APPENDIX

DATA SUMMARIZATION AND VISUALIZATION

Here, we present a very brief review of methods for summarizing and visualizing data. For deeper coverage, please see *Discovering Statistics*, by Daniel Larose (second edition, W.H. Freeman, New York, 2013).

PART 1 SUMMARIZATION 1: BUILDING BLOCKS OF DATA ANALYSIS

- **Descriptive statistics** refers to methods for summarizing and organizing the information in a data set.
  Consider Table A.1, which we will use to illustrate some statistical concepts.
- The entities for which information is collected are called the **elements**. In Table A.1, the elements are the 10 applicants. Elements are also called **cases** or **subjects**.
- A **variable** is a characteristic of an element, which takes on different values for different elements. The variables in Table A.1 are *marital status, mortgage, income, rank, year,* and *risk*. Variables are also called **attributes**.

Discovering Knowledge in Data: An Introduction to Data Mining, Second Edition.
By Daniel T. Larose and Chantal D. Larose.

TABLE A.1 Characteristics of 10 loan applicants

Applicant	Marital Status	Mortgage	Income ($)	Income Rank	Year	Risk
1	Single	y	38,000	2	2009	Good
2	Married	y	32,000	7	2010	Good
3	Other	n	25,000	9	2011	Good
4	Other	n	36,000	3	2009	Good
5	Other	y	33,000	4	2010	Good
6	Other	n	24,000	10	2008	Bad
7	Married	y	25,100	8	2010	Good
8	Married	y	48,000	1	2007	Good
9	Married	y	32,100	6	2009	Bad
10	Married	y	32,200	5	2010	Good

- The set of variable values for a particular element is an **observation**. Observations are also called **records**. The observation for Applicant 2 is:

Applicant	Marital Status	Mortgage	Income ($)	Income Rank	Year	Risk
2	Married	y	32,000	7	2010	Good

- Variables can be either *qualitative* or *quantitative*.
  - A **qualitative variable** enables the elements to be classified or categorized according to some characteristic. The qualitative variables in Table A.1 are *marital status, mortgage*, *rank,* and *risk*. Qualitative variables are also called **categorical variables**.
  - A **quantitative variable** takes numeric values and allows arithmetic to be meaningfully performed on it. The quantitative variables in Table A.1 are *income* and *year*. Quantitative variables are also called **numerical variables**.
- Data may be classified according to four *levels of measurement*: *nominal, ordinal, interval,* and *ratio*. Nominal and ordinal data are categorical; interval and ratio data are numerical.
  - **Nominal data** refer to names, labels, or categories. There is no natural ordering, nor may arithmetic be carried out on nominal data. The nominal variables in Table A.1 are *marital status, mortgage*, and *risk*.
  - **Ordinal data** can be rendered into a particular order. However, arithmetic cannot be meaningfully carried out on ordinal data. The ordinal variable in Table A.1 is *income rank*.
  - **Interval data** consist of quantitative data defined on an interval without a natural zero. Addition and subtraction may be performed on interval data. The interval variable in Table A.1 is *year*. (Note that there is no "year zero." The calendar goes from 1 B.C. to 1 A.D.)
  - **Ratio data** are quantitative data for which addition, subtraction, multiplication, and division may be performed. A natural zero exists for ratio data. The interval variable in Table A.1 is *income*.
- A numerical variable that can take either a finite or a countable number of values is a **discrete** variable, for which each value can be graphed as a

separate point, with space between each point. The discrete variable in Table A.1 is *year*.

