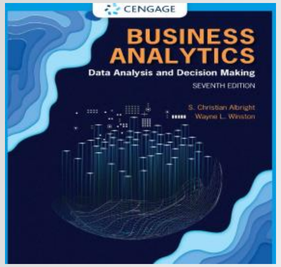
DS342 - Data Analytics



Chapter 2

Describing the Distribution of a Single Variable

Why Spreadsheets?

- Many commercial software packages can be used for Business Analytics.
- Spreadsheet software, such as Microsoft Excel, is widely available and used across all areas of business.
- Spreadsheets provide a flexible modeling environment for manipulating data and developing and solving models.

Basic Excel Skills

- Opening, saving, and printing files
- Using workbooks and worksheets
- Moving around a spreadsheet
- Selecting cells and ranges
- Inserting/deleting rows and columns
- Entering and editing text, data, and formulas
- Formatting data (number, currency, decimal)
- Working with text strings
- Formatting data and text
- Modifying the appearance of a spreadsheet

Basic Excel Functions

- =MIN(*range*)
- =MAX(*range*)
- =SUM(*range*)
- =AVERAGE(*range*)
- =COUNT(*range*)
- =COUNTIF(*range,criteria*)
  - Excel has other useful COUNT-type functions: COUNTA counts the number of nonblank cells in a range, and COUNTBLANK counts the number of blank cells in a range. In addition, COUNTIFS(*range1, criterion1, range2, criterion2,… range_n, criterion_n*) finds the number of cells within multiple ranges that meet specific criteria for each range.

Relative and Absolute References

- Cell references can be **relative** or **absolute**. Using a dollar sign before a row and/or column label creates an absolute reference.
  - Relative references: A2, C5, D10
  - Absolute references: A2, $C5, D$10
- Using a $ sign before a <u>row label</u> (for example, B$4) keeps the reference fixed to row 4 but allows the column reference to change if the formula is copied to another cell.
- Using a $ sign before a <u>column label</u> (for example, $B4) keeps the reference to column B fixed but allows the row reference to change.
- Using a $ sign before <u>both the row and column labels</u> (for example, B4) keeps the reference to cell B4 fixed no matter where the formula is copied.

Example 2.2 Using Basic Excel Functions

	A	B	C	D	E	F	G	H	I	J
1	**Purchase Orders**									
2										
3	**Supplier**	**Order No.**	**Item No.**	**Item Description**	**Item Cost**	**Quantity**	**Cost per order**	**A/P Terms (Months)**	**Order Date**	**Arrival Date**
4	Hulkey Fasteners	Aug11001	1122	Airframe fasteners	$ 4.25	19,500	$ 82,875.00	30	08/05/11	08/13/11
5	Alum Sheeting	Aug11002	1243	Airframe fasteners	$ 4.25	10,000	$ 42,500.00	30	08/08/11	08/14/11
6	Fast-Tie Aerospace	Aug11003	5462	Shielded Cable/ft.	$ 1.05	23,000	$ 24,150.00	30	08/10/11	08/15/11
7	Fast-Tie Aerospace	Aug11004	5462	Shielded Cable/ft.	$ 1.05	21,500	$ 22,575.00	30	08/15/11	08/22/11
8	Steelpin Inc.	Aug11005	5319	Shielded Cable/ft.	$ 1.10	17,500	$ 19,250.00	30	08/20/11	08/31/11
9	Fast-Tie Aerospace	Aug11006	5462	Shielded Cable/ft.	$ 1.05	22,500	$ 23,625.00	30	08/20/11	08/26/11
10	Steelpin Inc.	Aug11007	4312	Bolt-nut package	$ 3.75	4,250	$ 15,937.50	30	08/25/11	09/01/11
11	Durrable Products	Aug11008	7258	Pressure Gauge	$ 90.00	100	$ 9,000.00	45	08/25/11	08/28/11
12	Fast-Tie Aerospace	Aug11009	6321	O-Ring	$ 2.45	1,300	$ 3,185.00	30	08/25/11	09/04/11
96	Steelpin Inc.	Nov11009	5677	Side Panel	$ 195.00	110	$ 21,450.00	30	11/05/11	11/17/11
97	Manley Valve	Nov11010	9955	Door Decal	$ 0.55	125	$ 68.75	30	11/05/11	11/10/11
98										
99	Minimum Quantity	90		=MIN(F4:F97)						
100	Maximum Quantity	25,000		=MAX(F4:F97)						
101	Total Order Costs	$ 2,471,760.00		=SUM(G4:G97)						
102	Average Number of A/P Months	30.63829787		=AVERAGE(H4:H97)						
103	Number of Purchase Orders	94		=COUNT(B4:B97)						
104	Number of O-ring Orders	12		=COUNTIF(D4:D97,"=O-Ring")						
105	Number of A/P Terms < 30	17		=COUNTIF(H4:H97,"<30")						
106	Number of O-ring Orders Spacetime	3		=COUNTIFS(D4:D97,"O-Ring",A4:A97,"Spacetime Technologies")						

Other IF-Type Functions

- SUMIF, AVERAGEIF, SUMIFS, and AVERAGEIFS can be used to embed IF logic within mathematical functions.
- For instance, the syntax of SUMIF is
  - SUMIF(*range, criterion, [sum range]*). "Sum range" is an optional argument that allows you to add cells in a different range.
- Example: In the *Purchase Orders* database, to find the total cost of all airframe fasteners, use

  =SUMIF(D4:D97,"Airframe fasteners", G4:G97)

Logical Functions

- =IF(*condition, value if true, value if false*) – a returns one value if the condition is true and another if the condition is false,
- =AND(*condition1, condition2, …*) – returns TRUE if all conditions are true and FALSE if not,
- =OR(*condition1, condition2, …*) – returns TRUE if any condition is true and FALSE if not.

