BUAN 3500: Data Visualization and Descriptive Analytics

Descriptive Statistics

Lauren M. Nelsen, Ph.D.

University of Colorado Colorado Springs

Descriptive Statistics

- Much of this chapter is a review of things you learned in QUAN 2010.
- We'll look at some of the topics from the viewpoint of data analysis and practice some Excel techniques.

Section 2.1: Overview of Using Data: Definitions and Goals

Read this section on your own!

Section 2.2: Types of Data

- **Population vs Sample Data**
  - key concept in statistics and analytics

- **Quantitative vs Categorical Data**

  Examples:
  - marital status ← categorical
  - eye color ← categorical
  - height ← quantitative
  - square footage ← quantitative
  - smoking status ← categorical

This is an important distinction because we can't perform the same operations on quantitative and categorical data. So we need to be clear on what kind of data we are analyzing!

Section 2.2: Types of Data

- **Cross-Sectional vs Time Series Data**
  - **Cross-sectional** ← time of collection is not a variable relevant to the analysis
    - **Example:** current GPAs of everyone in our class
  - **Time-series** ← gathered over a period of time, and the time is relevant
    - **Example:** average student GPA over each of the past 10 semesters

Section 2.2: Types of Data

- **Sources of Data: experimental vs observational**
  - **Experimental** ← some variable(s) controlled or manipulated
    - **Example:** Online learning – provide some students with fast internet and others with slow internet to see if speed is a factor in online class performance
  - **Observational** ← no control over variables – just record what is observed
    - **Example:** Record type of internet and online class performance

Section 2.3: Modifying Data in Excel

Note: This is likely review for you, but it is useful!

Sorting and filtering data
→ These processes can help us to understand and identify patterns in our data.

Example

Open the data file “top20cars2019” in Canvas.

- The data is currently sorted by Feb 2019 sales. What if we want to sort by Feb 2018 sales?
- What if we want to see only Nissan data? (Filter.)

Section 2.3: Modifying Data in Excel

Conditional formatting → Gives us a way to highlight cells with certain properties

Example

Continued: Look at the file “Top20cars2019” again.

- Highlight Feb 2019 sales that are between 10,000 and 20,000.

Section 2.3: Modifying Data in Excel

Note: Our book shows the steps for this example very clearly so you can always look back at those!

Section 2.4: Creating Distributions from Data

Frequency Distributions for Categorical Data:

Definition

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**.

Section 2.4: Creating Distributions from Data

Example

Open the file “softdrinks” in Excel.

- If we make a frequency distribution, what should the bins be?
- Create a frequency distribution in Excel.

Section 2.4: Creating Distributions from Data

Relative Frequency and Percent Frequency Distributions:

Definition

- **relative frequency of a bin:** fraction or proportion of items belonging to a class

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

(for a data set with n observations)

Section 2.4: Creating Distributions from Data

Example:

Soft drink	**Frequency**
Coca-Cola	19
Diet Coke	8
Dr. Pepper	5
Pepsi	13
Sprite	5

Relative frequency for Sprite:

$$\frac{5}{50} = 10\%$$

Section 2.4: Creating Distributions from Data

Example

Continued:

Open the file “softdrinks” again in Excel.

- Construct a percent frequency distribution in Excel.

Frequency Distributions for Quantitative Data
We can also create frequency distributions for quantitative data, but we need to be more careful to be sure we're dividing the data into non-overlapping bins.

Example

Open the file “AuditTime”. (This shows the number of days to complete audits for 20 different customers.)

We would like to construct a frequency distribution for this data.

Section 2.4: Creating Distributions from Data

Steps necessary for defining classes for a frequency distribution with quantitative data:

1. Determine the number of non-overlapping bins.
2. Determine the width of each bin.
3. Determine the bin limits.

Section 2.4: Creating Distributions from Data

1. **Step 1:** Determine the number of non-overlapping bins.
    - Typically 5 to 20, depending on the number of data items
    - Our file is small (20 items), so choose 5 bins.

