CSEN 2141 : DATA ANALYTICS: DESCRIPTIVE, PREDICTIVE, PRESCRIPTIVE

Module-1

Introduction: Decision Making, Business Analytics Defined, A Categorization of Analytical Methods and Models, Big Data, Business Analytics in Practice, Legal and Ethical Issues in The Use of Data and Analytics.

Descriptive Statistics: Overview of Using Data: Definitions and Goals, Types of Data, Modifying Data in Excel, Creating Distributions from Data, Measures of Location, Measures of Variability, Analyzing Distributions, Measures of Association Between Two Variables.

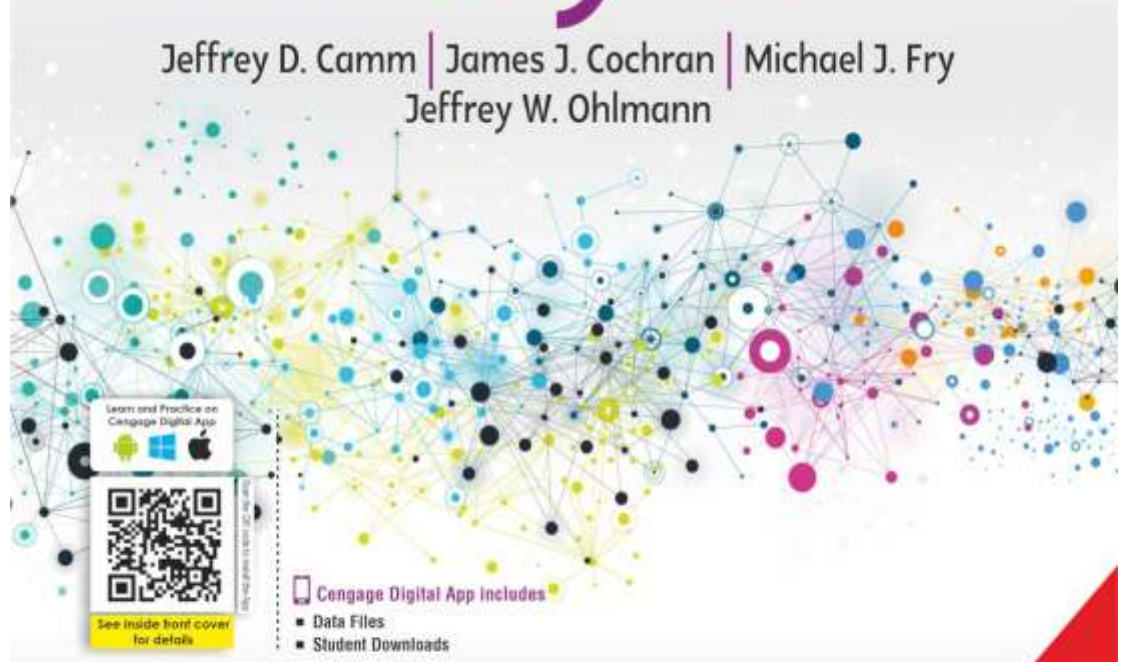
Cengage

Cengage MindTap

FOURTH EDITION

Business Analytics

Jeffrey D. Camm | James J. Cochran | Michael J. Fry
Jeffrey W. Ohlmann



MindTap Includes

Interactive eBook | Personalized Learning Path | Progress App | Homework Solutions | Mobile App | LMS Integration

See inside back cover for details

PART-1

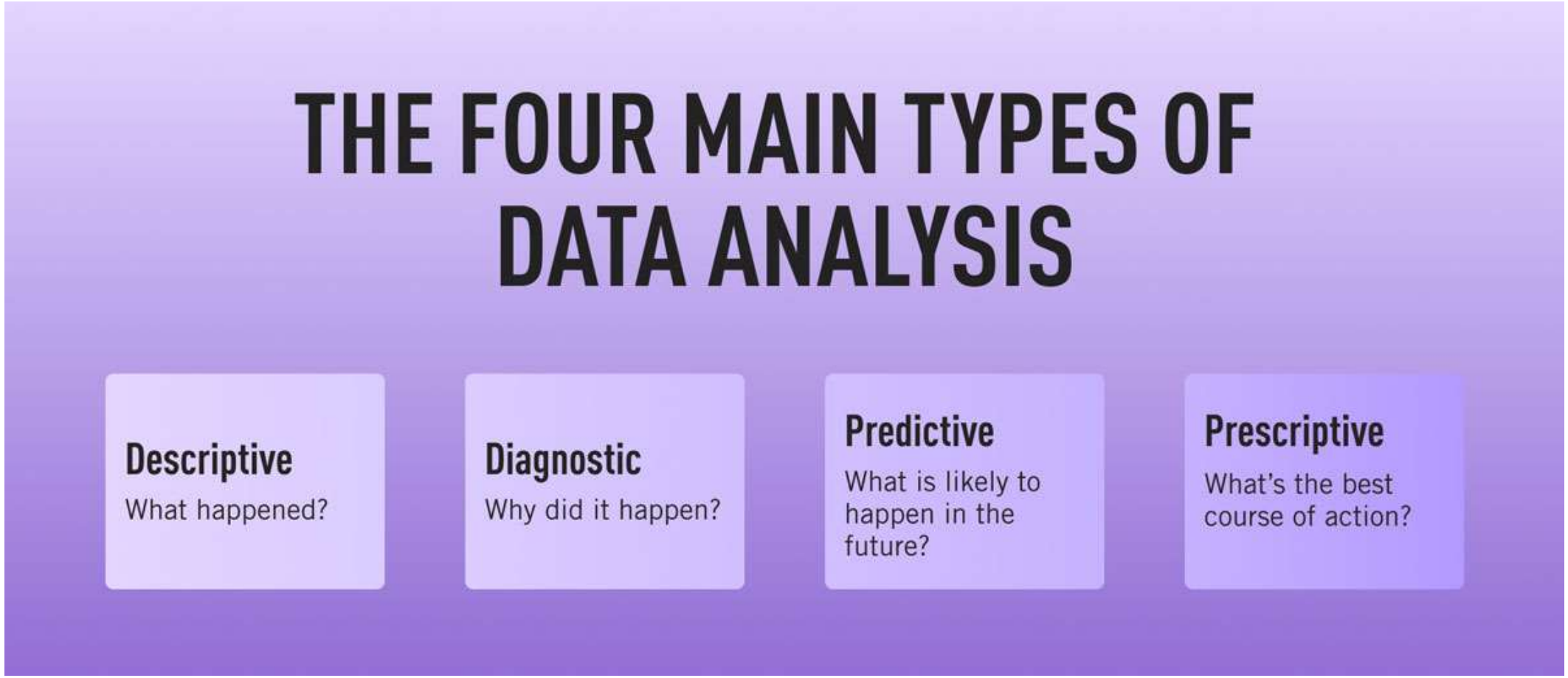
- **Introduction**: Decision Making, Business Analytics Defined, A Categorization of Analytical Methods and Models, Big Data, Business Analytics in Practice, Legal and Ethical Issues in The Use of Data and Analytics.

DATA ANALYTICS
?

DATA ANALYTICS

- Data analytics is the process of examining raw data to draw conclusions about that information. It involves collecting, transforming, and organizing data to reveal patterns, trends, and insights that can be used to make informed decisions.

