

Measures of Association Between Two Variables

Scatter Charts Covariance

Correlation Coefficient

96



Measures of Association Between Two Variables

Table 2.14: Data for Bottled Water Sales at Queensland Amusement Park for a Sample of 14 Summer Days

|  |  |
| --- | --- |
| **High Temperature (°F)** | **Bottled Water Sales**  **(cases)** |
| 78 | 23 |
| 79 | 22 |
| 80 | 24 |
| 80 | 22 |
| 82 | 24 |
| 83 | 26 |
| 85 | 27 |
| 86 | 25 |
| 87 | 28 |
| 87 | 26 |
| 88 | 29 |
| 88 | 30 |
| 90 | 31 |
| 92 | 31 |

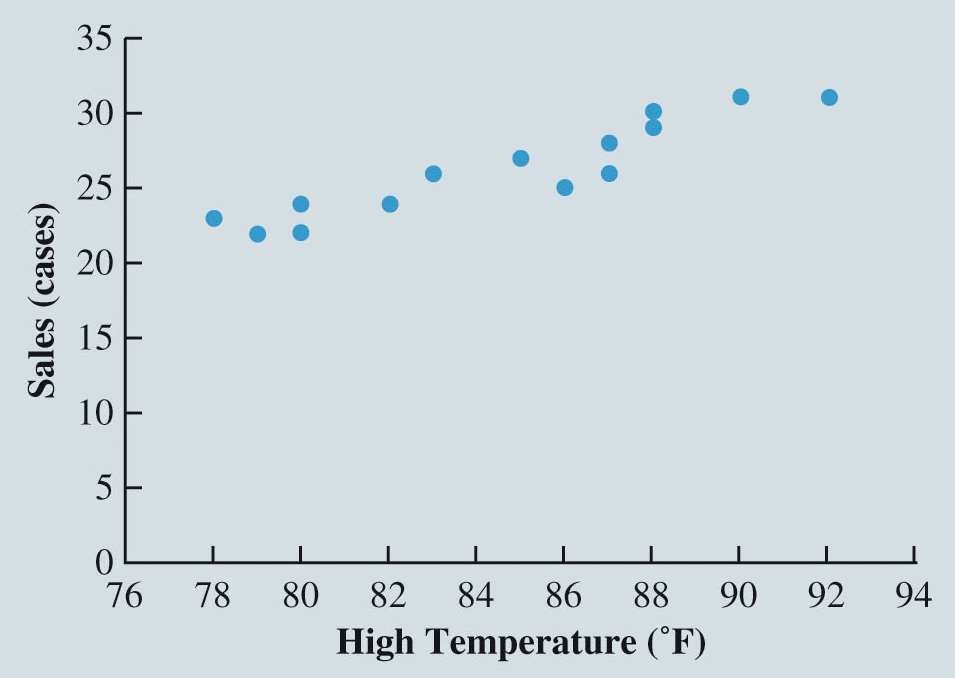
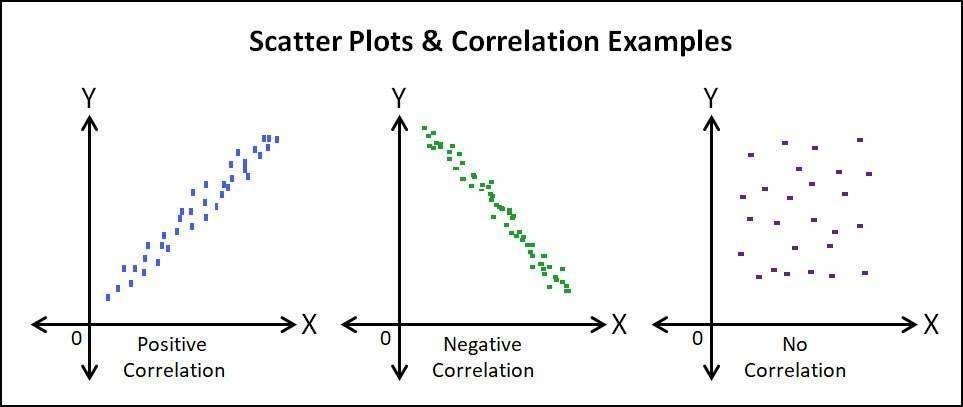
97



Measures of Association Between Two Variables

* A **scatter chart:**
  + is a useful graph for analyzing the relationship between two variables.