Section 2.4: Creating Distributions from Data

1. **Step 2:** Determine the width of each bin.
   - Find the range of the data
   - Divide the range by the number of bins.
   - Adjust if necessary

Approximate Bin Width:

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

In our example:

$$\frac{33-12}{5} = 4.2$$

Let's round to a bin width of 5.
(We could instead use 6 bins with a width of 4)

1. **Step 3:** Determine the bin limits (boundaries).
    - 5 bins with a width of 5 gives us a range from top to bottom of 25.
    - If (for convenience) we choose a bottom value of 10, our top value will be 35, and our largest data value will fit.
    - The bins must not overlap.
        - Since our data is integers, we can have our first bin run from 10 to 14 (5 values), the next from 15 to 19, and so on.

With the bins established, we can create the frequency distribution by counting the number of items in each bin with the FREQUENCY formula. (Do this in Excel.)

Histograms:

Example

Open the file “AuditTime” again.

Use the frequency distribution we found above to construct a histogram for this data.

Then determine if the distribution is skewed left, skewed right, or symmetric.

Cumulative Distributions:

Example

Open the file “AuditTime” again.

Use the frequency distribution we found above to construct the cumulative frequency distribution for this data.

Section 2.5: Measures of Location

Sections 2.5 and 2.6 also cover topics that you covered in QUAN 2010. You should review these sections on your own, but if we have time we will briefly review them in class.

Definition

- The **mean** (or **average value**) for a variable is given by

$$\overline{x} = \frac{\sum x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

- The **median** is the value in the middle when the data are arranged in ascending order.
- The **mode** is the value that occurs most frequently in a data set.

Section 2.5: Measures of Location

Example

Open the file "HomeSales" in Excel. This file shows a sample of home selling prices for 12 homes in a suburb of Cincinnati, Ohio.

Use Excel to find the mean (average), median, and mode(s) for the home sale values.

Section 2.5: Measures of Location

Definition

- **Sample Geometric Mean:**

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

  - This is often used in analyzing growth rates in financial data.

Example

Open the file "MutualFundReturns" in Excel. Assume we invest \$100 in the fund.

- What would the balance in the fund at the end of year 1 be?
  \$77.90
- What about the balance at the end of year 2?
  $100 \cdot (.779) = \$100.26$
- What would the balance be at the end of 10 years?
  $100 \cdot (.779)(1.287)\cdots(1.021) = 100 \cdot (1.3345) = \$133.45$

Section 2.5: Measures of Location

Geometric mean of the 10 growth factors:

$$\sqrt[10]{1.3345} = 1.029$$

So: annual returns grew at an average annual rate of $(1.029 - 1) \cdot 100$, or 2.9%.

So: a \$100 investment would grow to $\$100 \cdot (1.029)^{10}$ at the end of 10 years.

In Excel: GEOMEAN(C2:C11)

Section 2.6: Measures of Variability

Definition

- **range:**

$$\boxed{\text{range} = \text{largest value} - \text{smallest value}}$$

- **variance:** based on deviation about the mean
  For a random sample, the sample variance, s^2, is given by:

$$\boxed{s^2 = \frac{\sum (x_i - \overline{x})^2}{n-1}}$$

(For the population variance, σ^2, we divide by n instead of $n-1$.)

Section 2.6: Measures of Variability

Definition

- **standard deviation:** positive square root of the variance

$$\boxed{\text{sample standard deviation: } s = \sqrt{s^2}}$$

$$\boxed{\text{population standard deviation: } \sigma = \sqrt{\sigma^2}}$$

- **coefficient of variation:** measure that indicates how large the standard deviation is relative to the mean

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

Section 2.6: Measures of Variability

Example

Open the “HomeSales” file again in Excel. Calculate the following quantities:

- Range of home sale prices
  $\$348,250$
- The sample variance, s^2 for the 12 given home sale prices
  9037501420
- The sample standard deviation, s, for the 12 given home sale prices
  $\$95,065.77$
- The coefficient of variation for the home sales data
  43.22%

Tufte

Before we launch into Section 2.7 in our textbook, let's pause to talk about the book that you should be reading from this semester:

The Visual Display of Quantitative Information by Edward R. Tufte

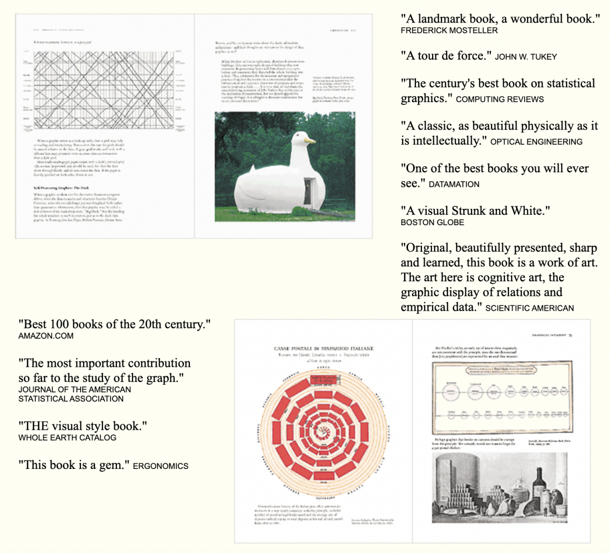
We'll talk more about your readings in future class meetings, but for now I want to introduce you to the author of the book.

Tufte

In 2013, the Financial Times said

> "Edward Tufte is the guru of graphics, the high priest of presentation."

Tufte – The Visual Display of Quantitative Information



"A landmark book, a wonderful book."
FREDERICK MOSTELLER

"A tour de force." JOHN W. TUKEY

"The century's best book on statistical graphics." COMPUTING REVIEWS

"A classic, as beautiful physically as it is intellectually." OPTICAL ENGINEERING

"One of the best books you will ever see." DATAMATION

"A visual Strunk and White."
BOSTON GLOBE

"Original, beautifully presented, sharp and learned, this book is a work of art. The art here is cognitive art, the graphic display of relations and empirical data." SCIENTIFIC AMERICAN

"Best 100 books of the 20th century."
AMAZON.COM

"The most important contribution so far to the study of the graph."
JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION

"THE visual style book."
WHOLE EARTH CATALOG

"This book is a gem." ERGONOMICS

(Image from edwardtufte.com)

Tufte – The Visual Display of Quantitative Information

Even though this book was published several decades ago, the principles it discusses are still important to consider now!

Tufte

Let’s watch the first two minutes of this PBS Digital Studios video that features Dr. Tufte:

https://youtu.be/AdSZJzb-aX8

Tufte – The Visual Display of Quantitative Information

If you haven't already started the assigned readings, do that this week!

Section 2.7: Analyzing Distributions

Definition

- If $n\%$ of the items in a distribution are less than a particular data item, we say that the data item is in the n**th percentile** of the distribution. **In Excel:** PERCENTILE.EXC
- The **percentile rank** identifies the percentile of a particular value within a data set. **In Excel:** PERCENTRANK.EXC

Section 2.7: Analyzing Distributions

Example

Open the “HomeSales” file again in Excel.

- Sort the sale prices in ascending order.
- Find the sale price that would be the 50th percentile using Excel. PERCENTILE.EXC(B2:B13,0.5)
- Find what percentile the sale price of $298,000 would be. PERCENTRANK.EXC(B2:B13,298000)

Location of the pth Percentile:

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

Example

Open the "HomeSales" file again in Excel.

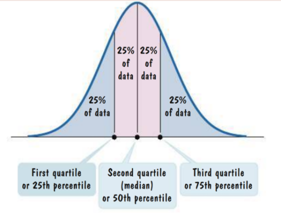
- Find the location of the 85th percentile, L_{85} using this formula. Does this make sense with your previous answer(s)?

$$L_{85} = \frac{85}{100}(12+1) = 11.05$$

Section 2.7: Analyzing Distributions

Definition

Quartiles are commonly encountered percentiles. Quartiles divide data sets into four equal parts.


25% of data
25% of data
25% of data
25% of data
First quartile or 25th percentile
Second quartile (median) or 50th percentile
Third quartile or 75th percentile

(Image from the book Thinking Mathematically by Blitzer)

The first, second, and third quartiles in a data set are the values for the 25th, 50th, and 75th percentiles, respectively. (These are denoted by Q_1, Q_2, and Q_3.)

In Excel: QUARTILE.EXC

Section 2.7: Analyzing Distributions

Example

Open the “HomeSales” file again. Find Q_1, Q_2, and Q_3 for home sales prices.

- $Q_1 = 139,000$
- $Q_2 = 203,750$
- $Q_3 = 256,625$

Reminder: Interquartile range (IQR) is defined as

$$\text{IQR} = Q_3 - Q_1$$

Section 2.7: Analyzing Distributions

z-**Scores:** You also learned about *z*-scores in QUAN 2010.