DATA ANALYTICS



DECISION MAKING

- Decision making is the process of identifying a problem or opportunity, gathering information, evaluating alternatives, and selecting the best course of action.
- It’s the responsibility of the managers to plan, coordinate, organize, and lead their organization to better performance.

Decision Making

1. Strategic Decisions:

- ✓Involve higher-level issues concerned with the overall design of the organization
- ✓Define the organization's overall goals and aspirations for the future

2. Tactical Decisions:

- ✓Concern about how the organization should achieve the goals and objectives set by its strategy
- ✓Focus on 1-2 years planning.
- ✓Are usually the responsibility of midlevel management

3.Operational Decisions:

- ✓Affect how the firm is run from day to day
- ✓Are the domain of operations managers, who are the closest to the customer

- **Decision-Making Process:**

1. Identify and define the problem

2. Determine the criteria that will be used to evaluate alternative solutions

3. Determine the set of alternative solutions

4. Evaluate the alternatives

5. Choose an alternative

- **Common Approaches to Making Decisions:**

-Tradition -Intuition

-Rules of thumb

-Using the relevant data available


Strategic
(Senior Management)
Policy Decisions
Long Term
Complex
Non-routine
Tactical
(Middle Management)
How to achieve policy
Medium term
Less complex
Operational
(Junior Management)
Day-to-day decisions
Simple
Routine

Business Analytics Defined

- **What makes decision-making difficult?**

-Dearth of data

-Enormous number of alternatives and we cannot evaluate them all .

- **Business Analytics:**

It is a scientific process of transforming data into insight for making better decisions.

-Used for data-driven or fact-based decision-making, which is often seen as more objective than other alternatives for decision-making

Tools of Business Analytics Can Aid Decision Making by:

-Creating insights from data .

-Improving our ability to forecast for planning more accurately .

-Helping us quantify risk .

-Yielding better alternatives through analysis and optimization.

Business Analytics in Practice

- Business analytics involves tools as simple as reports and graphs to those that are as sophisticated as optimization, data mining, and simulation.

➢-Financial Analytics

➢-Human Resource (HR) Analytics

➢-Marketing Analytics

➢-Health Care Analytics

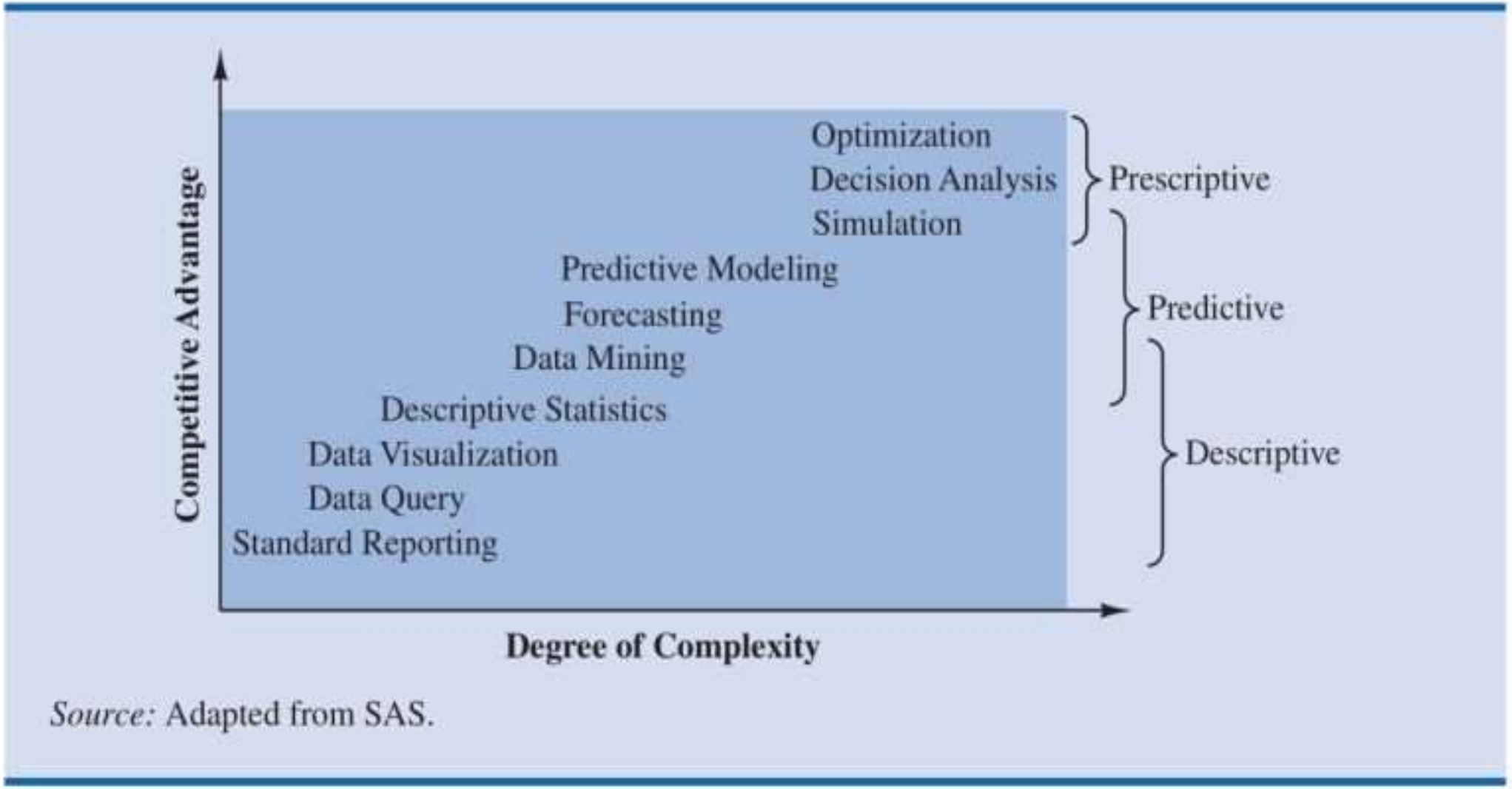
➢-Supply Chain Analytics

➢-Sports Analytics

➢-Web Analytics

Companies that apply analytics often follow a trajectory similar to the graph shown

FIGURE 1.2 THE SPECTRUM OF BUSINESS ANALYTICS



Source: Adapted from SAS.

A Categorization of Analytical Methods and Models:

1. **Descriptive Analytics**
2. **Predictive Analytics**
3. **Prescriptive Analytics**

- **Descriptive Analytics:** Encompasses the set of techniques that describe what has happened in the past;
- Examples include: Data queries, reports, descriptive statistics, data visualization including data boards, some data mining techniques etc.

Descriptive Analytics

- A data query is a request for information with certain characteristics from a database.
- Data dashboards are collections of tables,charts,maps and summary statistics that are updated as new data become available. Dashboards are used to help management monitor specific aspects of the company's performance related to their decision-making responsibilities.
- Data mining is the use of analytical techniques for better understanding patterns and relationships that exist in large data sets. For example , by analyzing text on social network platforms like Twitter, data mining techniques are used by companies to better understand their customers.

Predictive Analytics: Consists of techniques that use models constructed from past data to predict the future or ascertain the impact of one variable on another.