98

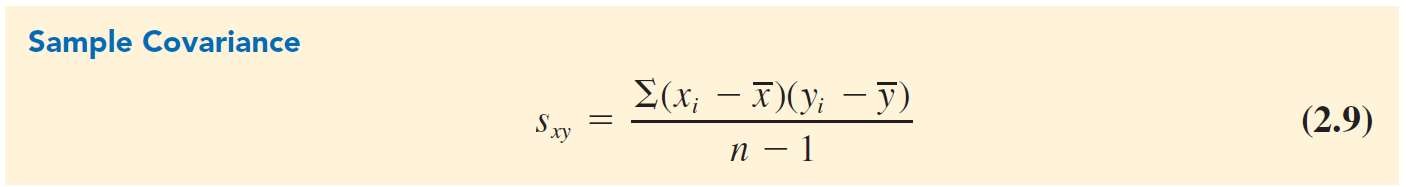


Measures of Association Between Two Variables

Sales and High Temperatures

* Positive relationship:
  + As high temperature increases, sales of bottled water generally also increases.
* Best Fit :
  + a straight line could be used as an approximation for the relationship between high temperature and sales of bottled water.

99



Measures of Association Between Two Variables

* **Covariance:**
  + is a descriptive measure of the linear association between two variables:

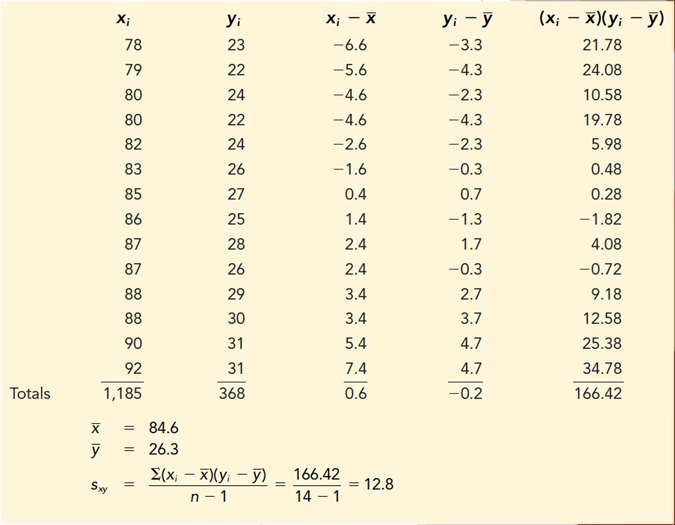
Population covariance, 𝜎xy =

Σ 𝑥i − 𝜇x Σ 𝑦i − 𝜇y

𝑁

.

100



Measures of Association Between Two Variables

Daily High Temperature and Bottled Water Sales at Queensland Amusement Park

* **If Covariance:**
  + Large and > 0 => Strong Positive Relationship
  + Near 0 => Not Linearly Related
  + Large and <0 => Strong Negative Relationship

101



Measures of Association Between Two Variables

The **correlation coefficient:**

* Measures the relationship between two variables.
* Not affected by the units of measurement for *x* and *y.*