IF Function

- =IF(*condition, value if true, value if false*)
- Conditions may include the following:

  = equal <> not equal to

  > greater than >= greater than or equal to

  < less than <= less than or equal to

- You may nest up to 7 IF functions, replacing the *value if false* with another IF function
- Example:

  =IF(A8 =2,(IF(B3 =5,"YES"," ")),15)

Lookup Functions for Database Queries

- These functions are useful for finding specific data in a spreadsheet.
- =VLOOKUP(*lookup_value, table_array, col_index_num, [range lookup]*) - looks up a value in the leftmost column of a table and returns a value in the same row from a column you specify
- =HLOOKUP(*lookup_value, table_array, row_index_num, [range lookup]*) - looks up a value in the top row of a table and returns a value in the same column from a row you specify.

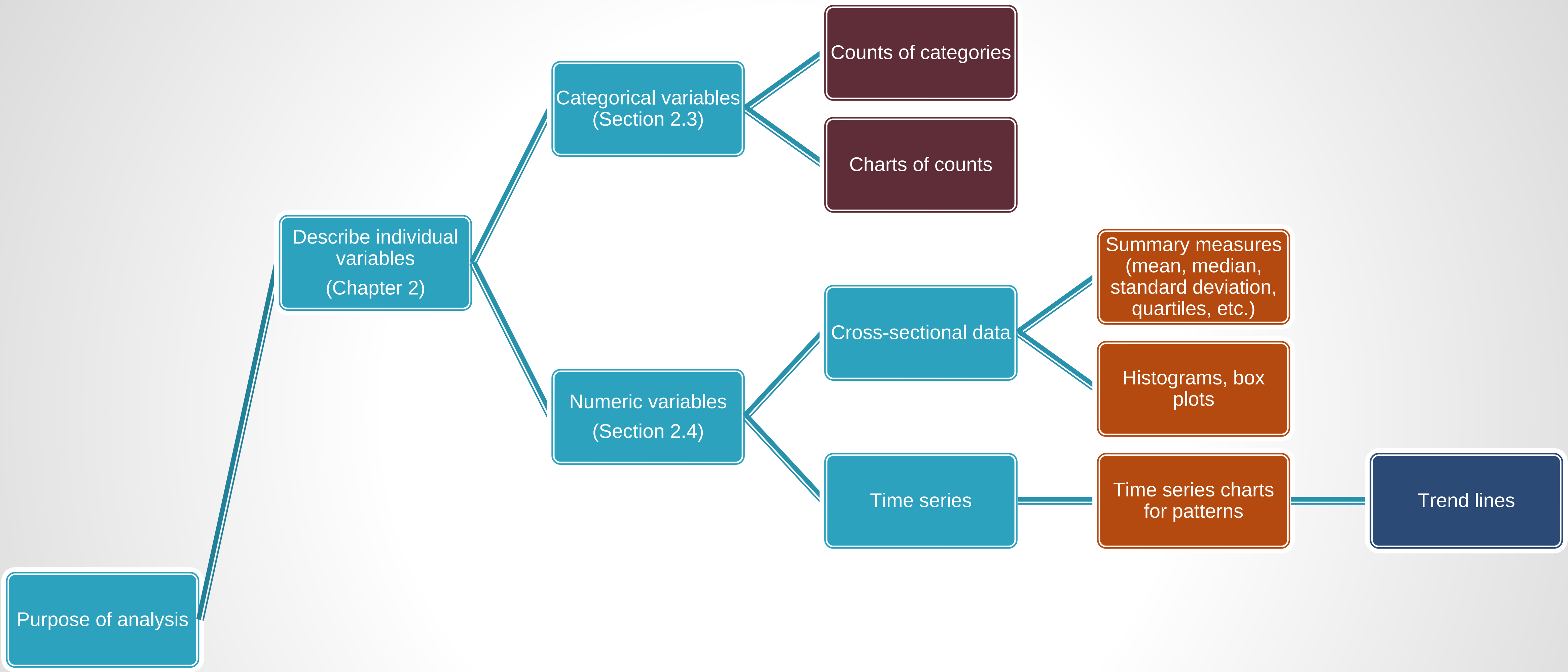
Important Notes on Lookup Functions

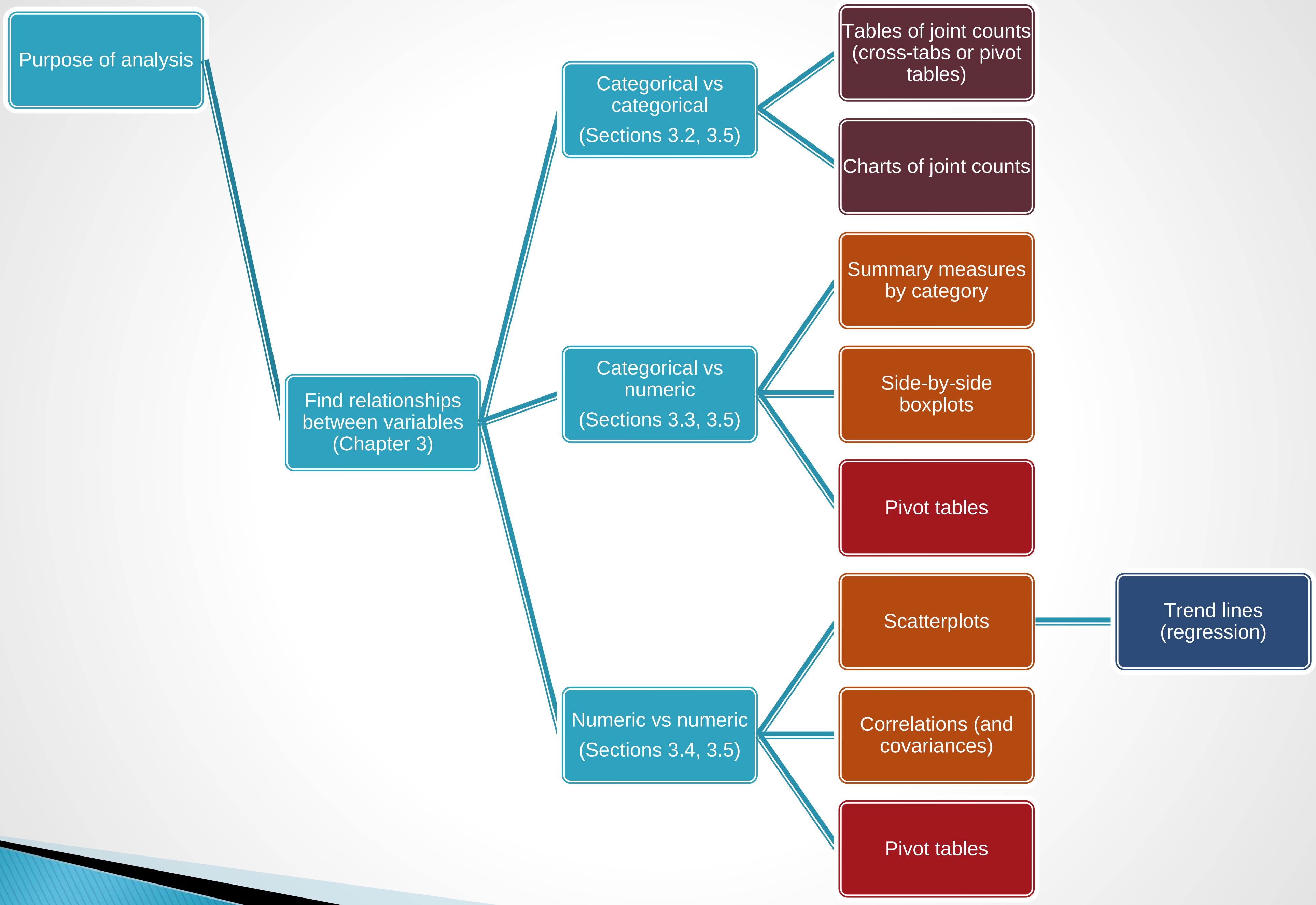
- In the VLOOKUP and HLOOKUP functions, *range lookup* is optional. If this is omitted or set as *True*, then the first column of the table must be sorted in ascending numerical order.
- If an exact match for the *lookup_value* is found in the first column, then Excel will return the value the *col_index_num* of that row. If an exact match is not found, Excel will choose the row with the largest value in the first column that is less than the l*ookup_value*.
- If range lookup is *False*, then Excel seeks an exact match in the first column of the table range. If no exact match is found, Excel will return #N/A (not available).
- We recommend that you specify the range lookup to avoid errors.

Example 2.4 Using the IF Function, Ex: Purchase Orders

