Chapter 4

Descriptive Data Mining

CONTENTS

ANALYTICS IN ACTION

Advice from a Machine[1]

The proliferation of data and increase in computing power have sparked the development of automated *recommender systems*, which provide consumers with suggestions for movies, music, books, clothes, restaurants, dating, and whom to follow on Twitter. The sophisticated, proprietary algorithms guiding recommender systems measure the degree of similarity between users or items to identify recommendations of potential interest to a user.

Netflix, a company that provides media content via DVD-by-mail and Internet streaming, provides its users with recommendations for movies and television shows based on each user's expressed interests and feedback on previously viewed content. As its business has shifted from renting DVDs by mail to streaming content online, Netflix has been able to track its customers' viewing behavior more closely. This allows Netflix's recommendations to account for differences in viewing behavior based on the day of the week, the time of day, the device used (computer, phone, television), and even the viewing location.

The use of recommender systems is prevalent in e-commerce. Using attributes detailed by the Music Genome Project, Pandora Internet Radio plays songs with properties similar to songs that a user "likes." In the online dating world, web sites such as eHarmony, Match.com, and OKCupid use different "formulas" to take into account hundreds of different behavioral traits to propose date "matches." Stitch Fix, a personal shopping service for women, combines recommendation algorithms and human input from its fashion experts to match its inventory of fashion items to its clients.

[1] "The Science Behind the Netflix Algorithms that Decide What You'll Watch Next," http://www.wired.com/2013/08/qq_netflix-algorithm. Retrieved on August 7, 2013; E. Colson, "Using Human and Machine Processing in Recommendation Systems," *First AAAI Conference on Human Computation and Crowdsourcing* (2013); K. Zhao, X. Wang, M. Yu, and B. Gao, "User Recommendation in Reciprocal and Bipartite Social Networks—A Case Study of Online Dating," *IEEE Intelligent Systems* 29, no. 2 (2014).

Over the past few decades, technological advances have led to a dramatic increase in the amount of recorded data. The use of smartphones, radio-frequency identification (RFID) tags, electronic sensors, credit cards, and the Internet has facilitated the collection of data from phone conversations, e-mails, business transactions, product and customer tracking, business transactions, and web browsing. The increase in the use of data-mining techniques in business has been caused largely by three events: the explosion in the amount of data being produced and electronically tracked, the ability to electronically warehouse these data, and the affordability of computer power to analyze the data. In this chapter, we discuss the analysis of large quantities of data in order to gain insight on customers and to uncover patterns to improve business processes.

We define an **observation**, or **record**, as the set of recorded values of variables associated with a single entity. An observation is often displayed as a row of values in a spreadsheet or database in which the columns correspond to the variables. For example, in a university's database of alumni, an observation may correspond to an alumnus's age, gender, marital status, employer, position title, as well as size and frequency of donations to the university.

Predictive data mining is discussed in Chapter 9.

In this chapter, we focus on descriptive data-mining methods, also called **unsupervised learning** techniques. In an unsupervised learning application, there is no outcome variable to predict; rather, the goal is to use the variable values to identify relationships between observations. Unsupervised learning approaches can be thought of as high-dimensional descriptive analytics because they are designed to describe patterns and relationships in large data sets with many observations of many variables. Without an explicit outcome (or one that is objectively known), there is no definitive measure of accuracy. Instead, qualitative assessments, such as how well the results match expert judgment, are used to assess and compare the results from an unsupervised learning method.

4.1 Cluster Analysis

The goal of clustering is to segment observations into similar groups based on the observed variables. Clustering can be employed during the data-preparation step to identify variables or observations that can be aggregated or removed from consideration. Cluster analysis is commonly used in marketing to divide consumers into different homogeneous groups, a process known as **market segmentation**. Identifying different clusters of consumers allows a firm to tailor marketing strategies for each segment. Cluster analysis can also be used to identify outliers, which in a manufacturing setting may represent quality-control problems and in financial transactions may represent fraudulent activity.