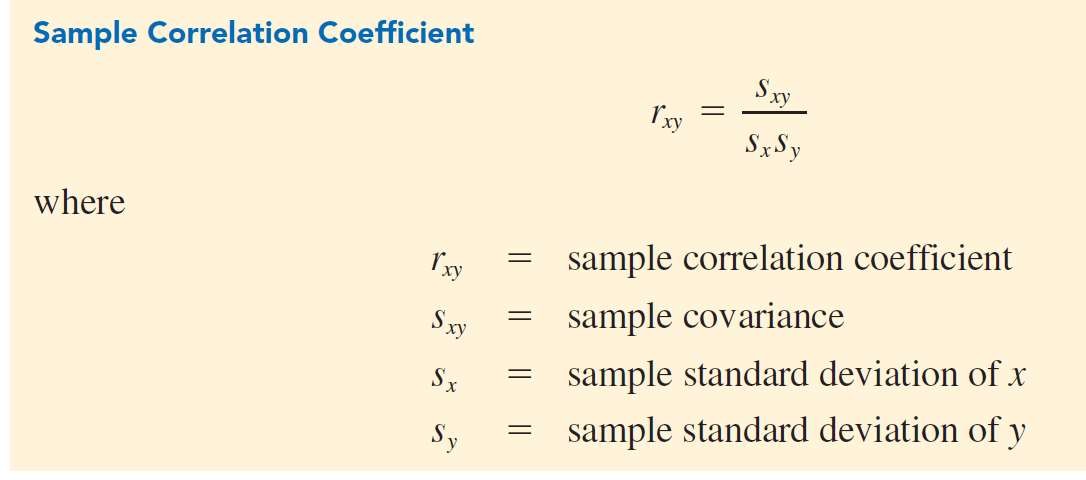
∑(𝑥i − 𝑥)(𝑦i − 𝑦¯)

𝑟xy = 𝑛 − 1

∑(𝑥i − 𝑥)2 ∑(𝑦i − 𝑦¯)2

𝑛 − 1 𝑛 − 1

102



Measures of Association Between Two Variables

Interpretation of Correlation Coefficient:

1  *r*  +1

|  |  |
| --- | --- |
| ***r* value** | **Relationship between the *x* and *y* variables** |
| < 0 | Negative linear |
| Near 0 | No linear relationship |
| > 0 | Positive linear |

103



Measures of Association Between Two Variables

Illustration:

* To determine the sample correlation coefficient for bottled water sales at Queensland Amusement Park:

*r*

*xy*

= *sxy*

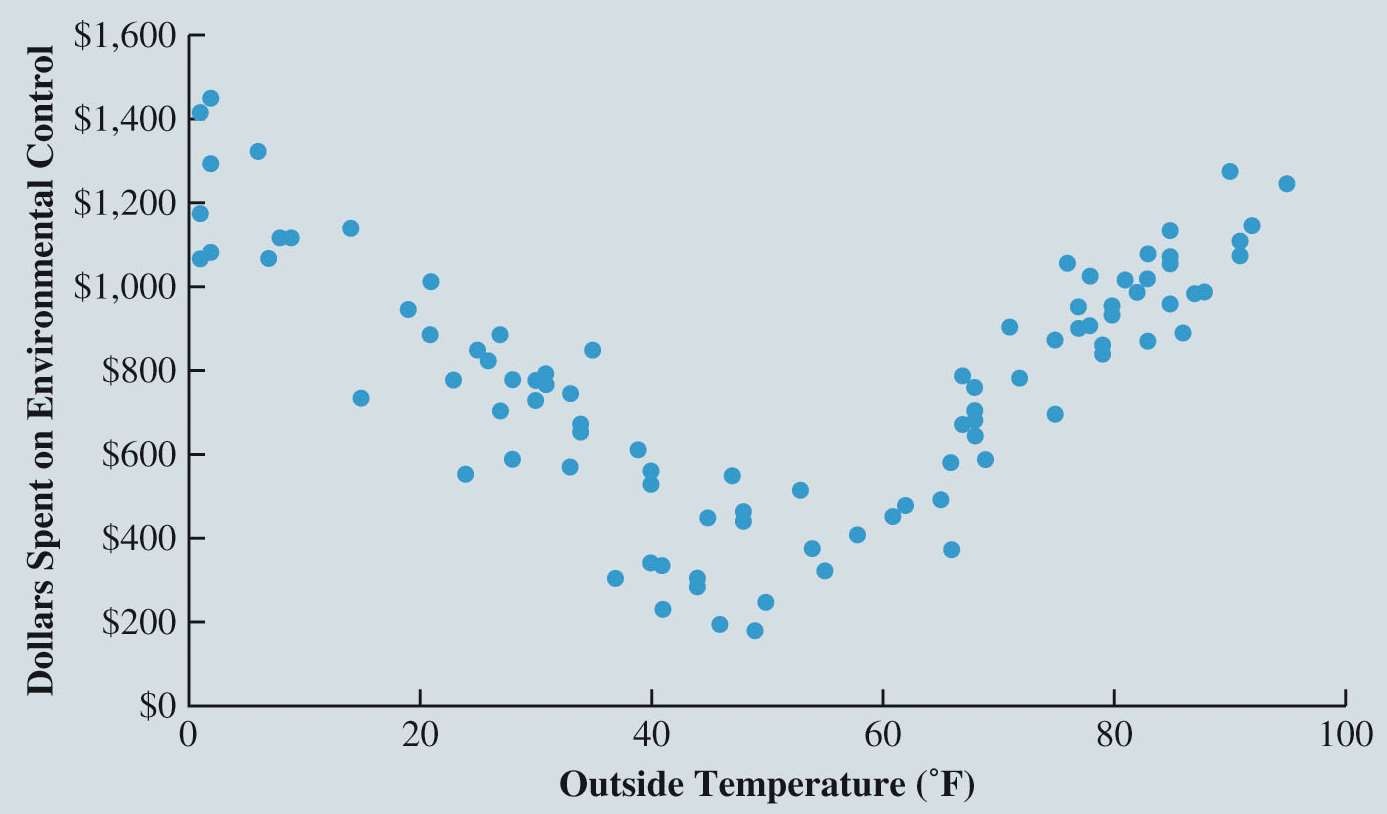
*s s* (4.36)(3.15)