-Survey data and past purchase behavior may be used to help predict the market share of a new product

Techniques used in Predictive Analytics include :

-Linear regression

-Time series analysis

-Data mining is used to find patterns or relationships among elements of the data in a large database; often used in predictive analytics

-Simulation involves the use of probability and statistics to construct a computer model to study the impact of uncertainty on a decision

Prescriptive Analytics: Indicates the best course of action to take

-The output of a prescriptive model is the best decision

-A forecast or prediction, when combined with a rule, becomes a prescriptive model.

-Prescriptive models that rely on a rule or set of rules are often referred to as rule-based models

-Another type of modeling in the prescriptive analytics category is simulation optimization which combines the use of probability and statistics to model uncertainty with optimization techniques to find good decisions in highly complex and highly uncertain settings.

Big Data

-There is no universally accepted definition of big data. Big data is any set of data that is too large or too complex to be handled by standard data-processing techniques and typical desktop software.

-IBM describes the phenomenon of big data through the four Vs: Volume, Velocity, Variety, and Veracity.

Volume

Data at Rest

Terabytes to exabytes of existing data to process

Velocity

Data in Motion

Streaming data, milliseconds to seconds to respond

Variety

Data in Many Forms

Structured, unstructured, text, multimedia

Veracity

Data in Doubt

Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

Legal and Ethical Issues in the Use of Data and Analytics

- Data privacy laws are designed to protect individual's data from being used against their wishes.
- One of the strictest data privacy laws is the General Data Protection Regulation(GDPR) which went into effect in the European Union in May 2018.
- **INFORMS (Institute for Operations Research and the Management Sciences)** provide ethical guidelines for analysts.
- INFORMS also offers a set of Ethics guidelines for its members which covers ethical behavior for analytics professionals in three domains:

  1. Society

  2.Organizations(Businesses, government, nonprofit organization and universities)

  3. Profession(operational research and analytics).

TABLE 1.2 INFORMS Ethics Guidelines

Relative to Society

Analytics professionals should aspire to be:

- **Accountable** for their professional actions and the impact of their work.
- **Forthcoming** about their assumptions, interests, sponsors, motivations, limitations, and potential conflicts of interest.
- **Honest** in reporting their results, even when they fail to yield the desired outcome.
- **Objective** in their assessments of facts, irrespective of their opinions or beliefs.
- **Respectful** of the viewpoints and the values of others.
- **Responsible** for undertaking research and projects that provide positive benefits by advancing our scientific understanding, contributing to organizational improvements, and supporting social good.

Relative to Organizations

Analytics professionals should aspire to be:

- **Accurate** in our assertions, reports, and presentations.
- **Alert** to possible unintended or negative consequences that our results and recommendations may have on others.
- **Informed** of advances and developments in the fields relevant to our work.
- **Questioning** of whether there are more effective and efficient ways to reach a goal.
- **Realistic** in our claims of achievable results, and in acknowledging when the best course of action may be to terminate a project.
- **Rigorous** by adhering to proper professional practices in the development and reporting of our work.

Relative to the Profession

Analytics professionals should aspire to be:

- **Cooperative** by sharing best practices, information, and ideas with colleagues, young professionals, and students.
- **Impartial** in our praise or criticism of others and their accomplishments, setting aside personal interests.
- **Inclusive** of all colleagues, and rejecting discrimination and harassment in any form.
- **Tolerant** of well-conducted research and well-reasoned results, which may differ from our own findings or opinions.
- **Truthful** in providing attribution when our work draws from the ideas of others.
- **Vigilant** by speaking out against actions that are damaging to the profession

PART-2

- **Descriptive Statistics:** Overview of Using Data: Definitions and Goals, Types of Data, Modifying Data in Excel, Creating Distributions from Data, Measures of Location, Measures of Variability, Analyzing Distributions, Measures of Association Between Two Variables.

Descriptive Statistics: **Overview of Data**

- The role of descriptive analytics is to collect and analyze data to gain a better understanding of variation and its impact on the business setting.
- **Data** are the facts and figures collected, analyzed, and summarized for presentation and interpretation.
- A characteristic or a quantity of interest that can take on different values is known as a **Variable.**
- In general, a quantity whose values are not known with certainty is called a **Random variable or uncertain variable.**
- An **observation** is a set of values corresponding to a set of variables.

TABLE 2.1 Data for Dow Jones Industrial Index Companies

Company	Symbol	Industry	Share Price ($)	Volume
Apple	AAPL	Technology	195.57	21,060,685
American Express	AXP	Financial	123.16	2,387,770
Boeing	BA	Manufacturing	369.32	3,002,708
Caterpillar	CAT	Manufacturing	133.71	3,747,782
Cisco Systems	CSCO	Technology	56.08	25,533,426
Chevron Corporation	CVX	Chemical, Oil, and Gas	123.64	4,705,879
Disney	DIS	Entertainment	139.94	14,670,995
Dow, Inc.	DOW	Chemical, Oil, and Gas	49.69	4,002,257
Goldman Sachs	GS	Financial	196.06	1,828,219
The Home Depot	HD	Retail	204.74	3,583,573
IBM	IBM	Technology	138.36	2,797,803
Intel	INTC	Technology	46.85	16,658,127
Johnson & Johnson	JNJ	Pharmaceuticals	144.24	7,516,973
JPMorgan Chase	JPM	Banking	107.76	18,654,861
Coca-Cola	KO	Food and Drink	51.76	11,517,843
McDonald's	MCD	Food and Drink	205.71	3,017,625
3M	MMM	Conglomerate	172.03	2,730,927
Merck	MRK	Pharmaceuticals	85.24	8,909,750
Microsoft	MSFT	Technology	133.43	33,328,420
Nike	NKE	Consumer Goods	82.62	7,335,836
Pfizer	PFE	Pharmaceuticals	43.76	26,952,088
Procter & Gamble	PG	Consumer Goods	111.72	6,795,912
Travelers	TRV	Insurance	153.13	1,295,768
UnitedHealth Group	UNH	Healthcare	247.66	3,178,942
United Technologies	UTX	Conglomerate	129.02	2,790,767
Visa	V	Financial	171.28	9,897,832
Verizon	VZ	Telecommunications	58.00	10,554,753
Walgreens Boots Alliance	WBA	Retail	52.95	8,535,442
Wal-Mart	WMT	Retail	110.72	6,104,935
ExxonMobil	XOM	Chemical, Oil, and Gas	76.27	9,722,688

- In this table variables are symbol, industry, share and price and volume.

  Each row in table corresponds to observation

Types of Data

- 1. Population and Sample Data
- 2. Quantitative and Categorical data
- 3. Cross-sectional and Time series Data

Types of Data

1. **Population and Sample Data**

- Data be categorized in several ways based on how they are collected and the type of data collected.
- It is not feasible to collect data from the population of all elements of interest.
- In such instances, we collect from a subset of the population known as a sample.

Types of Data

2. **Quantitative and Categorical data**

- Data are considered quantitative data if numeric and arithmetic operations, such as addition, subtraction, multiplication, and division, can be performed on them.
- If arithmetic operations cannot be performed on the data, they are considered categorical data. We can summarize categorical data by counting the number of observations or computing the proportions of observations in each category.

Types of Data

3. Cross-sectional and Time series Data

- For statistical analysis, it is important to distinguish between cross-sectional data and time series data. **Cross-sectional data** are collected from several entities at the same, or approximately the same, point in time. The data in Table 2.1 are cross-sectional because they describe the 30 companies that comprise the Dow at the same point in time (June 2019).
- **Time series data** are collected over several time periods. Graphs of time series data are frequently found in business and economic publications.