Definition

The *z*-**score** identifies the number of standard deviations a particular value is from the mean of its distribution.

$$\boxed{z_i = \frac{x_i - \overline{x}}{s}}$$

where

$$\begin{aligned} z_i &= \text{ the } z\text{-score for } x_i \\ \overline{x} &= \text{ the sample mean} \\ s &= \text{ the sample standard deviation} \end{aligned}$$

Section 2.7: Analyzing Distributions

Example

What does a z-score (or z-value) or -1.5 mean?

→ that x is 1.5 standard deviations below the mean of the data set

You can calculate z-values in Excel using the STANDARDIZE formula.

(The z-score is sometimes the **standardized value**.)

Example

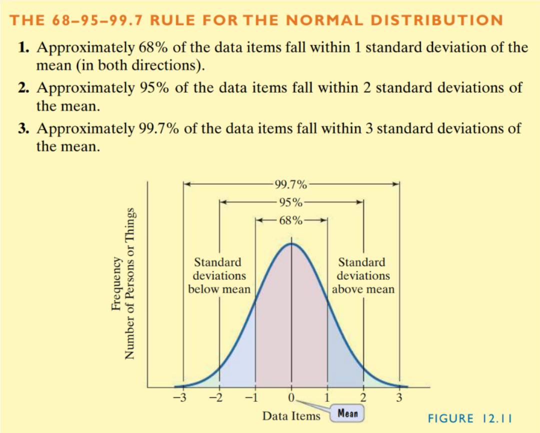
Find the z-values for the home sales prices in the Excel file.

Section 2.7: Analyzing Distributions

The Empirical Rule says that if a distribution follows a bell-shaped, symmetric curve centered around the mean, we should expect:

- approximately 68% of the data to fall within one standard deviation of the mean,
- approximately 95% of the data to flal within two standard deviations of the mean, and
- approximately 99.7% of the data to fall within three standard deviations of the mean.

The Empirical Rule:



THE 68–95–99.7 RULE FOR THE NORMAL DISTRIBUTION

1. Approximately 68% of the data items fall within 1 standard deviation of the mean (in both directions).
2. Approximately 95% of the data items fall within 2 standard deviations of the mean.
3. Approximately 99.7% of the data items fall within 3 standard deviations of the mean.

FIGURE 12.11

(Image from the book Thinking Mathematically by Blitzer)

Section 2.7: Analyzing Distributions

Identifying outliers:

- This is an important part of understanding the data!
- Each outlier should be investigated during analysis
  - could be a typo
  - could have been gathered improperly
  - could be real
- What is an outlier?
  - Rule of thumb: any data point with a *z*-value < -3 or $> +3$ is probably an outlier, since nearly all data should fall within 3 standard deviations of the mean (using the Empirical Rule)
    - OR: any data points outside of the interval

$$[Q_1 - (1.5)(IQR), Q3 + (1.5)(IQR)]$$

    (This is what boxplots do.)

  - For two related variables, a scatterplot will often reveal outliers.

Section 2.7: Analyzing Distributions

Boxplots:
The textbook talks quite a bit about boxplots (box and whisker plots) in this section.

We'll use a boxplot in the next problem, but if you don't remember what these charts are or how to create them in Excel, it's a good idea to read through this part of the textbook.

Example

A group of electronics stores wants to better understand how well a certain tablet is selling at their stores. Below is the number of those tables each store sold in a given day:

$$58, 67, 67, 82, 91, 92, 92, 103, 110, 178.$$

Find the IQR of this data set, and use it to identify any potential outliers. (Create a boxplot in Excel to see if this lines up with what you got.)

Section 2.8: Measures of Association Between Two Variables

So far, we've been looking at numerical methods used to summarize the data for one variable at a time.

Now we're going to look at descriptive measures and charts we can use to investigate and understand the relationship between two variables.

Section 2.8: Measures of Association Between Two Variables

A **scatter chart** (or plot) is useful for analyzing the relationship between two variables.

Example

Open the “BottledWater” file in Excel. This data shows the daily water bottle sales at Queensland Amusement Park and the high temperature for each of 14 summer days. The sales manager believes that daily bottled water sales in the summer are related to the outdoor temperature.

Create a scatter chart showing the relationship between sales and temperature.