= 12.8 = 0.93

*x y*

* There is a very strong linear relationship between high temperature and sales.

104



Measures of Association Between Two Variables

Nonlinear Relationship Producing a Correlation Coefficient Near Zero

*Correlation Coefficient = rxy* = –0.007

What does this indicate?

Is there a different type of relationship?

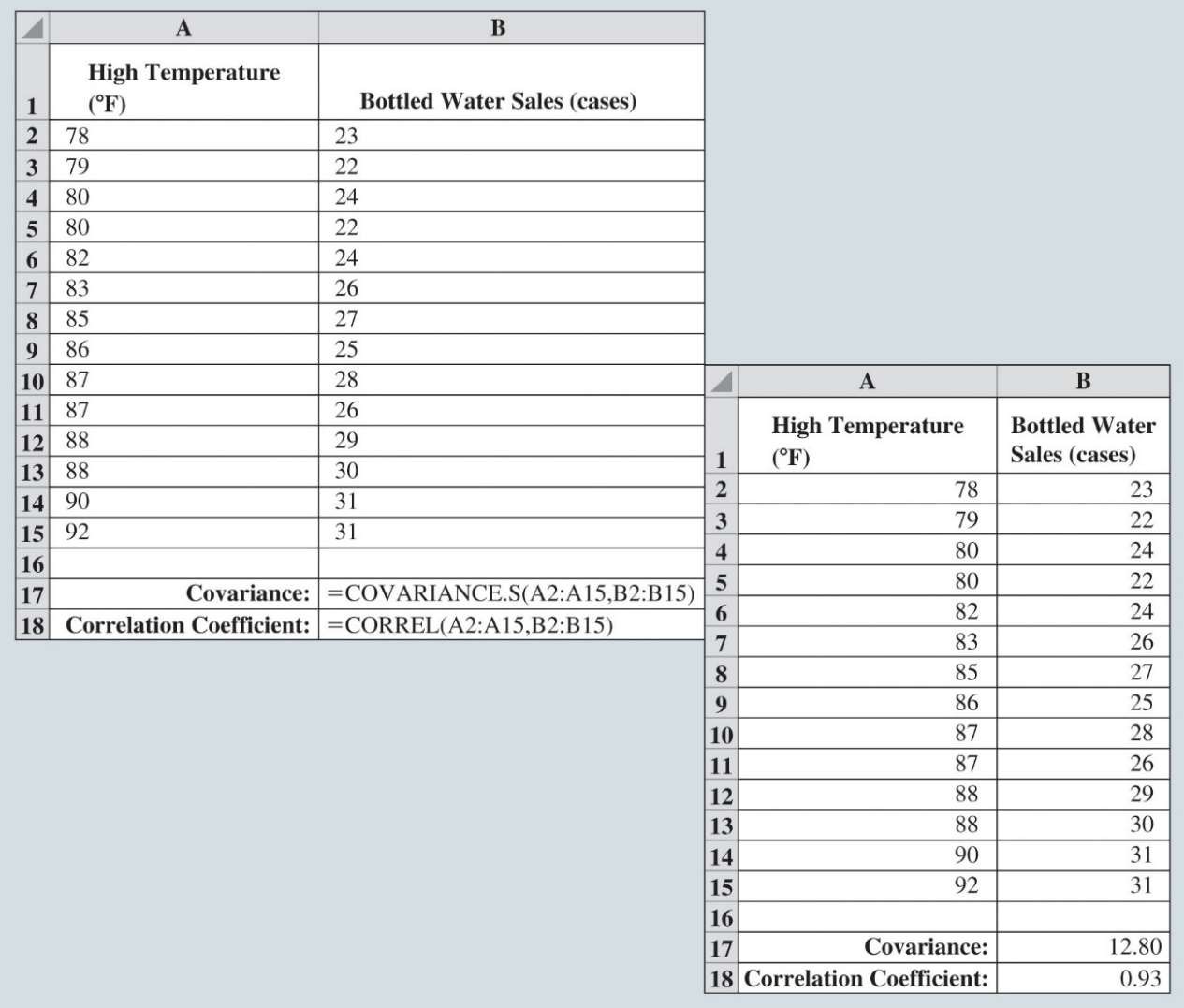
105



Measures of Association Between Two Variables

Figure 2.27: Calculating Covariance and Correlation Coefficient for Bottled Water Sales Using Excel

106



Data Cleaning

Missing Data Blakely Tires

Identification of Erroneous Outliers and other Erroneous Values Variable Representation

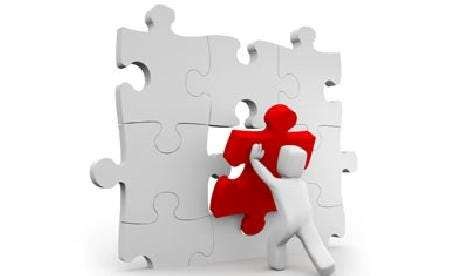
107



Data Cleaning

* Data sets commonly include observations with missing values for one or more variables.
* In some cases, missing data naturally occur; these are called:
* **Legitimately missing data.**
  + Generally, no remedial action is taken for legitimately missing data.

108



Data Cleaning