Sources of Data

FIGURE 2.2 Customer Opinion Questionnaire Used by Chops City Grill Restaurant

Chop's
CITY GRILL

Date: _____ Server Name: _____

Our customers are our top priority. Please take a moment to fill out our survey card, so we can better serve your needs. You may return this card to the front desk or return by mail. Thank you!

SERVICE SURVEY	Excellent	Good	Average	Fair	Poor
Overall Experience	❑	❑	❑	❑	❑
Greeting by Hostess	❑	❑	❑	❑	❑
Manager (Table Visit)	❑	❑	❑	❑	❑
Overall Service	❑	❑	❑	❑	❑
Professionalism	❑	❑	❑	❑	❑
Menu Knowledge	❑	❑	❑	❑	❑
Friendliness	❑	❑	❑	❑	❑
Wine Selection	❑	❑	❑	❑	❑
Menu Selection	❑	❑	❑	❑	❑
Food Quality	❑	❑	❑	❑	❑
Food Presentation	❑	❑	❑	❑	❑
Value for $ Spent	❑	❑	❑	❑	❑

What comments could you give us to improve our restaurant?

Thank you, we appreciate your comments. —The staff of Chops City Grill

Data necessary to analyze a business problem or opportunity can often be obtained with an appropriate study.

TABLE 2.1 Data for Dow Jones Industrial Index Companies

Company	Symbol	Industry	Share Price ($)	Volume
Apple	AAPL	Technology	195.57	21,060,685
American Express	AXP	Financial	123.16	2,387,770
Boeing	BA	Manufacturing	369.32	3,002,708
Caterpillar	CAT	Manufacturing	133.71	3,747,782
Cisco Systems	CSCO	Technology	56.08	25,533,426
Chevron Corporation	CVX	Chemical, Oil, and Gas	123.64	4,705,879
Disney	DIS	Entertainment	139.94	14,670,995
Dow, Inc.	DOW	Chemical, Oil, and Gas	49.69	4,002,257
Goldman Sachs	GS	Financial	196.06	1,828,219
The Home Depot	HD	Retail	204.74	3,583,573
IBM	IBM	Technology	138.36	2,797,803
Intel	INTC	Technology	46.85	16,658,127
Johnson & Johnson	JNJ	Pharmaceuticals	144.24	7,516,973
JPMorgan Chase	JPM	Banking	107.76	18,654,861
Coca-Cola	KO	Food and Drink	51.76	11,517,843
McDonald's	MCD	Food and Drink	205.71	3,017,625
3M	MMM	Conglomerate	172.03	2,730,927
Merck	MRK	Pharmaceuticals	85.24	8,909,750
Microsoft	MSFT	Technology	133.43	33,328,420
Nike	NKE	Consumer Goods	82.62	7,335,836
Pfizer	PFE	Pharmaceuticals	43.76	26,952,088
Procter & Gamble	PG	Consumer Goods	111.72	6,795,912
Travelers	TRV	Insurance	153.13	1,295,768
UnitedHealth Group	UNH	Healthcare	247.66	3,178,942
United Technologies	UTX	Conglomerate	129.02	2,790,767
Visa	V	Financial	171.28	9,897,832
Verizon	VZ	Telecommunications	58.00	10,554,753
Walgreens Boots Alliance	WBA	Retail	52.95	8,535,442
Wal-Mart	WMT	Retail	110.72	6,104,935
ExxonMobil	XOM	Chemical, Oil, and Gas	76.27	9,722,688

- In this table variables are symbol, industry, share and price and volume.

  Each row in table corresponds to observation

Modifying Data in Excel:- Sorting and filtering Data in excel

- Excel's Sort function, as shown in the following steps.
  - Step 1. Select cells A1:F21
  - Step 2. Click the Data tab in the Ribbon
  - Step 3. Click Sort in the Sort & Filter group
  - Step 4. Select the check box for My data has headers
  - Step 5. In the first Sort by dropdown menu, select Sales (February 2018)
  - Step 6. In the Order dropdown menu, select Largest to Smallest (see Figure 2.4)
  - Step 7. Click OK

FIGURE 2.3 Data for 20 Top-Selling Automobiles Entered into Excel with Percent Change in Sales from 2018

	A	B	C	D	E	F
1	**Rank (by February 2019 Sales)**	**Manufacturer**	**Model**	**Sales (February 2019)**	**Sales (February 2018)**	**Percent Change in Sales from 2018**
2	1	Toyota	Corolla	29016	25021	16.0%
3	2	Toyota	Camry	24267	30865	-21.4%
4	3	Honda	Civic	22979	25816	-11.0%
5	4	Honda	Accord	20254	19753	2.5%
6	5	Nissan	Sentra	17072	17148	-0.4%
7	6	Nissan	Altima	16216	19703	-17.7%
8	7	Ford	Fusion	13163	16721	-21.3%
9	8	Chevrolet Cruze	Malibu	10799	11890	-9.2%
10	9	Hyundai	Elantra	10304	15724	-34.5%
11	10	Kia	Soul	8592	6631	29.6%
12	11	Chevrolet	Cruze	7361	12875	-42.8%
13	12	Nissan	Versa	7410	7196	3.0%
14	13	Volkswagen	Jetta	7109	4592	54.8%
15	14	Kia	Optima	7212	6402	12.7%
16	15	Kia	Forte	6953	7662	-9.3%
17	16	Hyundai	Sonata	6481	6700	-3.3%
18	17	Tesla	Model 3	5750	2485	131.4%
19	18	Dodge	Charger	6547	7568	-13.5%
20	19	Ford	Mustang	5342	5800	-7.9%
21	20	Ford	Fiesta	5035	3559	41.5%

After sorting is performed

FIGURE 2.5 Top-Selling Automobiles Data Sorted by Sales in February 2018 Sales

	A	B	C	D	E	F
1	**Rank (by February 2019 Sales)**	**Manufacturer**	**Model**	**Sales (February 2019)**	**Sales (February 2018)**	**Percent Change in Sales from 2018**
2	2	Toyota	Camry	24267	30865	-21.38%
3	3	Honda	Civic	22979	25816	-10.99%
4	1	Toyota	Corolla	29016	25021	15.97%
5	4	Honda	Accord	20254	19753	2.54%
6	6	Nissan	Altima	16216	19703	-17.70%
7	5	Nissan	Sentra	17072	17148	-0.44%
8	7	Ford	Fusion	13163	16721	-21.28%
9	9	Hyundai	Elantra	10304	15724	-34.47%
10	11	Chevrolet	Cruze	7361	12875	-42.83%
11	8	Chevrolet Cruze	Malibu	10799	11890	-9.18%
12	15	Kia	Forte	6953	7662	-9.25%
13	18	Dodge	Charger	6547	7568	-13.49%
14	12	Nissan	Versa	7410	7196	2.97%
15	16	Hyundai	Sonata	6481	6700	-3.27%
16	10	Kia	Soul	8592	6631	29.57%
17	14	Kia	Optima	7212	6402	12.65%
18	19	Ford	Mustang	5342	5800	-7.90%
19	13	Volkswagen	Jetta	7109	4592	54.81%
20	20	Ford	Fiesta	5035	3559	41.47%
21	17	Tesla	Model 3	5750	2485	131.39%