- Suppose that orders with quantities of at least 10,000 units are classified as Large.
  - Cell K4: =IF(F4>=10000, “Large”, “Small”)
- Suppose that large orders with a total cost of at least $25,000 are considered critical.
  - Cell L4: =IF(AND(K4=“Large”, G4>=25000),“Critical”,“”)

	A	B	C	D	E	F	G	H	I	J	K	L
1	Purchase Orders											
2												
3	Supplier	Order No.	Item No.	Item Description	Item Cost	Quantity	Cost per order	A/P Terms (Months)	Order Date	Arrival Date	Order Size	Type
4	Hulkey Fasteners	Aug11001	1122	Airframe fasteners	$ 4.25	19,500	$ 82,875.00	30	08/05/11	08/13/11	Large	Critical
5	Alum Sheeting	Aug11002	1243	Airframe fasteners	$ 4.25	10,000	$ 42,500.00	30	08/08/11	08/14/11	Large	Critical
6	Fast-Tie Aerospace	Aug11003	5462	Shielded Cable/ft.	$ 1.05	23,000	$ 24,150.00	30	08/10/11	08/15/11	Large	
7	Fast-Tie Aerospace	Aug11004	5462	Shielded Cable/ft.	$ 1.05	21,500	$ 22,575.00	30	08/15/11	08/22/11	Large	
8	Steelpin Inc.	Aug11005	5319	Shielded Cable/ft.	$ 1.10	17,500	$ 19,250.00	30	08/20/11	08/31/11	Large	
9	Fast-Tie Aerospace	Aug11006	5462	Shielded Cable/ft.	$ 1.05	22,500	$ 23,625.00	30	08/20/11	08/26/11	Large	
10	Steelpin Inc.	Aug11007	4312	Bolt-nut package	$ 3.75	4,250	$ 15,937.50	30	08/25/11	09/01/11	Small	
11	Durrable Products	Aug11008	7258	Pressure Gauge	$ 90.00	100	$ 9,000.00	45	08/25/11	08/28/11	Small	
12	Fast-Tie Aerospace	Aug11009	6321	O-Ring	$ 2.45	1,300	$ 3,185.00	30	08/25/11	09/04/11	Small	
13	Fast-Tie Aerospace	Aug11010	5462	Shielded Cable/ft.	$ 1.05	22,500	$ 23,625.00	30	08/25/11	09/02/11	Large	
14	Steelpin Inc.	Aug11011	5319	Shielded Cable/ft.	$ 1.10	18,100	$ 19,910.00	30	08/25/11	09/05/11	Large	
15	Hulkey Fasteners	Aug11012	3166	Electrical Connector	$ 1.25	5,600	$ 7,000.00	30	08/25/11	08/29/11	Small	