- A numerical variable that can take infinitely many values is a **continuous variable**, whose possible values form an interval on the number line, with no space between the points. The continuous variable in Table A.1 is *income*.
- A **population** is the set of all elements of interest for a particular problem. A **parameter** is a characteristic of a population. For example, the population is the set of all American voters, and the parameter is the proportion of the population who supports a $1 per ton tax on carbon.
  - The value of a parameter is usually unknown, but it is a constant.
- A **sample** consists of a subset of the population. A characteristic of a sample is called a **statistic**. For example, the sample is the set of American voters in your classroom, and the statistic is the proportion of the sample who supports a $1 per ton tax on carbon.
  - The value of a statistic is usually known, but it changes from sample to sample.
- A **census** is the collection of information from every element in the population. For example, the census here would be to find from every American voter whether they support a $1 per ton tax on carbon. Such a census is impractical, so we turn to statistical inference.
- **Statistical inference** refers to methods for estimating or drawing conclusions about population characteristics based on the characteristics of a sample of that population. For example, suppose 50% of the voters in your classroom support the tax; using statistical inference, we would *infer* that 50% of all American voters support the tax. Obviously, there are problems with this. The sample is neither random nor representative. The estimate does not have a confidence level, and so on.
- When we take a sample for which each element has an equal chance of being selected, we have a **random sample**.
- A **predictor variable** is a variable whose value is used to help predict the value of the *response variable*. The predictor variables in Table A.1 are all the variables except *risk*.
- A **response variable** is a variable of interest whose value is presumably determined at least in part by the set of predictor variables. The response variable in Table A.1 is *risk*.

PART 2 VISUALIZATION: GRAPHS AND TABLES FOR SUMMARIZING AND ORGANIZING DATA

2.1 Categorical Variables

- The **frequency** (or **count**) of a category is the number of data values in each category. The **relative frequency** of a particular category for a categorical variable equals its frequency divided by the number of cases.

TABLE A.2 Frequency distribution and relative frequency distribution

Category of Marital Status	Frequency	Relative Frequency
Married	5	0.5
Other	4	0.4
Single	1	0.1
Total	10	1.0

- A (**relative**) **frequency distribution** for a categorical variable consists of all the categories that the variable assumes, together with the (relative) frequencies for each value. The frequencies sum to the number of cases; the relative frequencies sum to 1.

  For example Table A.2 contains the frequency distribution and relative frequency distribution for the variable *marital status* for the data from Table A.1.
- A **bar chart** is a graph used to represent the frequencies or relative frequencies for a categorical variable. Note that the bars do not touch.
  - A **Pareto chart** is a bar chart where the bars are arranged in decreasing order. Figure A.1 is an example of a Pareto chart.
- A **pie chart** is a circle divided into slices, with the size of each slice proportional to the relative frequency of the category associated with that slice. Figure A.2 shows a pie chart of *marital status*.

2.2 Quantitative Variables

- Quantitative data are grouped into **classes**. The **lower** (**upper**) **class limit** of a class equals the smallest (largest) value within that class. The **class width** is the difference between successive lower class limits.

Figure A.1 Bar chart for *marital status*.

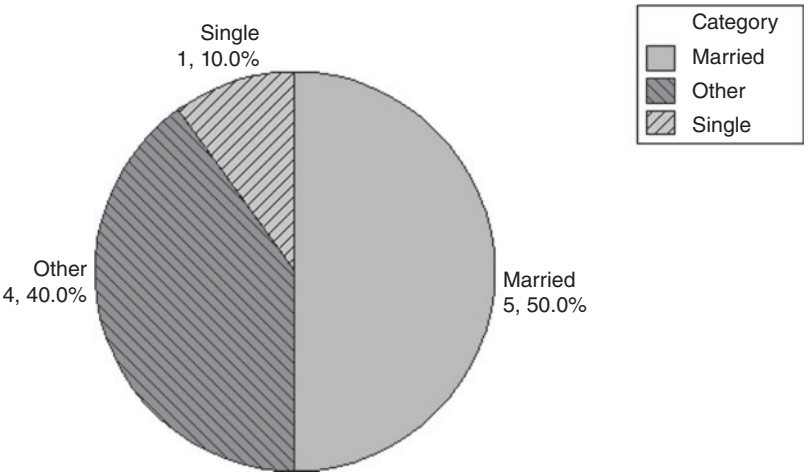

Figure A.2 Pie chart of *marital status*.