* **Illegitimately missing data.**
* Addressing such missing data:
  1. Discard observations (rows) with any missing values.
  2. Discard any variable (column) with missing values.
  3. Fill in missing entries with estimated values.
  4. Apply a data-mining algorithm that can handle missing values.

109

Data Cleaning

* **Missing completely at random**

(MCAR):

* whether data are missing does not depend on either the value of the missing data or the value of any other variable in the data.

**Probability of missing data is same for all cases**

**Rare in the Real World**

110



Data Cleaning

* **Missing at random** (MAR):
* The tendency for an observation to be missing a value for some variable is related to the value of some other variable(s) in the data.

**Probability of missing data is NOT same for all cases**

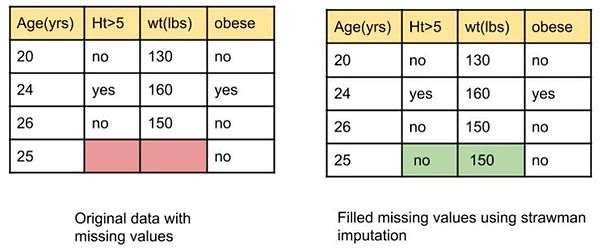
**More Realistic!**

111



Data Cleaning

* **Imputation**:
  + The systematic replacement of missing values with values that seem reasonable.