Purpose of analysis
Describe individual variables
(Chapter 2)
Categorical variables
(Section 2.3)
Counts of categories
Charts of counts
Numeric variables
(Section 2.4)
Cross-sectional data
Summary measures (mean, median, standard deviation, quartiles, etc.)
Histograms, box plots
Time series
Time series charts for patterns
Trend lines


Purpose of analysis
Find relationships between variables (Chapter 3)
Categorical vs categorical
(Sections 3.2, 3.5)
Categorical vs numeric
(Sections 3.3, 3.5)
Numeric vs numeric
(Sections 3.4, 3.5)
Tables of joint counts (cross-tabs or pivot tables)
Charts of joint counts
Summary measures by category
Side-by-side boxplots
Pivot tables
Scatterplots
Trend lines (regression)
Correlations (and covariances)
Pivot tables

Types of Data
(slide 1 of 5)

- A variable is **numerical** if meaningful arithmetic can be performed on it.
- Otherwise, the variable is **categorical**.
- There is also a third **data type**, a **date** variable.
  - Excel® stores dates as numbers, but dates are treated differently from typical numbers.
- A categorical variable is **ordinal** if there is a natural ordering of its possible values.
- If there is no natural ordering, it is **nominal**.

Types of Data
(slide 2 of 5)

- Categorical variables can be coded numerically or left uncoded.
- A **dummy variable** is a 0–1 coded variable for a specific category.
  - It is coded as 1 for all observations in that category and 0 for all observations not in that category.
- Categorizing a numerical variable by putting the data into discrete categories (called **bins**) is called **binning** or **discretizing**.
  - A variable that has been categorized in this way is called a **binned** or **discretized variable**.

Environmental Data
Using a Different Coding, Ex: Questionnaire

	A	B	C	D	E	F	G
1	Person	Age	Gender	State	Children	Salary	Opinion
2	1	Middle-aged	1	Minnesota	1	$65,400	Strongly agree
3	2	Elderly	0	Texas	2	$62,000	Strongly disagree
4	3	Middle-aged	1	Ohio	0	$63,200	Neutral
5	4	Middle-aged	1	Florida	2	$52,000	Strongly agree
6	5	Young	0	California	3	$81,400	Strongly disagree
7	6	Young	0	New York	3	$46,300	Strongly agree
8	7	Elderly	0	Minnesota	2	$49,600	Strongly disagree
9	8	Middle-aged	1	New York	1	$45,900	Strongly agree
10	9	Middle-aged	1	Texas	3	$47,700	Agree
11	10	Young	0	Texas	1	$59,900	Agree
12	11	Middle-aged	1	New York	1	$48,100	Agree
13	12	Middle-aged	0	Virginia	0	$58,100	Neutral
14	13	Middle-aged	0	Illinois	2	$56,000	Strongly disagree
15	14	Middle-aged	0	Virginia	2	$53,400	Strongly disagree
16	15	Middle-aged	0	New York	2	$39,000	Disagree
17	16	Middle-aged	1	Michigan	1	$61,500	Disagree
18	17	Middle-aged	1	Ohio	0	$37,700	Strongly disagree
19	18	Middle-aged	0	Michigan	2	$36,700	Agree
28	27	Young	1	Illinois	3	$45,400	Disagree
29	28	Elderly	1	Michigan	2	$53,900	Strongly disagree
30	29	Middle-aged	1	California	1	$44,100	Neutral
31	30	Middle-aged	0	New York	2	$31,000	Agree

Note the formulas in columns B, C, and G that generate this recoded data. The formulas in columns B and G are based on the lookup tables below.

Age lookup table (range name AgeLookup)

0	Young
35	Middle-aged
60	Elderly

Opinion lookup table (range name OpinionLookup)

1	Strongly disagree
2	Disagree
3	Neutral
4	Agree
5	Strongly agree

Types of Data
(slide 4 of 5)

- A numerical variable is **discrete** if it results from a count, such as the number of children.
- A **continuous** variable is the result of an essentially continuous measurement, such as weight or height.
- **Cross-sectional** data are data on a cross section of a population at a distinct point in time.
- **Time series** data are data collected over time.