- For quantitative data, a **(relative) frequency distribution** divides the data into nonoverlapping classes of equal class width. Table A.3 shows the frequency distribution and relative frequency distribution of the continuous variable *income* from Table A.1.
- A **cumulative (relative) frequency distribution** shows the total number (relative frequency) of data values less than or equal to the upper class limit. See Table A.4.
- A **distribution** of a variable is a graph, table, or formula that specifies the values and frequencies of the variable for all elements in the data set. For example, Table A.3 represents the distribution of the variable *income*.
- A **histogram** is a graphical representation of a (relative) frequency distribution for a quantitative variable. See Figure A.3. Note that histograms represent a simple version of *data smoothing* and can thus vary in shape depending on the number and width of the classes. Therefore, histograms should be interpreted with caution. See *Discovering Statistics*, by Daniel Larose (W.H. Freeman) section 2.4 for an example of a data set presented as *both* symmetric and right-skewed by altering the number and width of the histogram classes.
- A **stem-and-leaf display** shows the shape of the data distribution while retaining the original data values in the display, either exactly or approximately. The

TABLE A.3 Frequency distribution and relative frequency distribution of *income*

Class of *Income*	Frequency	Relative Frequency
$24,000–$29,999	3	0.3
$30,000–$35,999	4	0.4
$36,000–$41,999	2	0.2
$42,000–$48,999	1	0.1
Total	10	1.0

TABLE A.4 Cumulative frequency distribution and cumulative relative frequency distribution of *income*

Class of *Income*	Cumulative Frequency	Cumulative Relative Frequency
\$24,000–\$29,999	3	0.3
\$30,000–\$35,999	7	0.7
\$36,000–\$41,999	9	0.9
\$42,000–\$48,999	10	1.0

leaf units are defined to equal a power of 10, and the stem units are 10 times the leaf units. Then each leaf represents a data value, through a stem-and-leaf combination. For example, in Figure A.4, the leaf units (right-hand column) are 1000s and the stem units (left-hand column) are 10,000s. So "2 4" represents $2 \times 10,000 + 4 \times 1000 = \$24,000$, while "2 55" represents two equal incomes of \$25,000 (one of which is exact, the other approximate, \$25,100). Note that Figure A.4, turned 90 degrees to the left, presents the shape of the data distribution.

- In a **dotplot,** each dot represents one or more data values, set above the number line. See Figure A.5.
- A distribution is **symmetric** if there exists an axis of symmetry (a line) that splits the distribution into two halves that are approximately mirror images of each other (Figure A.6a).
- **Right-skewed** data have a longer tail on the right than the left (Figure A.6b). **Left-skewed** data have a longer tail on the left than the right (Figure A.6c).

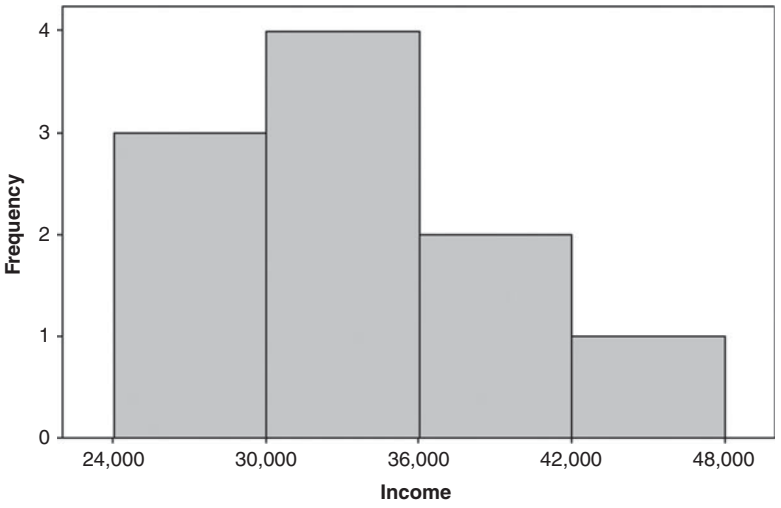

Figure A.3 Histogram of *income*.

```
Stem-and-leaf of Income
Leaf Unit = 1000

     2  4
     2  55
     3  2223
     3  68
     4
     4  8
```

Figure A.4 Stem-and-leaf display of *income*.