Modifying Data in Excel:- Conditional Formatting of Data in Excel

- Conditional formatting in Excel can make it easy to identify data that satisfy certain conditions in a data set.
  - Step 1. Starting with the original data shown select cells F1:F21
  - Step 2. Click the Home tab in the Ribbon
  - Step 3. Click Conditional Formatting in the Styles group
  - Step 4. Select Highlight Cells Rules, and click Less Than . . . from the dropdown menu
  - Step 5. Enter 0% in the Format cells that are LESS THAN: box
  - Step 6. Click OK

FIGURE 2.7 Using Conditional Formatting in Excel to Highlight Automobiles with Declining Sales from February 2018

	A	B	C	D	E	F
1	**Rank (by February 2019 Sales)**	**Manufacturer**	**Model**	**Sales (February 2019)**	**Sales (February 2018)**	**Percent Change in Sales from 2018**
2	1	Toyota	Corolla	29016	25021	15.97%
3	2	Toyota	Camry	24267	30865	-21.38%
4	3	Honda	Civic	22979	25816	-10.99%
5	4	Honda	Accord	20254	19753	2.54%
6	5	Nissan	Sentra	17072	17148	-0.44%
7	6	Nissan	Altima	16216	19703	-17.70%
8	7	Ford	Fusion	13163	16721	-21.28%
9	8	Chevrolet Cruze	Malibu	10799	11890	-9.18%
10	9	Hyundai	Elantra	10304	15724	-34.47%
11	10	Kia	Soul	8592	6631	29.57%
12	12	Nissan	Versa	7410	7196	2.97%
13	11	Chevrolet	Cruze	7361	12875	-42.83%
14	14	Kia	Optima	7212	6402	12.65%
15	13	Volkswagen	Jetta	7109	4592	54.81%
16	15	Kia	Forte	6953	7662	-9.25%
17	18	Dodge	Charger	6547	7568	-13.49%
18	16	Hyundai	Sonata	6481	6700	-3.27%
19	17	Tesla	Model 3	5750	2485	131.39%
20	19	Ford	Mustang	5342	5800	-7.90%
21	20	Ford	Fiesta	5035	3559	41.47%

Data Bars from the Conditional Formatting dropdown menu in the Styles Group of the Home tab in the Ribbon.

FIGURE 2.8 Using Conditional Formatting in Excel to Generate Data Bars for the Top-Selling Automobiles Data

	A	B	C	D	E	F
1	Rank (by February 2019 Sales)	Manufacturer	Model	Sales (February 2019)	Sales (February 2018)	Percent Change in Sales from 2018
2	1	Toyota	Corolla	29016	25021	15.97%
3	2	Toyota	Camry	24267	30865	-21.38%
4	3	Honda	Civic	22979	25816	-10.99%
5	4	Honda	Accord	20254	19753	2.54%
6	5	Nissan	Sentra	17072	17148	-0.44%
7	6	Nissan	Altima	16216	19703	-17.70%
8	7	Ford	Fusion	13163	16721	-21.28%
9	8	Chevrolet Cruze	Malibu	10799	11890	-9.18%
10	9	Hyundai	Elantra	10304	15724	-34.47%
11	10	Kia	Soul	8592	6631	29.57%
12	12	Nissan	Versa	7410	7196	2.97%
13	11	Chevrolet	Cruze	7361	12875	-42.83%
14	14	Kia	Optima	7212	6402	12.65%
15	13	Volkswagen	Jetta	7109	4592	54.81%
16	15	Kia	Forte	6953	7662	-9.25%
17	18	Dodge	Charger	6547	7568	-13.49%
18	16	Hyundai	Sonata	6481	6700	-3.27%
19	17	Tesla	Model 3	5750	2485	131.39%
20	19	Ford	Mustang	5342	5800	-7.90%
21	20	Ford	Fiesta	5035	3559	41.47%

Creating Distributions from Data:

- Distributions help summarize many characteristics of a data set by describing how often certain values for a variable appear in that data set.
- Distributions can be created for both categorical and quantitative data, and they assist the analyst in gauging variation.

1. **Frequency Distributions for Categorical Data**
2. **Relative Frequency and Percent Frequency Distributions**
3. **Frequency Distributions for Quantitative Data**
4. **Histograms**
5. **Cumulative Distributions**

Creating Distributions from Data:

1. Frequency Distributions for Categorical Data

- It is often useful to create a frequency distribution for a data set.
- A **frequency distribution** is a summary of data that shows the number (frequency) of observations in each of several nonoverlapping classes, typically referred to as **bins**.

TABLE 2.3 Data from a Sample of 50 Soft Drink Purchases

Coca-Cola	Sprite	Pepsi
Diet Coke	Coca-Cola	Coca-Cola
Pepsi	Diet Coke	Coca-Cola
Diet Coke	Coca-Cola	Coca-Cola
Coca-Cola	Diet Coke	Pepsi
Coca-Cola	Coca-Cola	Dr. Pepper
Dr. Pepper	Sprite	Coca-Cola
Diet Coke	Pepsi	Diet Coke
Pepsi	Coca-Cola	Pepsi
Pepsi	Coca-Cola	Pepsi
Coca-Cola	Coca-Cola	Pepsi
Dr. Pepper	Pepsi	Pepsi
Sprite	Coca-Cola	Coca-Cola
Coca-Cola	Sprite	Dr. Pepper
Diet Coke	Dr. Pepper	Pepsi
Coca-Cola	Pepsi	Sprite
Coca-Cola	Diet Coke	

TABLE 2.4 Frequency Distribution of Soft Drink Purchases

Soft Drink	Frequency
Coca-Cola	19
Diet Coke	8
Dr. Pepper	5
Pepsi	13
Sprite	5
Total	50

Creating Distributions from Data:

2. Relative Frequency and Percent Frequency Distributions

- The *relative frequency* of a bin equals the fraction or proportion of items belonging to a class. For a data set with *n* observations, the relative frequency of each bin can be determined as follows:

$$\text{Relative frequency of a bin} = \frac{\text{Frequency of the bin}}{n}$$

- A **relative frequency distribution** is a tabular summary of data showing the relative frequency for each bin.
- A **percent frequency distribution** summarizes the percent frequency of the data for each bin.

TABLE 2.5 Relative Frequency and Percent Frequency Distributions of Soft Drink Purchases

Soft Drink	Relative Frequency	Percent Frequency (%)
Coca-Cola	0.38	38
Diet Coke	0.16	16
Dr. Pepper	0.10	10
Pepsi	0.26	26
Sprite	0.10	10
Total	1.00	100

3. Frequency Distributions for Quantitative Data

TABLE 2.6	Year-End Audit Times (Days)		
12	14	19	18
15	15	18	17
20	27	22	23
22	21	33	28
14	18	16	13

Table 2.6. These data show the time in days required to complete year-end audits for a sample of 20 clients of Sanderson and Clifford, a small public accounting firm.

The three steps necessary to define the classes for a frequency distribution with quantitative data are as follows:

1.Determine the number of nonoverlapping bins.

2. Determine the width of each bin.

3. Determine the bin limits.

1. Number of Bins Bins are formed by specifying the ranges used to group the data. As a general guideline, we recommend using from 5 to 20 bins.(5)

2. Width of the bins

APPROXIMATE BIN WIDTH

$$\frac{\text{Largest data value} - \text{smallest data value}}{\text{Number of bins}}$$

Approximate bin width of (33 -12)/5 =4.25 We therefore decided to round up and use a bin width of five days in the frequency distribution.