Descriptive Measures for Categorical Variables

- There are only a few possibilities for describing a categorical variable, all based on *counting*:
  - Count the number of categories.
  - Give the categories names.
  - Count the number of observations in each category (referred to as the **count of categories**).
    - Once you have the counts, you can display them graphically, usually in a column chart or a pie chart.

Example 2.2:
Supermarket Transactions.xlsx (slide 1 of 3)

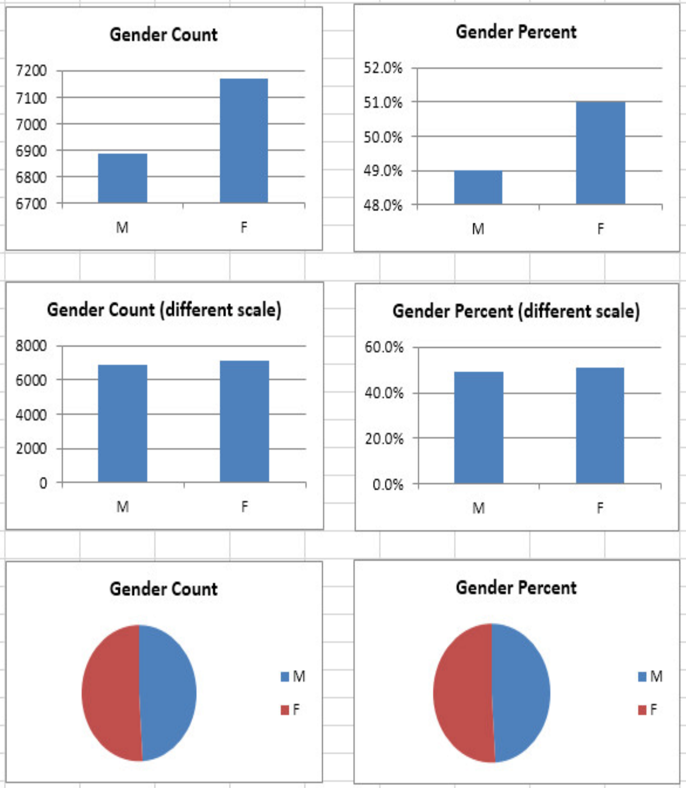
- **Objective**: To summarize categorical variables in a large data set.
- **Solution**: Data set contains transactions made by supermarket customers over a two-year period.
- Children, Units Sold, and Revenue are numerical.
- Purchase Date is a date variable.
- Transaction and Customer ID are used only for identification.
- All of the other variables are categorical.

	A	B	C	D	E	F	G	H	I	J	K	O	P
1	Transaction	Purchase Date	Customer ID	Gender	Marital Status	Homeowner	Children	Annual Income	City	State or Province	Country	Units Sold	Revenue
2	1	12/18/2011	7223	F	S	Y	2	$30K - $50K	Los Angeles	CA	USA	5	$27.38
3	2	12/20/2011	7841	M	M	Y	5	$70K - $90K	Los Angeles	CA	USA	5	$14.90
4	3	12/21/2011	8374	F	M	N	2	$50K - $70K	Bremerton	WA	USA	3	$5.52
5	4	12/21/2011	9619	M	M	Y	3	$30K - $50K	Portland	OR	USA	4	$4.44
6	5	12/22/2011	1900	F	S	Y	3	$130K - $150K	Beverly Hills	CA	USA	4	$14.00
7	6	12/22/2011	6696	F	M	Y	3	$10K - $30K	Beverly Hills	CA	USA	3	$4.37
8	7	12/23/2011	9673	M	S	Y	2	$30K - $50K	Salem	OR	USA	4	$13.78
9	8	12/25/2011	354	F	M	Y	2	$150K +	Yakima	WA	USA	6	$7.34
10	9	12/25/2011	1293	M	M	Y	3	$10K - $30K	Bellingham	WA	USA	1	$2.41
11	10	12/25/2011	7938	M	S	N	1	$50K - $70K	San Diego	CA	USA	2	$8.96

Example 2.2: Supermarket Transactions.xlsx (slide 2 of 3)

- To get the counts in column S, use Excel's *COUNTIF* function.
- To get the percentages in column T, divide each count by the total number of observations.
- When creating charts, be careful to use appropriate scales.

Categorical summaries		
Gender	Count	Percent
M	6889	49.0%
F	7170	51.0%
		100.0%
Marital Status	Count	Percent
S	7193	51.2%
M	6866	48.8%
		100.0%
Homeowner	Count	Percent
Y	8444	60.1%
N	5615	39.9%
		100.0%
Annual Income	Count	Percent
$10K - $30K	3090	22.0%
$30K - $50K	4601	32.7%
$50K - $70K	2370	16.9%
$70K - $90K	1709	12.2%
$90K - $110K	613	4.4%
$110K - $130K	643	4.6%
$130K - $150K	760	5.4%
$150K +	273	1.9%
		100.0%



Example 2.2:
Supermarket Transactions.xlsx (slide 3 of 3)

- Another efficient way to find counts for a categorical variable is to use dummy (0–1) variables.
  - Recode each variable so that one category is replaced by 1 and all others by 0.
    - This can be done using a simple IF formula.
  - Find the count of that category by summing the 0s and 1s.
  - Find the percentage of that category by averaging the 0s and 1s.