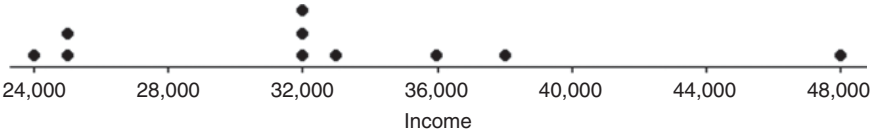

Figure A.5 Dotplot of *income*.

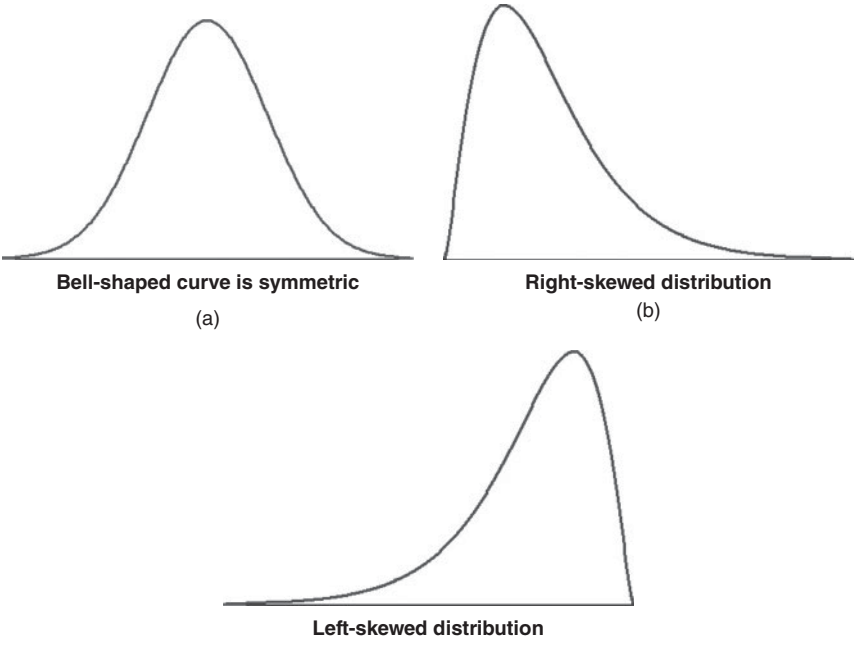

Figure A.6 Symmetric and skewed curves.

PART 3 SUMMARIZATION 2: MEASURES OF CENTER, VARIABILITY, AND POSITION

- The *summation notation* $\sum^{x}$ means to add up all the data values x. The sample size is n and the population size is N.
- **Measures of center** indicate where on the number line the central part of the data is located. The measures of center we will learn are the *mean*, the *median*, the *mode*, and the *midrange*.
  - The **mean** is the *arithmetic average* of a data set. To calculate the mean, add up the values and divide by the number of values. The mean income from Table A.1 is

$$\frac{38{,}000 + 32{,}000 + \cdots + 32{,}200}{10} = \frac{325{,}400}{10} = \$32{,}540$$

  - The **sample mean** is the arithmetic average of a sample, and is denoted $\bar{x}$ ("*x-bar*").
  - The **population mean** is the arithmetic average of a population, and is denoted μ ("*myu*", the Greek letter for *m*).
  - The **median** is the middle data value, when there is an odd number of data values and the data have been sorted into ascending order. If there is an even number, the median is the mean of the two middle data values. When the income data are sorted into ascending order, the two middle values are \$32,100 and \$32,200, the mean of which is the median income, \$32,150.
  - The **mode** is the data value that occurs with the greatest frequency. Both quantitative and categorical variables can have modes, but only quantitative variables can have means or medians. Each income value occurs only once, so there is no mode. The mode for *year* is 2010, with a frequency of 4.
  - The **midrange** is the average of the maximum and minimum values in a data set. The midrange income is

$$\begin{aligned}\text{midrange (income)} &= \frac{(\max(\text{income}) + \min(\text{income}))}{2} = \frac{48{,}000 + 24{,}000}{2} \\ &= \$36{,}000\end{aligned}$$