3. Frequency Distributions for Quantitative Data

3. Bin Limits must be chosen so that each data item belongs to one and only one class. The lower bin limit identifies the smallest possible data value assigned to the bin. (10) The upper bin limit identifies the largest possible data value assigned to the class. (34).

TABLE 2.7 Frequency, Relative Frequency, and Percent Frequency Distributions for the Audit Time Data

Audit Times (days)	Frequency	Relative Frequency	Percent Frequency
10–14	4	0.20	20
15–19	8	0.40	40
20–24	5	0.25	25
25–29	2	0.10	10
30–34	1	0.05	5

- Download the audit data in excel.
- We can use the FREQUENCY function in Excel to count the number of observations in each bin.

  Step 1. Select cells D2:D6

  Step 2. Type the formula FREQUENCY(A2:A21, C2:C6). The range A2:A21 defines the data set, and the range C2:C6 defines the bins.

  Step 3. Press CTRL+SHIFT+ENTER after typing the formula in Step 2

FIGURE 2.11 Using Excel to Generate a Frequency Distribution for Audit Times Data

	A	B	C	D
1	**Audit Times (in Days)**		**Bin**	**Frequency**
2	12		14	=FREQUENCY(A2:A21,C2:C6)
3	15		19	=FREQUENCY(A2:A21,C2:C6)
4	20		24	=FREQUENCY(A2:A21,C2:C6)
5	22		29	=FREQUENCY(A2:A21,C2:C6)
6	14		34	=FREQUENCY(A2:A21,C2:C6)
7	14			
8	15			
9	27			
10	21			
11	18			
12	19			
13	18			
14	22			
15	33			
16	16			
17	18			
18	17			
19	23			
20	28			
21	13			

	A	B	C	D
1	**Audit Times (in Days)**		**Bin**	**Frequency**
2	12		14	4
3	15		19	8
4	20		24	5
5	22		29	2
6	14		34	1
7	14			
8	15			
9	27			
10	21			
11	18			
12	19			
13	18			
14	22			
15	33			
16	16			
17	18			
18	17			
19	23			
20	28			
21	13			

4. Histograms

A common graphical presentation of quantitative data is a **histogram**. This graphical summary can be prepared for data previously summarized in either a frequency, a relative frequency, or a percent frequency distribution.

Histograms can be created in Excel using the Data Analysis Tool Pak.

Step 1. Click the Data tab in the Ribbon

Step 2. Click Data Analysis in the Analyze group

Step 3. When the Data Analysis dialog box opens, choose Histogram from the list of

- Analysis Tools, and click OK
- In the Input Range: box, enter A2:A21
- In the Bin Range: box, enter C2:C6
- Under Output Options:, select New Worksheet Ply:
- Select the check box for Chart Output
- Click OK

4. Histograms

To remove the gaps between the columns in the histogram created by Excel, follow these steps:

Step 1. Right-click on one of the columns in the histogram

Select Format Data Series…

Step 2. When the Format Data Series pane opens, click the Series Options button, Set the Gap Width to 0%

One of the most important uses of a histogram is to provide information about the shape, or form, of a distribution. Skewness, or the lack of symmetry, is an important characteristic of the shape of a distribution.

5. Cumulative Distributions

A variation of the frequency distribution that provides another tabular summary of quantitative data is the **cumulative frequency distribution.**

TABLE 2.8 Cumulative Frequency, Cumulative Relative Frequency, and Cumulative Percent Frequency Distributions for the Audit Time Data

Audit Time (days)	Cumulative Frequency	Cumulative Relative Frequency	Cumulative Percent Frequency
Less than or equal to 14	4	0.20	20
Less than or equal to 19	12	0.60	60
Less than or equal to 24	17	0.85	85
Less than or equal to 29	19	0.95	95
Less than or equal to 34	20	1.00	100

Measures of Location

- Mean(Arithmetic Mean)
- Median
- Mode
- Geometric mean

TABLE 2.9 Data on Home Sales in a Cincinnati, Ohio, Suburb

Home Sale	Selling Price ($)
1	138,000
2	254,000
3	186,000
4	257,500
5	108,000
6	254,000
7	138,000
8	298,000
9	199,500
10	208,000
11	142,000
12	456,250

Mean (Arithmetic Mean)

The most commonly used measure of location is the **mean (arithmetic mean)**, or average value, for a variable. The mean provides a measure of central location for the data. If the data are for a sample (typically the case), the mean is denoted by $\bar{x}$. The sample mean is a point estimate of the (typically unknown) population mean for the variable of interest. If the data for the entire population are available, the population mean is computed in the same manner, but denoted by the Greek letter μ.

In statistical formulas, it is customary to denote the value of variable x for the first observation by x_1, the value of variable x for the second observation by x_2, and so on. In general, the value of variable x for the ith observation is denoted by x_i. For a sample with n observations, the formula for the sample mean is as follows.

SAMPLE MEAN

$$\bar{x} = \frac{\sum x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n} \tag{2.2}$$

Median

The **median**, another measure of central location, is the value in the middle when the data are arranged in ascending order (smallest to largest value). With an odd number of observations, the median is the middle value. An even number of observations has no single middle value. In this case, we follow convention and define the median as the average of the values for the middle two observations.

Let us apply this definition to compute the median class size for a sample of five college classes. Arranging the data in ascending order provides the following list:

32 42 46 46 54

Because $n = 5$ is odd, the median is the middle value. Thus, the median class size is 46 students. Even though this data set contains two observations with values of 46, each observation is treated separately when we arrange the data in ascending order.

- Suppose we also compute the median value for the 12 home sales in Table 2.9. We first arrange the data in ascending order.

108,000 138,000 138,000 142,000 186,000 199,500 208,000 254,000 254,000 257,500 298,000 456,250

Middle Two Values

Because n 5 is even, the median is the average of the middle two values: 199,500 and 208,000.

$$\text{Median} = \frac{199{,}500 + 208{,}000}{2} = 203{,}750$$

The median of a data set can be found in Excel using the function MEDIAN. In Figure 2.16, the value for the median in cell E3 is found using the formula MEDIAN(B2:B13)

Mode

- A third measure of location, the mode, is the value that occurs most frequently in a data set.
- The Excel MODE.SNGL function will return only a single most-often-occurring value.
- To find both of the modes in Excel, we take these steps:
- Step 1. Select cells E4 and E5
- Step 2. Type the formula 5MODE.MULT(B2:B13)
- Step 3. Press CTRL+SHIFT+ENTER after typing the formula in Step 2.

Geometric Mean

The **geometric mean** is a measure of location that is calculated by finding the nth root of the product of n values. The general formula for the sample geometric mean, denoted $\bar{x}_g$, follows.

SAMPLE GEOMETRIC MEAN

$$\bar{x}_g = \sqrt[n]{(x_1)(x_2)\cdots(x_n)} = [(x_1)(x_2)\cdots(x_n)]^{1/n} \qquad (2.3)$$

- The geometric mean is often used in analyzing growth rates in financial data. In these
- types of situations, the arithmetic mean or average value will provide misleading results.