In this section, we consider the use of cluster analysis to assist a company called Know Thy Customer (KTC), a financial advising company that provides personalized financial advice to its clients. As a basis for developing this tailored advising, KTC would like to segment its customers into several groups (or clusters) so that the customers within a group are similar with respect to key characteristics and are dissimilar to customers that are not in the group. For each customer, KTC has an observation consisting of the following variables:

DATA file
DemoKTC

Age	= age of the customer in whole years
Female	= 1 if female, 0 if not
Income	= annual income in dollars
Married	= 1 if married, 0 if not
Children	= number of children
Loan	= 1 if customer has a car loan, 0 if not
Mortgage	= 1 if customer has a mortgage, 0 if not

We present two clustering methods using a small sample of data from KTC. We first consider bottom-up **hierarchical clustering** that starts with each observation belonging to its own cluster and then sequentially merges the most similar clusters to create a series of nested clusters. The second method, ***k*-means clustering**, assigns each observation to one of k clusters in a manner such that the observations assigned to the same cluster are as similar as possible. Because both methods depend on how two observations are similar, we first discuss how to measure similarity between observations.

Measuring Similarity Between Observations

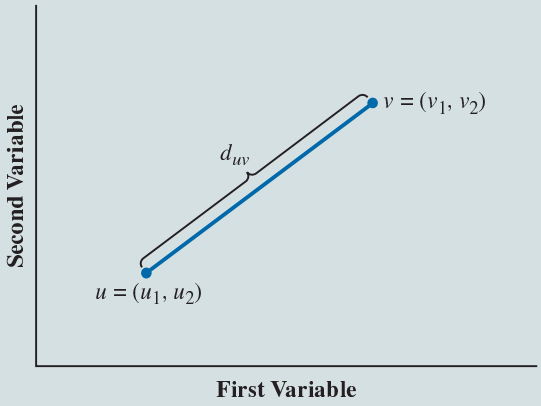
The goal of cluster analysis is to group observations into clusters such that observations within a cluster are similar and observations in different clusters are dissimilar. Therefore, to formalize this process, we need explicit measurements of similarity or, conversely, dissimilarity. Some metrics track similarity between observations, and a clustering method using such a metric would seek to maximize the similarity between observations. Other metrics measure dissimilarity, or distance, between observations, and a clustering method using one of these metrics would seek to minimize the distance between observations in a cluster.

When observations include numerical variables, **Euclidean distance** is the most common method to measure dissimilarity between observations. Let observations $u = (u_1, u_2, \ldots, u_q)$ and $v = (v_1, v_2, \ldots, v_q)$ each comprise measurements of q variables. The Euclidean distance between observations u and v is

$$d_{uv} = \sqrt{(u_1 - v_1)^2 + (u_2 - v_2)^2 + \cdots + (u_q - v_q)^2}$$

Figure 4.1 depicts Euclidean distance for two observations consisting of two variable measurements. Euclidean distance becomes smaller as a pair of observations become more similar with respect to their variable values. Euclidean distance is highly influenced by the scale on which variables are measured. For example, consider the task of clustering customers on the basis of the variables Age and Income. Let observation u = (23, $20,375) correspond to a 23-year old customer with an annual income of $20,375 and observation

FIGURE 4.1 Euclidean Distance



v = (36, $19,475) correspond to a 36-year old with an annual income of $19,475. As measured by Euclidean distance, the dissimilarity between these two observations is

$$d_{uv} = \sqrt{(23 - 36)^2 + (20{,}375 - 19{,}475)^2} = \sqrt{169 + 811{,}441} = 901$$

Refer to Chapter 2 for a discussion of z-scores.

Thus, we see that when using the raw variable values, the amount of dissimilarity between observations is dominated by the Income variable because of the difference in the magnitude of the measurements. Therefore, it is common to standardize the units of each variable j of each observation u. That is, u_j, the value of variable j in observation u, is replaced with its z-score z_j. For the data in *DemoKTC*, the standardized (or normalized) values of observations u and v are (–1.76, –0.56) and (–0.76, –0.62), respectively. The dissimilarity between these two observations based on standardized values is

$$\begin{aligned}(standardized)\ d_{uv} &= \sqrt{(-1.76 - (-0.76))^2 + (-0.56 - (-0.62))^2} \\ &= \sqrt{0.994 + 0.004} = 0.998\end{aligned}$$

Based on standardized variable values, we observe that observations u and v are actually much more different in age than in income.

The conversion to z-scores also makes it easier to identify outlier measurements, which can distort the Euclidean distance between observations. After conversion to z-scores, unequal weighting of variables can also be considered by multiplying the variables of each observation by a selected set of weights. For instance, after standardizing the units on customer observations so that income and age are expressed as their respective z-scores (instead of expressed in dollars and years), we can multiply the income z-scores by 2 if we wish to treat income with twice the importance of age. In other words, standardizing removes bias due to the difference in measurement units, and variable weighting allows the analyst to introduce appropriate bias based on the business context.