- **Skewness and measures of center**. The following are tendencies, and not strict rules.
  - For symmetric data, the mean and the median are approximately equal.
  - For right-skewed data, the mean is greater than the median.
  - For left-skewed data, the median is greater than the mean.
- **Measures of variability** quantify the amount of *variation, spread*, or *dispersion* present in the data. The measures of variability we will learn are the *range*, the *variance*, the *standard deviation*, and, later, the *interquartile range*.

- The **range** of a variable equals the difference between the maximum and minimum values. The range of *income* is

$$\text{Range} = \max(\text{income}) - \min(\text{income}) = 48{,}000 - 24{,}000 = \$24{,}000.$$

- A **deviation** is the signed difference between a data value, and the mean value. For Applicant 1, the deviation in *income* equals $x - \bar{x} = 38{,}000 - 32{,}540 = 5{,}460$. For any conceivable data set, the *mean deviation* always equals zero, because the sum of the deviations equals zero.
- The **population variance** is the mean of the squared deviations, denoted as σ^2 ("*sigma-squared*"):

$$\sigma^2 = \frac{\sum (x - \mu)^2}{N}$$

- The **population standard deviation** is the square root of the population variance: $\sigma = \sqrt{\sigma^2}$.
- The **sample variance** is approximately the mean of the squared deviations, with n replaced by $n - 1$ in the denominator in order to make it an *unbiased estimator* of σ^2. (An **unbiased estimator** is a statistic whose expected value equals its target parameter.)

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

- The **sample standard deviation** is the square root of the sample variance: $s = \sqrt{s^2}$.
- The variance is expressed in *units squared*, an interpretation that may be opaque to nonspecialists. For this reason, the standard deviation, which is expressed in the original units, is preferred when reporting results. For example, the sample variance of *income* is $s^2 = 51{,}860{,}444$ *dollars squared*, the meaning of which may be unclear to clients. Better to report the sample standard deviation $s = \$7201$.
- The sample standard deviation s is interpreted as the size of the *typical deviation*, that is, the size of the typical difference between data values and the mean data value. For example, incomes typically deviate from their mean by \$7201.

- **Measures of position** indicate the relative position of a particular data value in the data distribution. The measures of position we cover here are the *percentile*, the *percentile rank*, the *Z-score*, and the *quartiles*.
  - The ***p*th percentile** of a data set is the data value such that p percent of the values in the data set are at or below this value. The 50th percentile is the median. For example, the median *income* is \$32,150, and 50% of the data values lie at or below this value.
  - The **percentile rank** of a data value equals the percentage of values in the data set that are at or below that value. For example, the percentile rank

of Applicant 1's income of \$38,000 is 90%, since that is the percentage of incomes equal to or less than \$38,000.

- The ***Z*-score** for a particular data value represents how many standard deviations the data value lies above or below the mean. For a sample, the *Z*-score is

$$Z\text{-score} = \frac{x - \bar{x}}{s}$$

For Applicant 6, the *Z*-score is

$$\frac{24{,}000 - 32{,}540}{7201} \approx -1.2$$

The income of Applicant 6 lies 1.2 standard deviations below the mean.

- We may also find data values, given a *Z*-score. Suppose no loans will be given to those with incomes more than 2 standard deviations below the mean. Here, $Z\text{-score} = -2$, and the corresponding minimum income is

$$\text{Income} = Z\text{-score} \cdot s + \bar{x} = (-2)(7201) + 32{,}540 = \$18{,}138$$

No loans will be provided to the applicants with incomes below \$18,138.