FIGURE 2.16 Calculating the Mean, Median, and Modes for the Home Sales Data Using Excel

	A	B	C	D	E
1	**Home Sale**	**Selling Price ($)**			
2	1	138,000		**Mean:**	=AVERAGE(B2:B13)
3	2	254,000		**Median:**	=MEDIAN(B2:B13)
4	3	186,000		**Mode 1:**	=MODE.MULT(B2:B13)
5	4	257,500		**Mode 2:**	=MODE.MULT(B2:B13)
6	5	108,000			
7	6	254,000			
8	7	138,000			
9	8	298,000			
10	9	199,500			
11	10	208,000			
12	11	142,000			
13	12	456,250			

	A	B	C	D	E
1	**Home Sale**	**Selling Price ($)**			
2	1	138,000		**Mean:**	$ 219,937.50
3	2	254,000		**Median:**	$ 203,750.00
4	3	186,000		**Mode 1:**	$ 138,000.00
5	4	257,500		**Mode 2:**	$ 254,000.00
6	5	108,000			
7	6	254,000			
8	7	138,000			
9	8	298,000			
10	9	199,500			
11	10	208,000			
12	11	142,000			
13	12	456,250			

Measures of Variability

- It is often desirable to consider measures of variability or dispersion.
- **Range:** The simplest measure of variability is the **range**. The range can be found by subtracting the smallest value from the largest value in a data set. The range can be calculated in Excel using the MAX and MIN functions
- **Variance:** The **variance** is a measure of variability that utilizes all the data. The variance is based on the *deviation of the mean*, which is the difference between the value of each observation (*xi*) and the mean. The variance in cell E8 is calculated using the formula VAR.S(B2:B13)

SAMPLE VARIANCE

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1}$$

Measures of Variability

- **Standard Deviation:** The **standard deviation** is the positive square root of the variance.

SAMPLE STANDARD DEVIATION

$$s = \sqrt{s^2}$$

- **Coefficient of Variation:** In some situations, we may be interested in a descriptive statistic that indicates how large the standard deviation is relative to the mean. This measure is called the **coefficient of variation** and is usually expressed as a percentage. Excel calculation for the sample standard deviation of the home
- sales data, which can be calculated using Excel's STDEV.S function. The sample standard deviation in cell E9 is calculated using the formula STDEV.S(B2:B13).

COEFFICIENT OF VARIATION

$$\left(\frac{\text{Standard deviation}}{\text{Mean}} \times 100\right)\%$$

FIGURE 2.19 Calculating Variability Measures for the Home Sales Data in Excel

	A	B	C	D	E
1	Home Sale	Selling Price ($)			
2	1	138000		Mean:	=AVERAGE(B2:B13)
3	2	254000		Median:	=MEDIAN(B2:B13)
4	3	186000		Mode 1:	=MODE.MULT(B2:B13)
5	4	257500		Mode 2:	=MODE.MULT(B2:B13)
6	5	108000			
7	6	254000		Range:	=MAX(B2:B13)-MIN(B2:B13)
8	7	138000		Variance:	=VAR.S(B2:B13)
9	8	298000		Standard Deviation:	=STDEV.S(B2:B13)
10	9	199500			
11	10	208000		Coefficient of Variation:	=E9/E2
12	11	142000			
13	12	456250		85th Percentile:	=PERCENTILE.EXC(B2:B13,0.85)
14					
15				1st Quartile:	=QUARTILE.EXC(B2:B13,1)
16				2nd Quartile:	=QUARTILE.EXC(B2:B13,2)
17				3rd Quartile:	=QUARTILE.EXC(B2:B13,3)
18					
19				IQR:	=E17-E15

	A	B	C	D	E
1	Home Sale	Selling Price ($)			
2	1	138000		Mean:	$ 219,937.50
3	2	254000		Median:	$ 203,750.00
4	3	186000		Mode 1:	$ 138,000.00
5	4	257500		Mode 2:	$ 254,000.00
6	5	108000			
7	6	254000		Range:	$ 348,250.00
8	7	138000		Variance:	9037501420
9	8	298000		Standard Deviation:	$ 95,065.77
10	9	199500			
11	10	208000		Coefficient of Variation:	43.22%
12	11	142000			
13	12	456250		85th Percentile:	$ 305,912.50
14					
15				1st Quartile:	$ 139,000.00
16				2nd Quartile:	$ 203,750.00
17				3rd Quartile:	$ 256,625.00
18					
19				IQR:	$ 117,625.00

Analysing Distributions:

- Distributions are beneficial for interpreting and analyzing data. A distribution describes the overall variability of the observed values of a variable.
- **Percentiles** : A **percentile** is the value of a variable at which a specified (approximate) percentage of observations are below that value.

LOCATION OF THE *p*th PERCENTILE

$$L_p = \frac{p}{100}(n+1)$$

The pth percentile can also be calculated in Excel using the function PERCENTILE.EXC

- **Quartiles:** It is often desirable to divide data into four parts, with each part containing approximately. one-fourth, or 25 percent, of the observations. These division points are referred to as the **quartiles** and are defined as follows:
- *Q*1 first quartile, or 25th percentile
- *Q*2 second quartile, or 50th percentile (also the median)
- *Q*3 third quartile, or 75th percentile

The difference between the third and first quartiles is often referred to as the **interquartile range.**

A quartile can be computed in Excel using the function QUARTILE.EXC.

- ***z*-Scores :** A ***z*-score** allows us to measure the relative location of a value in the data set. More specifically,

a *z*-score helps us determine how far a particular value is from the mean relative to the data set's standard deviation. The z-score can be calculated in Excel using the function STANDARDIZE.

z-SCORE

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

where

z_i = the z-score for x_i
$\bar{x}$ = the sample mean
s = the sample standard deviation

FIGURE 2.20 Calculating *z*-Scores for the Home Sales Data in Excel

	A	B	C
1	**Home Sale**	**Selling Price ($)**	**z-Score**
2	1	138000	=STANDARDIZE(B2,B15,B16)
3	2	254000	=STANDARDIZE(B3,B15,B16)
4	3	186000	=STANDARDIZE(B4,B15,B16)
5	4	257500	=STANDARDIZE(B5,B15,B16)
6	5	108000	=STANDARDIZE(B6,B15,B16)
7	6	254000	=STANDARDIZE(B7,B15,B16)
8	7	138000	=STANDARDIZE(B8,B15,B16)
9	8	298000	=STANDARDIZE(B9,B15,B16)
10	9	199500	=STANDARDIZE(B10,B15,B16)
11	10	208000	=STANDARDIZE(B11,B15,B16)
12	11	142000	=STANDARDIZE(B12,B15,B16)
13	12	456250	=STANDARDIZE(B13,B15,B16)
14			
15	**Mean:**	=AVERAGE(B2:B13)	
16	**Standard Deviation:**	=STDEV.S(B2:B13)	

	A	B	C
1	**Home Sale**	**Selling Price ($)**	**z-Score**
2	1	138,000	–0.862
3	2	254,000	0.358
4	3	186,000	–0.357
5	4	257,500	0.395
6	5	108,000	–1.177
7	6	254,000	0.358
8	7	138,000	–0.862
9	8	298,000	0.821
10	9	199,500	–0.215
11	10	208,000	–0.126
12	11	142,000	–0.820
13	12	456,250	2.486
14			
15	**Mean:**	$ 219,937.50	
16	**Standard Deviation:**	$ 95,065.77	

- **Empirical Rule :** The **empirical rule** can be used to determine the percentage of data values that are within a specified number of standard deviations of the mean.
- **Identifying Outliers :** Sometimes a data set will have one or more observations with unusually large or unusually small values. These extreme values are called **outliers**.
- **Boxplots :** A **boxplot** is a graphical summary of the distribution of data.