112



Data Cleaning

* **Missing not at random** (MNAR):

The tendency for the value of a variable to be missing is related to the value that is missing.

**Heavier weight => Missing Values**

Others: Weaker opinionated people are less likely to respond to survey

113



Data Cleaning

**Identification of Erroneous Outliers and other Erroneous Values:**

* Uncover Data=quality Issues using:
  + summary statistics, frequency distributions, bar charts and histograms, z-scores, scatter plots, correlation coefficients, and other tools
* Many software ignore missing values when calculating various summary statistics.
* Warning:
  + Sometime missing values are indicated with a unique value (such as 9999999)
    - These values may skew summary statistics.
  + Find missing data before running summary statistics OR
  + Look over summary statistics very carefully!

114



Data Cleaning



# Variable Representation:

* Data-Mining
  + May get too many variables to work with
* **Dimension reduction:**
  + is the process of removing variables from the analysis without losing crucial information.

# Critical:

* + Determine how to represent measurements of variables and which variables to consider.
  + Consider relationships or combinations of variables for more insights

115

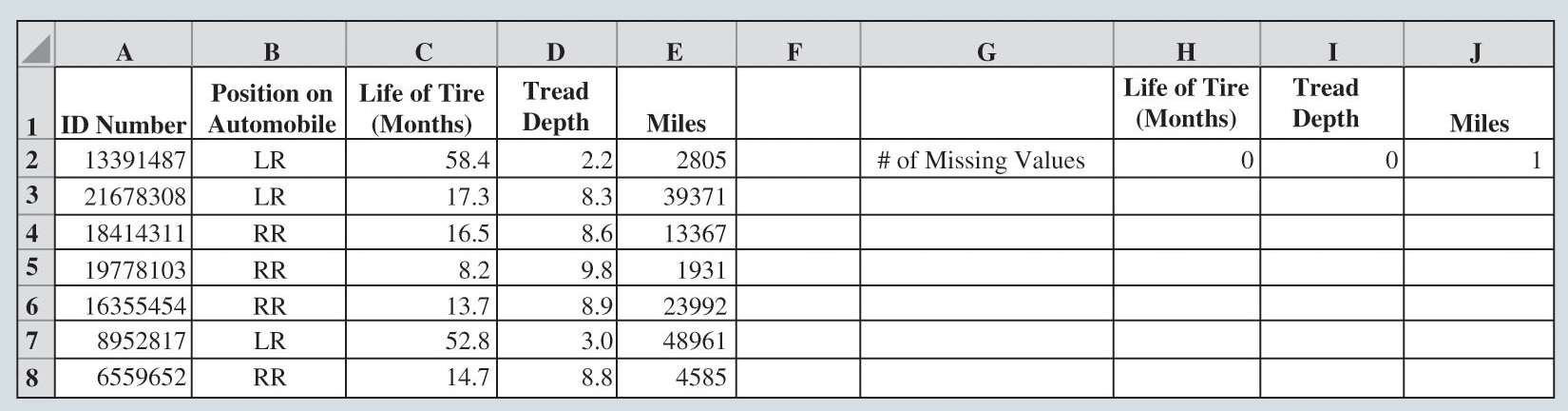


Data Cleaning

Blakely Tires:

* A U.S. producer of automobile tires wants to learn about the conditions of its tires on automobiles in Texas.
* The data obtained includes the position of the tire on the automobile, age of the tire, mileage on the tire, and depth of the remaining tread on the tire.
* Begin assessing the quality of these data by determining which (if any) observations have missing values (see Figure 2.30).