- If the data distribution is normal, then the **Empirical Rule** states:
  - About 68% of the data lies within 1 standard deviation of the mean,
  - About 95% of the data lies within 2 standard deviations of the mean,
  - About 99.7% of the data lies within 3 standard deviations of the mean.
- The **first quartile (Q1)** is the 25th percentile of a data set; the **second quartile (Q2)** is the 50th percentile (median); and the **third quartile (Q3)** is the 75th percentile.
- The **interquartile range (IQR)** is a measure of variability that is not sensitive to the presence of outliers. $\text{IQR} = Q3 - Q1$.
- In the **IQR method for detecting outliers**, a data value x is an outlier if either
  - $x \leq Q1 - 1.5(IQR)$, or
  - $x \geq Q3 + 1.5(IQR)$.

- The **five-number summary** of a data set consists of the *minimum*, *Q1*, the *median*, *Q3*, and the *maximum*.
- The **boxplot** is a graph based on the five-number summary, useful for recognizing symmetry and skewness. Suppose for a particular data set (not from Table A.1), we have $Min = 15$, $Q1 = 29$, $Median = 36$, $Q3 = 42$, and $Max = 47$. Then the boxplot is shown in Figure A.7.
  - The box covers the "middle half" of the data from Q1 to Q3.
  - The left whisker extends down to the minimum value which is not an outlier.
  - The right whisker extends up to the maximum value that is not an outlier.
  - When the left whisker is longer than the right whisker, then the distribution is left-skewed. And vice versa.

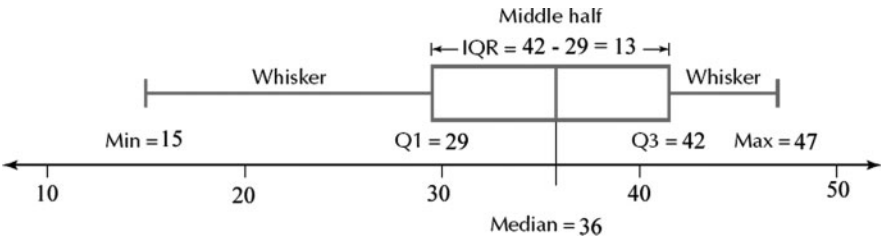

Figure A.7 Boxplot of left-skewed data.

- When the whiskers are about equal in length, the distribution is symmetric. The distribution in Figure A.7 shows evidence of being left-skewed.

PART 4 SUMMARIZATION AND VISUALIZATION OF BIVARIATE RELATIONSHIPS

- A **bivariate relationship** is the relationship between two variables.
- The relationship between two categorical variables is summarized using a **contingency table**, which is a crosstabulation of the two variables, and contains a cell for every combination of variable values (i.e., for every contingency). Table A.5 is the contingency table for the variables *mortgage* and *risk*. The total column contains the **marginal distribution** for *risk*, that is, the frequency distribution for this variable alone. Similarly the total row represents the marginal distribution for *mortgage*.
- Much can be learned from a contingency table. The *baseline proportion* of *bad risk* is 2/10 = 20%. However, the proportion of *bad risk* for applicants without a mortgage is 1/3 = 33%, which is higher than the baseline; and the proportion of *bad risk* for applicants with a mortgage is only 1/7 = 1%, which is lower than the baseline. Thus, whether or not the applicant has a mortgage is useful for predicting risk.
- A **clustered bar chart** is a graphical representation of a contingency table. Figure A.8 shows the clustered bar chart for *risk*, clustered by *mortgage*. Note that the disparity between the two groups is immediately obvious.
- To summarize the relationship between a quantitative variable and a categorical variable, we calculate summary statistics for the quantitative variable for each level of the categorical variable. For example, Minitab provided the following