The step-by-step directions below illustrate how to create boxplots in Excel for both a single variable and multiple variables. First we will create a boxplot for a single variable using the *HomeSales* file.

Step 1. Select cells B1:B13
Step 2. Click the **Insert** tab on the Ribbon
Click the **Insert Statistic Chart** button in the **Charts** group
Choose the **Box and Whisker** chart from the drop-down menu

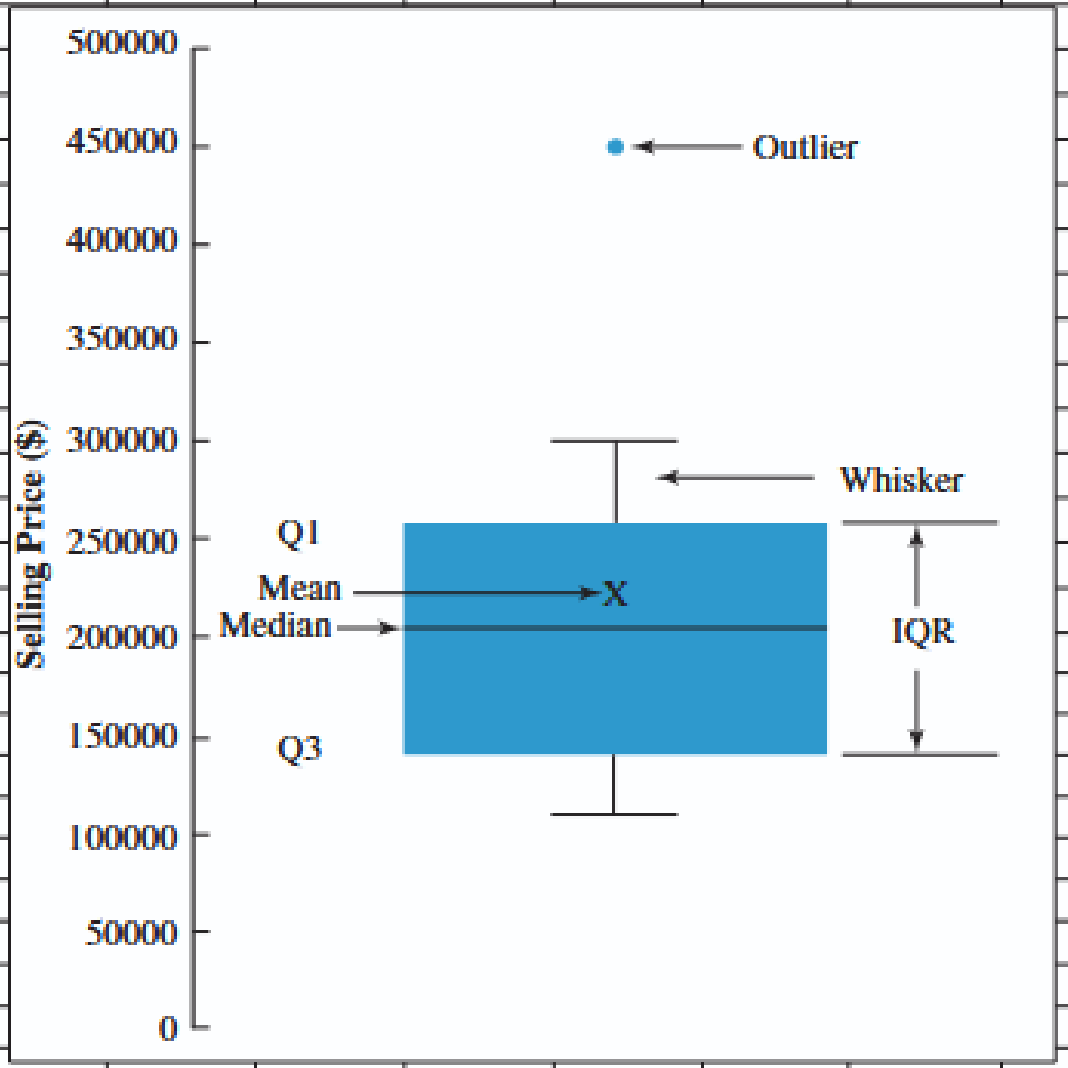
The resulting boxplot created in Excel is shown in Figure 2.24. Comparing this figure to Figure 2.22, we see that all the important elements of a boxplot are generated here. Excel orients the boxplot vertically, and by default it also includes a marker for the mean.

Next we will use the *HomeSalesComparison* file to create boxplots in Excel for multiple variables similar to what is shown in Figure 2.26.

Step 1. Select cells B1:F11
Step 2. Click the **Insert** tab on the Ribbon
Click the **Insert Statistic Chart** button in the **Charts** group
Choose the **Box and Whisker** chart from the drop-down menu

FIGURE 2.24 Boxplot Created in Excel for Home Sales Data

	A	B	C	D	E	F	G	H	I	J
1	Home Sale	Selling Price ($)								
2	1	138000								
3	2	254000								
4	3	186000								
5	4	257500								
6	5	108000								
7	6	254000								
8	7	138000								
9	8	298000								
10	9	199500								
11	10	208000								
12	11	142000								
13	12	456250								
14										
15										
16										
17										
18										
19										
20										
21										
22										
23										
24										
25										
26										



Measures of Association Between Two Variables:

- **Scatter Charts :** It is a useful graph for analyzing the relationship between two variables.
- **Covariance :** It is a descriptive measure of the linear association between two variables. For a sample of size *n* with the observations (*x*1 , *y*1), (*x*2 , *y*2), and so on, the sample covariance is defined as follows:

SAMPLE COVARIANCE

$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

The covariance is calculated in cell B17 using the formula COVARIANCE.S(A2:A15, B2:B15).

Note: If the covariance is near 0, then the *x* and *y* variables are not linearly related. If the covariance is less than 0, then the *x* and *y* variables are negatively related, which means that as *x* increases, *y* generally decreases.

- **Correlation Coefficient:** The **correlation coefficient** measures the relationship between two variables, and, unlike covariance, the relationship between two variables is not affected by the units of measurement for *x* and *y*. For sample data, the correlation coefficient is defined as follows:

SAMPLE CORRELATION COEFFICIENT

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

where

r_{xy} = sample correlation coefficient
s_{xy} = sample covariance
s_x = sample standard deviation of x
s_y = sample standard deviation of y

The correlation coefficient is computed using the formula CORREL(A2:A15, B2:B15), where A2:A15 defines the range for the x variable and B2:B15 defines the range for the y variable.

Modifying data in excel

- Sort, Filter, conditional Formatting, Data bars

Creating Distributions from data

- Frequency(range),Histogram,

Measures of location

- Mean =AVERAGE(range).
- Median =MEDIAN(range)
- Mode =MODE.MULT(range)
- Geometric mean =GEOMEAN(range)

Measures of Variability

- Range =MAX(range) − MIN(range).
- Variance =VAR.S(range)
- Standard variation =STDEV.S(range)
- Coefficient of variance (Standard deviation/mean)*100

Analysing Distributions

- Percentile =PERCENTILE.EXC(range,percentilevalue)
- Quartile = QUARTILE.EXC(range,quartile number)
- Z-score = STANDARDIZE(range),
- Boxplot

Measures of association between two variables

- Scatter charts
- Covariance =COVARIANCE.S(range,range).
- correlation coefficient = CORREL(range, range)