116



Data Cleaning

Figure 2.30: Portion of Excel Spreadsheet Showing Number of Missing Values for Variables in

*TreadWear* Data

117



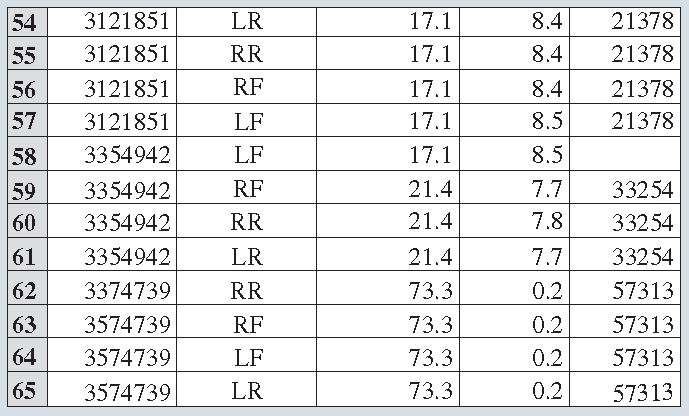
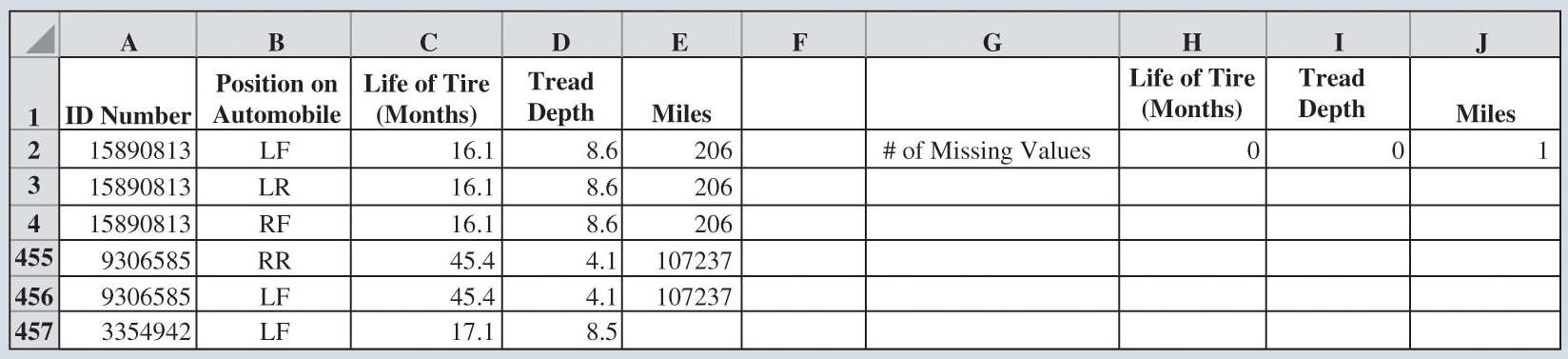
Data Cleaning

Blakely Tires (cont.):

* Sort all of Blakely’s data on Miles from smallest to largest value to determine which observation is missing its value of this variable.

Figure 2.31: Portion of Excel Spreadsheet Showing *TreadWear* Data Sorted on Miles from Lowest to Highest Value

118



Data Cleaning

Figure 2.32: Portion of Excel Spreadsheet Showing *TreadWear* Data Sorted from Lowest to Highest by ID Number