TABLE A.5 Contingency table for *mortgage* versus *risk*

		Mortgage		
		Yes	No	Total
Risk	Good	6	2	8
	Bad	1	1	2
	Total	7	3	10

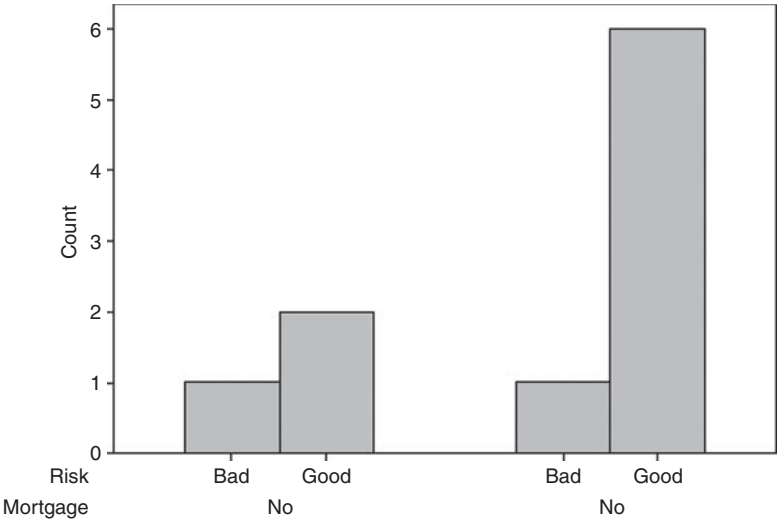

Figure A.8 Clustered bar chart for *risk*, clustered by *mortgage*.

summary statistics for *income*, for records with *bad risk* and for records with *good risk*. All summary measures are larger for *good risk*. Is the difference significant? We need to perform a hypothesis test to find out (Chapter 4).

Descriptive Statistics: Income

Variable	Risk	Mean	StDev	Minimum	Median	Maximum
Income	Bad	28050	5728	24000	28050	32100
	Good	33663	7402	25000	32600	48000

- To visualize the relationship between a quantitative variable and a categorical variable, we may use an **individual value plot**, which is essentially a set of vertical dotplots, one for each category in the categorical variable. Figure A.9 shows the individual value plot for *income* versus *risk*, showing that incomes for *good risk* tend to be larger.

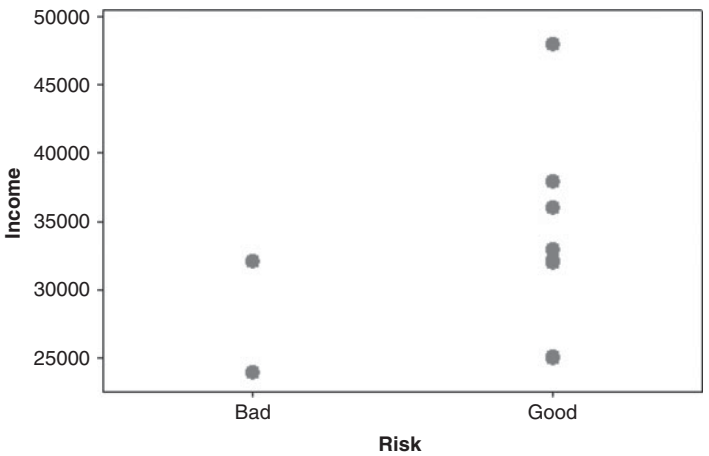

Figure A.9 Individual value plot of *income* versus *risk*.