Section 2.8: Measures of Association Between Two Variables

Definition

Covariance: descriptive measure of the linear association between two variables

$$\text{sample covariance} = s_{xy} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{n-1}$$

- This calculation yields a positive or negative number which indicated a positive or negative linear relationship
- a number close to zero indicates no linear relationship
- the magnitude of the covariance is difficult to interpret because the units of the two variables might be difference (like $ versus pounds of chocolate)
- In Excel use COVARIANCE.S (for "sample") to calculate

Section 2.8: Measures of Association Between Two Variables

Example

Open the “BottledWater” file again in Excel. Find the sample covariance using the formula COVARIANCE.S(A2:A15, B2:B15).

Section 2.8: Measures of Association Between Two Variables

Definition

Correlation coefficient: measures the relationship between two variables, and, unlike covariance, the relationship is not affected by the units of measurement for x and y

$$\text{sample correlation coefficient} = r_{xy} = \frac{s_{xy}}{s_x s_y}$$

where

$$\begin{aligned} s_{xy} &= \text{ sample covariance} \\ s_x &= \text{ sample standard deviation of } x \\ s_y &= \text{ sample standard deviation of } y \end{aligned}$$

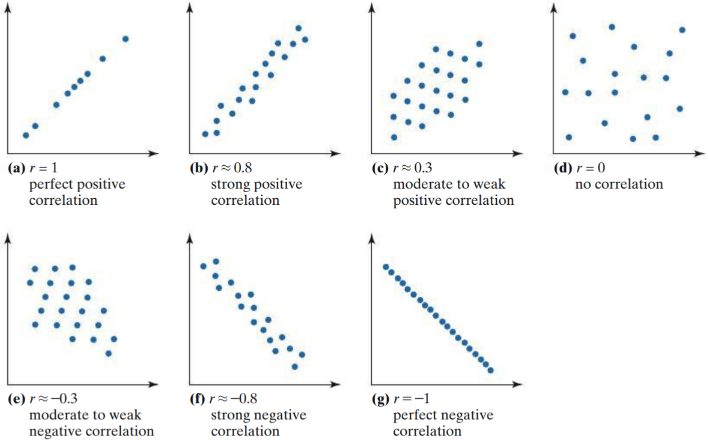
Section 2.8: Measures of Association Between Two Variables

Correlation coefficient:

- Not affected by the variable units, as covariance is
- Calculation results in values between -1 and 1.
  - Near zero: no linear relationship
  - Close to -1 or 1: indicates a strong negative or positive linear relationship between the two variables
- In Excel, use the CORREL formula to calculate

Section 2.8: Measures of Association Between Two Variables

Correlation coefficient Examples:



(Image from the book Thinking Mathematically by Blitzer)

Section 2.8: Measures of Association Between Two Variables

Example

Open the “BottledWater” file again in Excel. You can find the correlation coefficient for the sales of bottled water using the formula CORREL(A2:A15, B2:B15).

(A2:A15 defines the range for the x variable, and B2:B15 defines the range for the y variable.)

Data Cleansing

(This is discussed more in depth in Chapter 4 in the textbook.)

> To gain an understanding of the data you are analyzing, start by checking over the data.

Missing Data:

- Legitimately missing (a sensor was not turned on, so no data)
- Illegitimately missing (a sensor fails to send a reading, a respondent does not answer a survey question)

Data Cleansing

Illegitimately missing data:

Four primary options:

1. discard observations with any missing values (throw out the survey from that respondent)
2. discard any variable with missing values (throw out all data for a survey question that has missing answers)
3. fill in missing data with estimated values (like the average of the existing values)
4. use an analysis algorithm that can handle missing values

Data Cleansing

> “Deciding on a strategy for dealing with missing data requires some understanding of why the data are missing and the potential impact these missing values might have on an analysis.”
> –Our textbook

Data Cleansing

Why is a data point missing? (The following are standard definitions and acronyms.)

- **Missing Completely at Random (MCAR)**
  The fact that the data is missing is not related to the value of the data or any other variable in the data set. (Literally random, like someone forgot to fill in an answer.)

- **Missing at Random (MAR):** (not a very helpful name)
  The missing data point is not related to the missing data, but is related to the values of some other data.

  Example: patient visits are missing the results of a diagnostic test whenever the doctor deems the patient too sick to undergo the testing

Data Cleansing

- **Missing Not at Random (MNAR):**
  The tendency for the data to be missing is related to the value that is missing.