	A	B	C	D	E
1	Transaction	Purchase Date	Customer ID	Gender	Gender Dummy for M
2	1	12/18/2011	7223	F	0
3	2	12/20/2011	7841	M	1
4	3	12/21/2011	8374	F	0
5	4	12/21/2011	9619	M	1
6	5	12/22/2011	1900	F	0
7	6	12/22/2011	6696	F	0
8	7	12/23/2011	9673	M	1
9	8	12/25/2011	354	F	0
10	9	12/25/2011	1293	M	1
11	10	12/25/2011	7938	M	1
14055	14054	12/29/2013	2032	F	0
14056	14055	12/29/2013	9102	F	0
14057	14056	12/29/2013	4822	F	0
14058	14057	12/31/2013	250	M	1
14059	14058	12/31/2013	6153	F	0
14060	14059	12/31/2013	3656	M	1
14061				Count	6889
14062				Percent	49.0%

Descriptive Measures for Numerical Variables

- There are many ways to summarize numerical variables, both with numerical summary measures and with charts.
- To learn how the values of a variable are distributed, ask:
  - What are the most "typical" values?
  - How spread out are the values?
  - What are the "extreme" values on either end?
  - Is the chart of the values symmetric about some middle value, or is it skewed in some direction? Does it have any other peculiar features besides possible skewness?

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

- **Objective**: To learn how salaries are distributed across all 2011 MLB players.
- **Solution**: Data set contains data on 843 Major League Baseball players in the 2011 season.
- Variables are player's name, team, position, and salary.
- Create summary measures of baseball salaries using Excel functions.

	A	B	C	D
1	**Player**	**Team**	**Position**	**Salary**
2	A.J. Burnett	New York Yankees	Pitcher	$16,500,000
3	A.J. Ellis	Los Angeles Dodgers	Catcher	$421,000
4	A.J. Pierzynski	Chicago White Sox	Catcher	$2,000,000
5	Aaron Cook	Colorado Rockies	Pitcher	$9,875,000
6	Aaron Crow	Kansas City Royals	Pitcher	$1,400,000
7	Aaron Harang	San Diego Padres	Pitcher	$3,500,000
8	Aaron Heilman	Arizona Diamondbacks	Pitcher	$2,000,000
9	Aaron Hill	Toronto Blue Jays	Second Baseman	$5,000,000
10	Aaron Laffey	Seattle Mariners	Pitcher	$431,600
11	Aaron Miles	Los Angeles Dodgers	Second Baseman	$500,000
12	Aaron Rowand	San Francisco Giants	Outfielder	$13,600,000
13	Adam Dunn	Chicago White Sox	Designated Hitter	$12,000,000
14	Adam Everett	Cleveland Indians	Shortstop	$700,000

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

	A	B	C	D	E	F
1	**Measures of central tendency**				**Measures of variability**	
2	Mean	$3,305,055			Range	$31,586,000
3	Median	$1,175,000			Interquartile range	$3,875,925
4	Mode	$414,000	57		Variance	20,563,887,478,833
5					Standard deviation	$4,534,742
6	**Min, max, percentiles, quartiles**				Mean absolute deviation	$3,249,917
7	Min	$414,000				
8	Max	$32,000,000			**Measures of shape**	
9	P01	$414,000	0.01		Skewness	2.2568
10	P05	$414,000	0.05		Kurtosis	5.7233
11	P10	$416,520	0.10			
12	P20	$424,460	0.20		**Percentages of values less than given values**	
13	P50	$1,175,000	0.50		Value	Percentage less than
14	P80	$5,500,000	0.80		$1,000,000	46.38%
15	P90	$9,800,000	0.90		$1,500,000	54.69%
16	P95	$13,590,000	0.95		$2,000,000	58.36%
17	P99	$20,000,000	0.99		$2,500,000	63.23%
18	Q1	$430,325	1		$3,000,000	66.55%
19	Q2	$1,175,000	2			
20	Q3	$4,306,250	3			

Measures of Central Tendency
(slide 1 of 3)

- The **mean** is the average of all values.
  - If the data set represents a sample from some larger population, this measure is called the **sample mean** and is denoted by $\overline{X}$.
  - If the data set represents the entire population, it is called the **population mean** and is denoted by μ.

$$\text{Mean} = \frac{\sum_{i=1}^{n} X_i}{n}$$

- In Excel, the mean can be calculated with the *AVERAGE* function.

Measures of Central Tendency

(slide 2 of 3)

- The **median** is the middle observation when the data are sorted from smallest to largest.
  - If the number of observations is odd, the median is literally the middle observation.
  - If the number of observations is even, the median is usually defined as the average of the two middle observations.
- In Excel, the median can be calculated with the *MEDIAN* function.

Measures of Central Tendency
(slide 3 of 3)

- The **mode** is the value that appears most often.
  - In most cases where a variable is essentially continuous, the mode is not very interesting because it is often the result of a few lucky ties.
  - However, it is not always a result of luck and may reveal interesting information.
- In Excel, the mode can be calculated with the *MODE* function.

Minimum, Maximum, Percentiles, and Quartiles

- For any percentage p, the pth **percentile** is the value such that a percentage p of all values are less than it.
- The **quartiles** divide the data into four groups, each with (approximately) a quarter of all observations.
  - The first, second and third quartiles are the percentiles corresponding to p = 25%, p = 50%,
    and p = 75%.
  - By definition, the second quartile (p = 50%) is equal to the median.
- The **minimum** and **maximum** values can be calculated with Excel's *MIN* and *MAX* functions, and the percentiles and quartiles with Excel's *PERCENTILE* and *QUARTILE* functions.