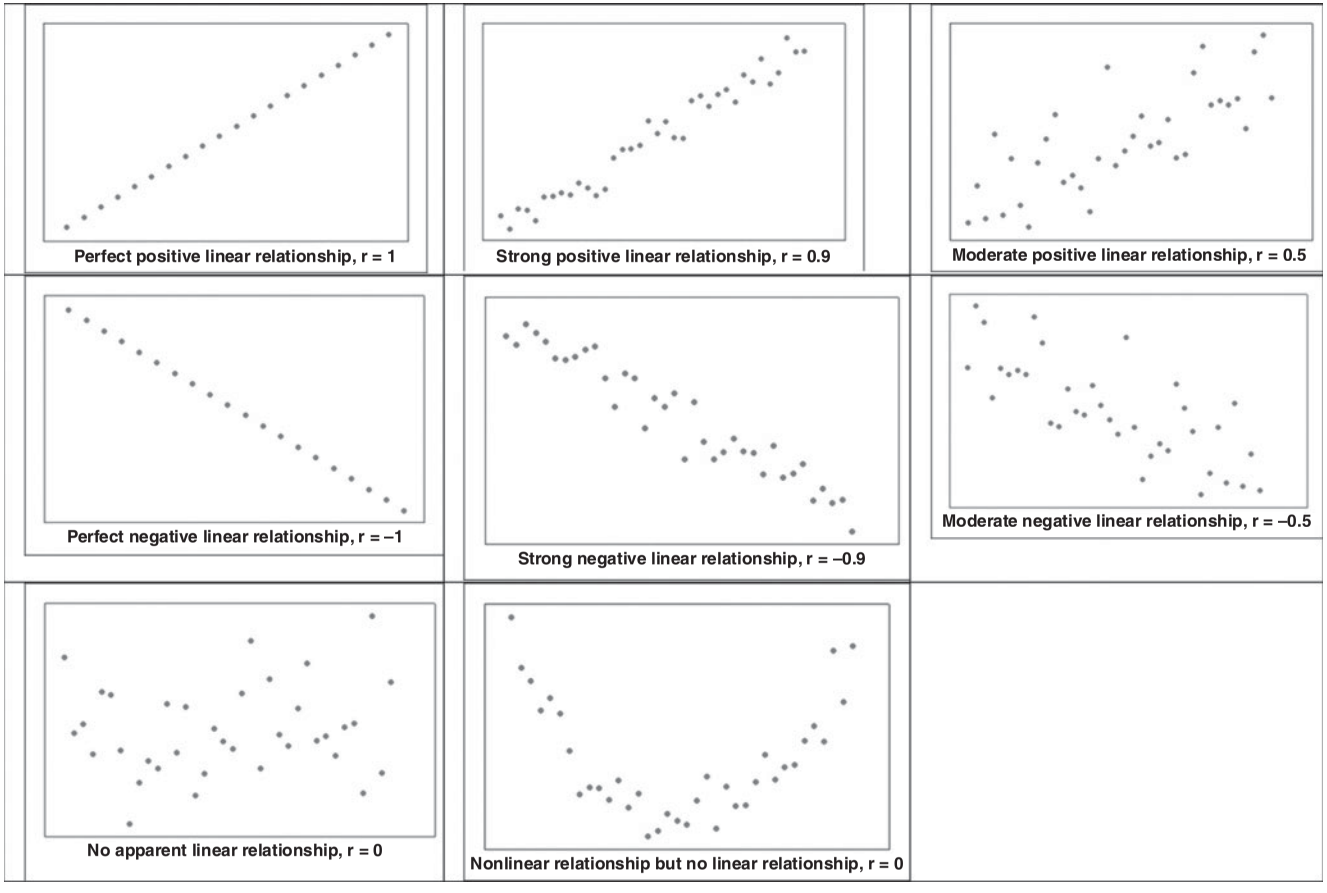

Figure A.10 Some possible relationships between x and y.

- A **scatter plot** is used to visualize the relationship between two quantitative variables, x and y. Each (x, y) point is graphed on a Cartesian plane, with the x axis on the horizontal and the y axis on the vertical. Figure A.10 shows eight scatter plots, showing some possible types of relationships between the variables, along with the value of the *correlation coefficient r*.
- The **correlation coefficient *r*** quantifies the strength and direction of the linear relationship between two quantitative variables. The correlation coefficient is defined as

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{(n-1)\, s_x s_y}$$

  where s_x and s_y represent the standard deviation of the x-variable and the y-variable, respectively. $-1 \leq r \leq 1$.
  - In data mining, where there are a large number of records (over 1000), even small values of r, such as $-0.1 \leq r \leq 0.1$ may be statistically significant.
  - If r is positive and significant, we say that x and y are *positively correlated*. An increase in x is associated with an increase in y.
  - If r is negative and significant, we say that x and y are *negatively correlated*. An increase in x is associated with a decrease in y.