When clustering observations solely on the basis of categorical variables encoded as 0–1 (or dummy variables), a better measure of similarity between two observations can be achieved by counting the number of variables with matching values. The simplest overlap measure is called the **matching coefficient** and is computed as follows:

MATCHING COEFFICIENT

$$\frac{\text{number of variables with matching value for observations } u \text{ and } v}{\text{total number of variables}}$$

One weakness of the matching coefficient is that if two observations both have a 0 entry for a categorical variable, this is counted as a sign of similarity between the two observations. However, matching 0 entries do not necessarily imply similarity. For instance, if the categorical variable is Own A Minivan, then a 0 entry in two different observations does not mean that these two people own the same type of car; it means only that neither owns a minivan. To avoid misstating similarity due to the absence of a feature, a similarity measure called **Jaccard's coefficient** does not count matching zero entries and is computed as follows:

JACCARD'S COEFFICIENT

$$\frac{\text{number of variables with matching nonzero value for observations } u \text{ and } v}{(\text{total number of variables}) - (\text{number of variables with matching zero values for observations } u \text{ and } v)}$$

For five customer observations from the file *DemoKTC*, Table 4.1 contains observations of the binary variables Female, Married, Loan, and Mortgage and the distance matrixes corresponding to the matching coefficient and Jaccard's coefficient, respectively. Based on the matching coefficient, Observation 1 and Observation 4 are more similar (0.75) than Observation 2 and Observation 3 (0.5) because 3 out of 4 variable values match between Observation 1 and Observation 4 versus just 2 matching values out of 4 for Observation 2 and Observation 3. However, based on Jaccard's coefficient, Observation 1 and Observation 4 are equally similar (0.5) as Observation 2 and Observation 3 (0.5) as Jaccard's coefficient discards the matching zero values for the Loan and Mortgage variables for Observation 1 and Observation 4. In the context of this example, choice of the matching coefficient or Jaccard's coefficient depends on whether KTC believes that matching 0 entries implies similarity or not. That is, KTC must gauge whether meaningful similarity is implied if a pair of observations are not female, not married, do not have a car loan, or do not have a mortgage.

TABLE 4.1 Comparison of Similarity Matrixes for Observations with Binary Variables

Observation	Female	Married	Loan	Mortgage
1	1	0	0	0
2	0	1	1	1
3	1	1	1	0
4	1	1	0	0
5	1	1	0	0

Similarity Matrix Based on Matching Coefficient

Observation	1	2	3	4	5
1	1				
2	0	1			
3	0.5	0.5	1		
4	0.75	0.25	0.75	1	
5	0.75	0.25	0.75	1	1

Similarity Matrix Based on Jaccard's Coefficient

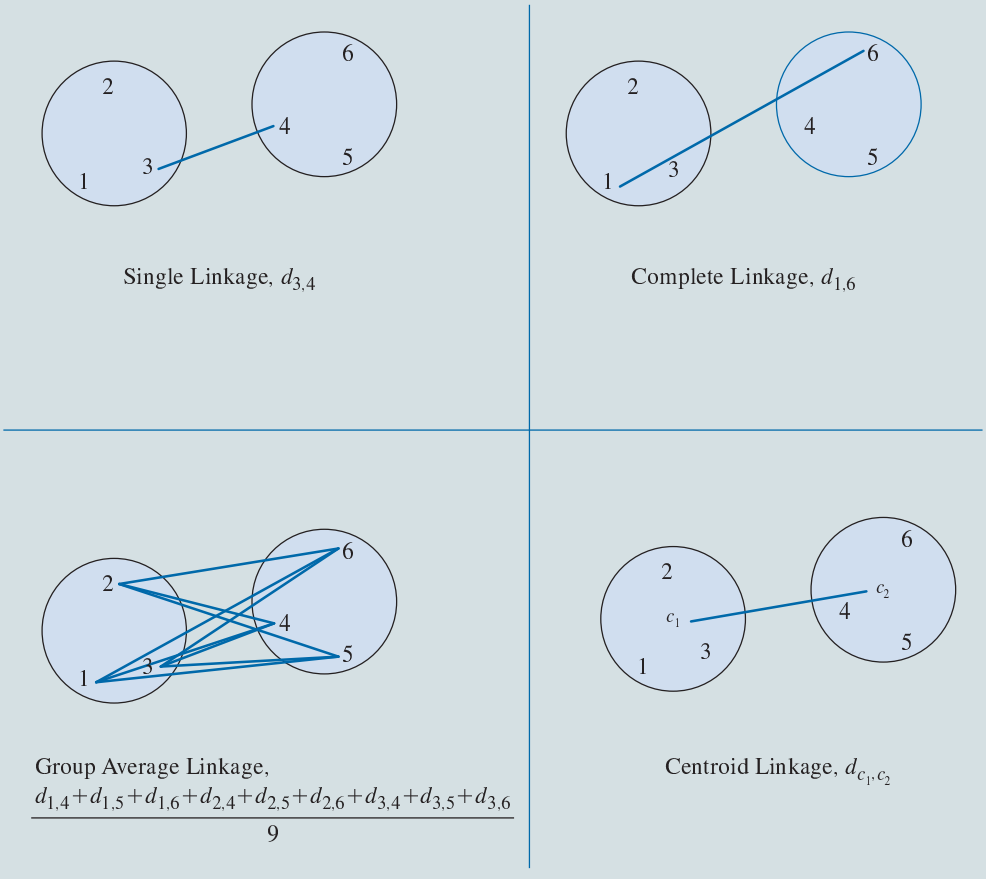
Observation	1	2	3	4	5
1	1				
2	0	1			
3	0.333	0.5	1		
4	0.5	0.25	0.667	1	
5	0.5	0.25	0.667	1	1