Measures of Variability

(slide 1 of 3)

- The **range** is the maximum value minus the minimum value.
- The **interquartile range** (**IQR**) is the third quartile minus the first quartile.
  - Thus, it is the range of the middle 50% of the data.
  - It is less sensitive to extreme values than the range.
- The **variance** is essentially the average of the squared deviations from the mean.
  - If X_i is a typical observation, its squared deviation from the mean is $(X_i - mean)^2$.

Measures of Variability
(slide 2 of 3)

- The **sample variance** is denoted by s^2, and the **population variance** by σ^2.

$$s^2 = \frac{\sum_{i=1}^{n} (X_i - \text{mean})^2}{n - 1} \qquad \sigma^2 = \frac{\sum_{i=1}^{n} (X_i - \text{mean})^2}{n}$$

  - If all observations are close to the mean, their squared deviations from the mean—and the variance—will be relatively small.
  - If at least a few of the observations are far from the mean, their squared deviations from the mean—and the variance—will be large.
  - In Excel, use the *VAR* function to obtain the sample variance and the *VARP* function to obtain the population variance.

Measures of Variability
(slide 3 of 3)

- A fundamental problem with variance is that it is in squared units (e.g., \$ → $\$^2$).
- A more natural measure is the **standard deviation**, which is the square root of variance.
  - The **sample standard deviation**, denoted by *s*, is the square root of the sample variance.
  - The **population standard deviation**, denoted by *σ,* is the square root of the population variance.
  - In Excel, use the *STDEV* function to find the sample standard deviation or the *STDEVP* function to find the population standard deviation.

Coefficient of Variation

- The **coefficient of variation (CV)** provides a relative measure of dispersion in data relative to the mean:

$$CV = \frac{\text{standard deviation}}{\text{mean}}$$

  - Expressed as a percentage.
  - Provides a relative measure of risk to return.

Calculating Variance and Standard Deviation

	A	B	C	D	E	F
1	**Low variability supplier**				**High variability supplier**	
2						
3	Diameter1	Sq dev from mean			Diameter2	Sq dev from mean
4	102.61	6.610041			103.21	9.834496
5	103.25	10.310521			93.66	41.139396
6	96.34	13.682601			120.87	432.473616
7	96.27	14.205361			110.26	103.754596
8	103.77	13.920361			117.31	297.079696
9	97.45	6.702921			110.23	103.144336
10	98.22	3.308761			70.54	872.257156
11	102.76	7.403841			39.53	3665.575936
12	101.56	2.313441			133.22	1098.657316
13	98.16	3.530641			101.91	3.370896
14						
15	Mean				Mean	
16	100.039				100.074	
17						
18	Sample variance				Sample variance	
19	9.1098	9.1098			736.3653	736.3653
20						
21	Population variance				Population variance	
22	8.1988	8.1988			662.7287	662.7287
23						
24	Sample standard deviation				Sample standard deviation	
25	3.0182	3.0182			27.1361	27.1361
26						
27	Population standard deviation				Population standard deviation	
28	2.8634	2.8634			25.7435	25.7435

Excel *Descriptive Statistics* Tool

This tool provides a summary of numerical statistical measures for sample data.

Data >
Data Analysis >
Descriptive Statistics

- Enter *Input Range*
- *Labels* (optional)
- Check *Summary Statistics* box

<u>Note:</u>
Results of the *Analysis Toolpak* <u>do not change</u> when changes are made to the data.

- The data must be in a <u>single row or column</u>. If the data are in multiple columns, the tool treats each row or column as a **separate data set**

Measures of Shape
(slide 1 of 2)

- **Skewness** occurs when there is a lack of symmetry.
  - A variable can be **skewed to the right** (or **positively skewed**) because of some really *large* values (e.g., really large baseball salaries).
  - Or it can be **skewed to the left** (or **negatively skewed**) because of some really *small* values (e.g., temperature lows in Antarctica).
- In Excel, a measure of skewness can be calculated with the *SKEW* function.

Coefficient of Skewness

- Coefficient of Skewness (CS):

$$CS = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^3}{\sigma^3} \quad (4.11)$$

- Excel function: =SKEW(*data range*)
  - CS is negative for left-skewed data.
  - CS is positive for right-skewed data.
  - $|CS| > 1$ suggests high degree of skewness.
  - $0.5 \leq |CS| \leq 1$ suggests moderate skewness.
  - $|CS| < 0.5$ suggests relative symmetry.

Measures of Shape

(slide 2 of 2)

- **Kurtosis** has to do with the "fatness" of the tails of the distribution relative to the tails of a normal distribution.
- A distribution with high kurtosis has many more extreme observations.
- In Excel, kurtosis can be calculated with the *KURT* function.

Measures of Shape: Kurtosis

- **Kurtosis** refers to the peakedness (i.e., high, narrow) or flatness (i.e., short, flat-topped) of a histogram.
- The coefficient of kurtosis (CK) measures the degree of kurtosis of a population

$$\mathrm{CK} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^4}{\sigma^4} \qquad (4.12)$$

  - CK < 3 indicates the data is somewhat flat with a wide degree of dispersion.
  - CK > 3 indicates the data is somewhat peaked with less dispersion.
  - Excel function: =KURT(*data range*).

