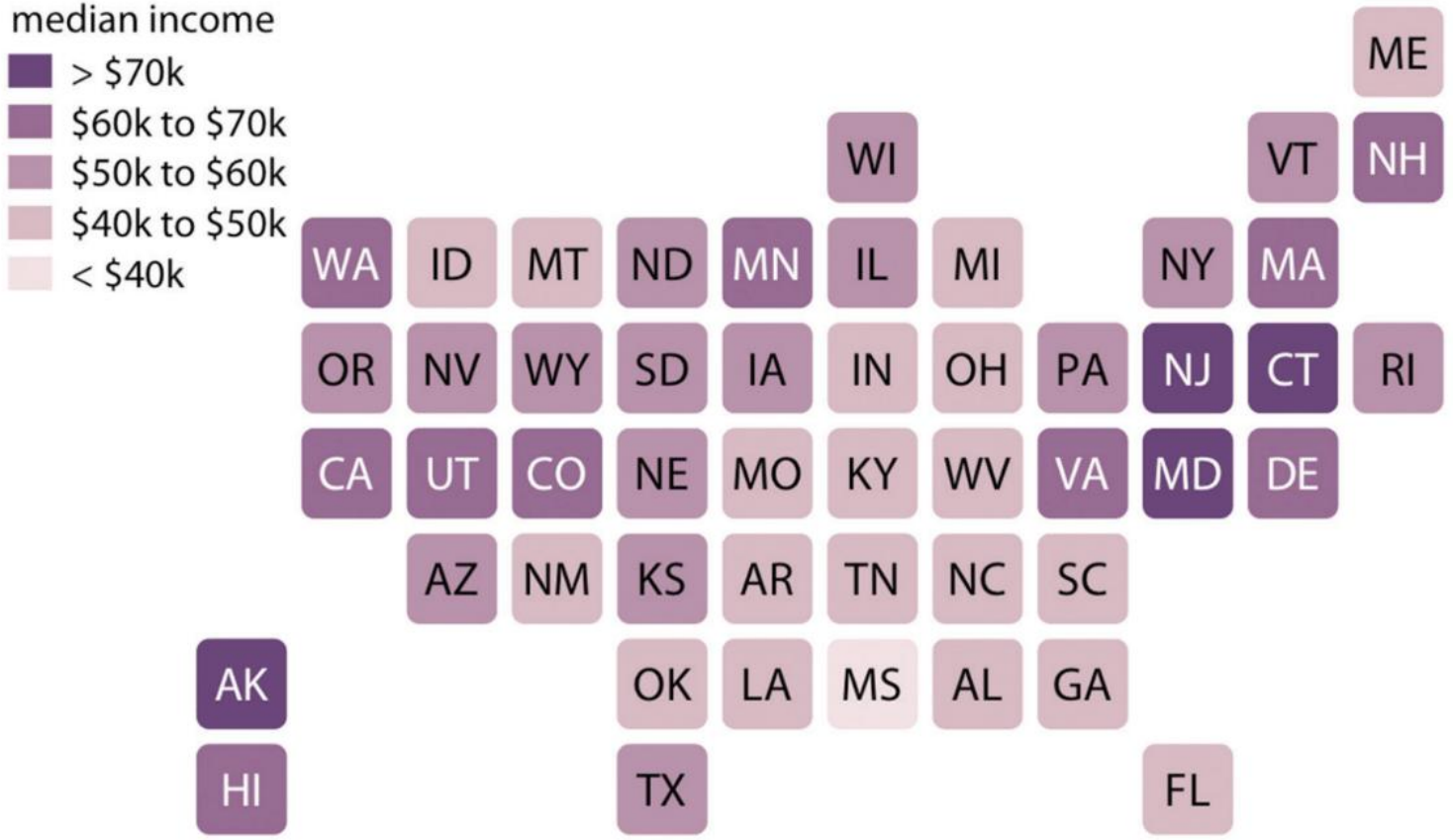
CENGAGE

3.5 Equal-Area Cartogram of the United States

An equal-area cartogram may be used to remedy the distortions a cartogram introduces.

An **equal-area cartogram** provides a balanced visual representation of each state while maintaining fidelity to relative geographic positioning.

Cartograms are not available in Excel.



3.6 Data Dashboards

In a business, **key performance indicators (KPIs)** are metrics indicative of current operating characteristics.

A **data dashboard** is a data visualization tool that illustrates multiple KPIs and automatically updates as new data becomes available.

For effective data dashboards, the following principles apply:

- Show on a single screen all KPIs related to one aspect of the operations.
- Create multiple dashboards of related KPIs to prevent scrolling.
- The KPIs should convey clear meaning and be related to decisions.
- Dashboards must adhere to the principles of effective data visualization.
- Use color to call attention and differentiate between categorical variables.

3.6 An Application of Data Dashboards

Grogan Oil Company has offices in Houston, Dallas, and Austin (HQs). It uses data dashboards to monitor the operations of its IT call center in Austin, Texas.

- The stacked-column chart shows the call volume per type of problem over time.
- The column chart shows the % time that employees spent on each problem type.
- The bar charts in the middle show unresolved cases by time, and call volume by the office.
- The histogram shows the distribution of time needed to resolve a case.

Summary

- In this chapter, we covered techniques and tools related to data visualization.
- We discussed principles related to preattentive attributes and data-ink ratio.
- We introduced techniques for enhancing visual presentations.
- We explained when tables are preferable to charts for data visualization.
- We introduced crosstabulation as a form of a table for two variables and explained how to use Excel to create a PivotTable.
- We presented a variety of standard and specialized charts that can be used for many different data visualization applications.
- We presented choropleth maps and cartograms as visualizations that can be used with geospatial data.
- We introduced data dashboards as a data-visualization tool that summarizes a firm’s operations in visual form.