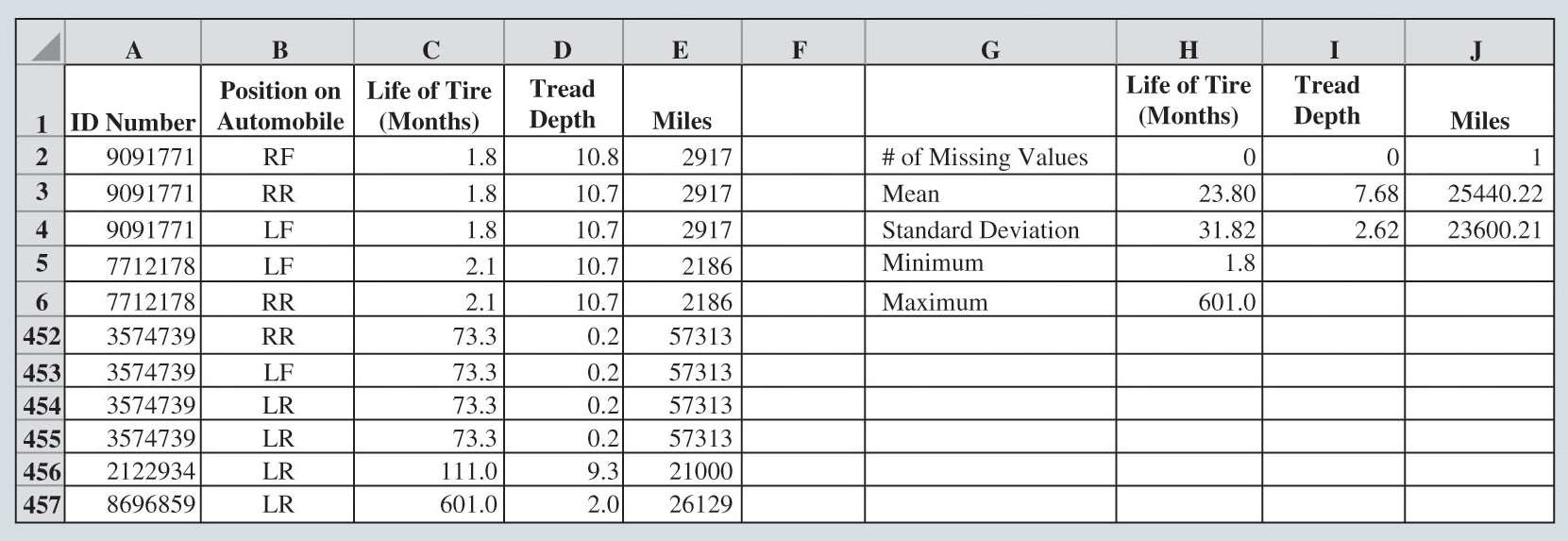
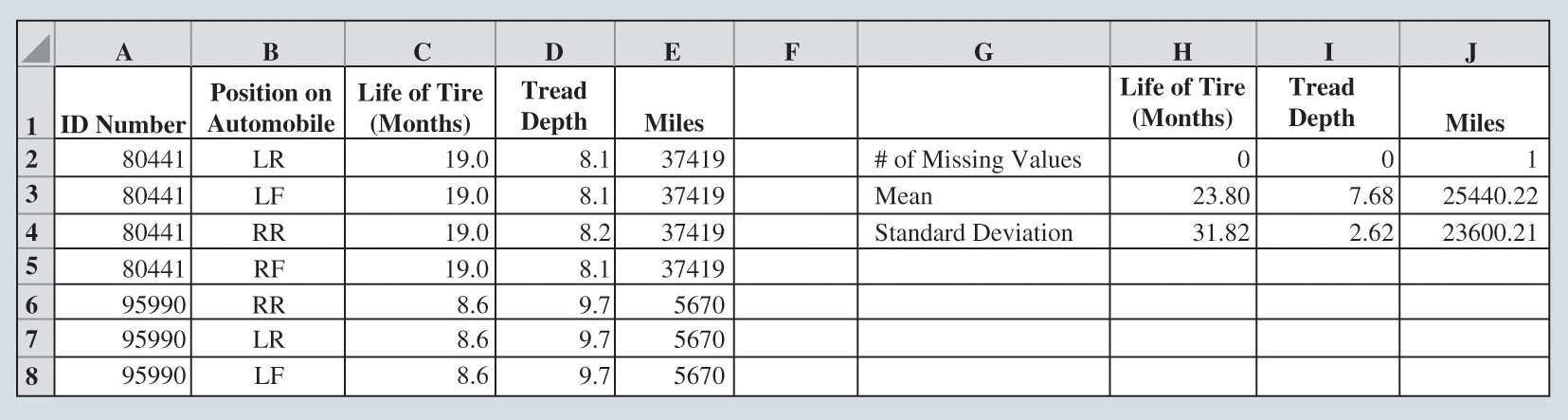
119



Data Cleaning

Figure 2.33: Portion of Excel Spreadsheet Showing the Mean and Standard Deviation for Each Variable in the *TreadWear* Data

120



Data Cleansing

Figure 2.34: Portion of Excel Spreadsheet Showing the *TreadWear* Data Sorted on Life of Tires (Months) from Lowest to Highest Value

121



Data Cleaning

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

122