  Example: People are asked on a survey to self-report their income. People with high incomes might be less inclined to respond to that question than people with low incomes.

Data Cleansing

Why is it important to know why data is missing?

- If there are a small number of MCAR or MAR data points we can usually ignore them.
- Missing values that are MNAR cannot be ignored – they will bias the analysis.

Data Cleansing

Definition

imputation: filling in a missing data point with a reasonable value (such as the mean of all of the existing data points)

Data Cleansing

Example

Blakely Tires (This is on pg. 64-65 of our textbook.)

Open the Excel file “treadwear” that is on Canvas.

We will likely not work through this entire example that is in the book. I recommend reading through this on your own.

Data Cleansing

Example

Blakely Tires (Continued)

Blakely Tires is a U.S. producer of automobile tires. In an attempt to learn about the conditions of its tires on automobiles in Texas, the company has obtained information for each of the four tires from 116 automobiles with Blakely brand tires that have been collected through recent state automobile inspection facilities in Texas. The data obtained by Blakely includes the position of the tire on the automobile (left front, left rear, right front, right rear), age of the tire, mileage on the tire, and depth of the remaining tread on the tire. Before Blakely management attempts to learn more about its tires on automobiles in Texas, it wants to assess the quality of these data.

Data Cleansing

Example

Blakely Tires (Continued)

The tread depth of a tire is a vertical measurement between the top of the tread rubber to the bottom of the tire's deepest grooves, and is measured in 32nds of an inch in the United States. New Blakely brand tires have a tread depth of 10/32nds of an inch, and a tire's tread depth is considered insufficient if it is 2/32nds of an inch or less. Shallow tread depth is dangerous as it results in poor traction and so makes steering the automobile more difficult. Blakely's tires generally last for four to five years or 40,000 to 60,000 miles.

Let's assess the quality of these data by determining which (if any) observations have missing values for any of the variables in the data.

Data Cleansing

How do we identify erroneous values?

- We have a whole set of statistical tools that can be used to find "data quality" issues and outliers:
  - summary statistics
  - scatterplots
  - z-scores
  - etc.

Data Cleansing

How do we identify erroneous values? (Continued)

- The Blakely Tires example:
  - Take the mean and standard deviation of each variable. (like Life of Tires)
    - Do the values seem reasonable?
    - Are there individual data values that stray especially far from the mean?

Data Cleansing

How do we identify erroneous values? (Continued)

- The Blakely Tires example:
  - Look at the min and max values for each variable. In this example, Tire Life min = 1.8 months and Tire Life max = 601 months.
    - Is this reasonable? (600 months = 50 years)
    - In this case we can sort the data by car ID# and find that the life of the other 3 tires on the car is 60.1 months, so the value of 601 months is likely a typo.

Data Cleansing

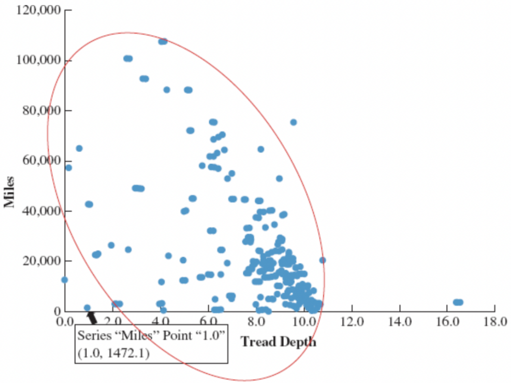
How do we identify erroneous values? (Blakely Tires Continued)

- If two variables are related, we can make a scatterplot to see if there are outliers (that defy the relationship).

  The figure on the next slide shows Tire Miles vs. Tread Depth. (Do we expect these to be related?)

Data Cleansing

Figure 2.35 **Scatter Chart of Tread Depth and Miles for the *TreadWear* Data**



In this chart, values near the origin represent tires with low miles and low tread depth, which doesn't make sense.
(Be sure you understand your data!)

Data Cleansing

Variable representation

- It's possible to have so many variables in a data set that it is difficult to analyze
  - If some pairs or groups of variables are closely related (correlated), then they are providing similar information, and some can be removed from the analysis.

Definition

Dimension reduction is the process of removing variables from the analysis without losing crucial information.