Numerical Summary Measures in the Status Bar and with Data Analysis add-in

- If you select multiple cells, summary measures appear for the selected cells in the status bar at the bottom of the Excel window.
  - You can choose the summary measures that appear by right-clicking the status bar and selecting your favorites.
- Although Excel's built-in functions can be used to calculate a number of summary measures, a much quicker way is to use the ***Data Analysis*** add-in.

Standardized Values

- A **standardized value**, commonly called a ***z*-score**, provides a relative measure of the distance an observation is from the mean, which is independent of the units of measurement.
- The *z*-score for the i^{th} observation in a data set is calculated as follows:

$$z_i = \frac{x_i - \bar{x}}{s}$$

  - Excel function:

Properties of z-Scores

- The numerator represents the distance that x_i is from the sample mean; a negative value indicates that x_i lies to the left of the mean, and a positive value indicates that it lies to the right of the mean. By dividing by the standard deviation, *s*, we scale the distance from the mean to express it in units of standard deviations. Thus,
  - a *z*-score of 1.0 means that the observation is one standard deviation to the right of the mean;
  - a *z*-score of -1.5 means that the observation is 1.5 standard deviations to the left of the mean.

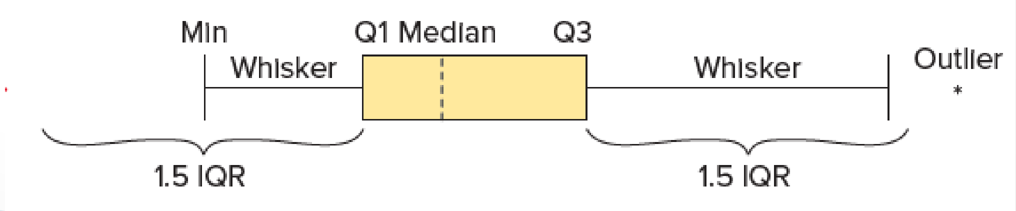
$$z_i = \frac{x_i - \bar{x}}{s}$$

Outliers

- An **outlier** is a value or an entire observation (row) that lies well outside of the norm.
- There is no standard definition of what constitutes an outlier.
- Some typical rules of thumb:
  - ✓ *z*-scores greater than +3 or less than -3
  - ✓ Extreme outliers are more than 3*IQR to the left of Q_1 or right of Q_3
  - ✓ Mild outliers are between 1.5*IQR and 3*IQR to the left of Q_1 or right of Q_3
- When dealing with outliers, it is best to run the analyses two ways: with the outliers and without them.

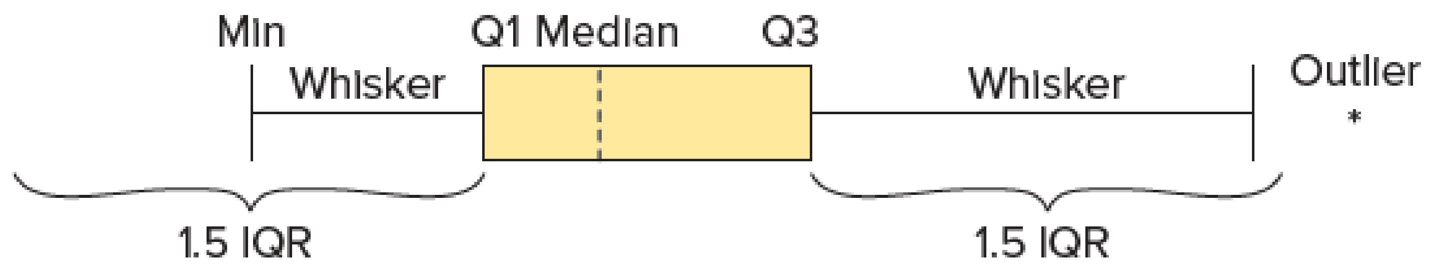
Outliers

- A common way to quickly summarize a variable is to use a five-number summary.
- A five-number summary shows the minimum, the quartiles (Q1, Q2, and Q3), and the maximum.
- A boxplot, also referred to as a box-and-whisker plot, is a way to graphically display a five-number summary.
  - Draw a box encompassing the first and third quartiles.
  - Draw a dashed vertical line in the box at the median.
  - Calculate the IQR. Draw a whisker that extends from Q1 to the minimum value that is not further from 1.5*IQR from Q1.
  - Similarly, draw a line that extends from Q3 to the maximum value that is not farther than 1.5*IQR from Q3.
  - Use an asterisk (or another symbol) to indicate observations that are farther than 1.5*QQR from the box. These observations are considered outliers.



Outliers

- A boxplot is also used to informally gauge the shape of the distribution.
- Symmetry is implied if the median is in the center of the box and the left/right whiskers are equidistant from their respective quartiles.
- If the median is left of center and the right whisker is longer than the left whisker, then the distribution is positively skewed.
- Similarly, if the median is right of center and the left whisker is longer than the right whisker, then the distribution is negatively skewed.
- If outliers exist, we need to include them when comparing the lengths of the left and right whiskers.



Missing Values

- Most real data sets have gaps in the data.
- There are two issues: how to detect these **missing values** and what to do about them.
- The more important issue is what to do about them:
  - One option is to simply ignore them. Then you will have to be aware of how the software deals with missing values.
  - Another option is to fill in missing values with the average of nonmissing values, but this isn’t usually a very good option.
  - A third option is to examine the nonmissing values in the *row* of a missing value; these values might provide clues on what the missing value should be.