Hierarchical Clustering

We consider a bottom-up hierarchical clustering approach that starts with each observation in its own cluster and then iteratively combines the two clusters that are the most similar into a single cluster. Each iteration corresponds to an increased level of aggregation by decreasing the number of distinct clusters. Hierarchical clustering determines the similarity of two clusters by considering the similarity between the observations composing either cluster. Given a way to measure similarity between observations (Euclidean distance, matching coefficients, or Jaccard's coefficients), there are several hierarchical clustering method alternatives for comparing observations in two clusters to obtain a cluster similarity measure. Using Euclidean distance to illustrate, Figure 4.2 provides a two-dimensional depiction of four methods we will discuss.

When using the **single linkage** clustering method, the similarity between two clusters is defined by the similarity of the pair of observations (one from each cluster) that are the most similar. Thus, single linkage will consider two clusters to be close if an observation in one of the clusters is close to at least one observation in the other cluster. However, a cluster formed by merging two clusters that are close with respect to single linkage may also consist of pairs of observations that are very different. The reason is that there is no consideration of how different an observation may be from other observations in a cluster as long as it is similar to at least one observation in that cluster. Thus, in two dimensions (variables), single linkage clustering can result in long, elongated clusters rather than compact, circular clusters.

FIGURE 4.2 Measuring Similarity Between Clusters



The **complete linkage** clustering method defines the similarity between two clusters as the similarity of the pair of observations (one from each cluster) that are the most different. Thus, complete linkage will consider two clusters to be close if their most-different pair of observations are close. This method produces clusters such that all member observations of a cluster are relatively close to each other. The clusters produced by complete linkage have approximately equal diameters. However, clustering created with complete linkage can be distorted by outlier observations.

The single linkage and complete linkage methods define between-cluster similarity based on the single pair of observations in two different clusters that are most similar or least similar. In contrast, the **group average linkage** clustering method defines the similarity between two clusters to be the average similarity computed over *all* pairs of observations between the two clusters. If Cluster 1 consists of n_1 observations and Cluster 2 consists of n_2 observations, the similarity of these clusters would be the average of $n_1 \times n_2$ similarity measures. This method produces clusters that are less dominated by the similarity between single pairs of observations. The **median linkage** method is analogous to group average linkage except that it uses the median of the similarities computed between all pairs of observations between the two clusters. The use of the median reduces the effect of outliers.

Centroid linkage uses the averaging concept of cluster centroids to define between-cluster similarity. The centroid for cluster k, denoted as c_k, is found by calculating the average value for each variable across all observations in a cluster; that is, a centroid is the average observation of a cluster. The similarity between cluster k and cluster j is then defined as the similarity of the centroids c_k and c_j.

Ward's method merges two clusters such that the dissimilarity of the observations within the resulting single cluster increases as little as possible. It tends to produce clearly defined clusters of similar size. For a pair of clusters under consideration for aggregation, Ward's method computes the centroid of the resulting merged cluster and then calculates the sum of squared dissimilarity between this centroid and each observation in the union of the two clusters. Representing observations within a cluster with the centroid can be viewed as a loss of information in the sense that the individual differences in these observations will not be captured by the cluster centroid. Hierarchical clustering using Ward's method results in a sequence of aggregated clusters that minimizes this loss of information between the individual observation level and the cluster centroid level.

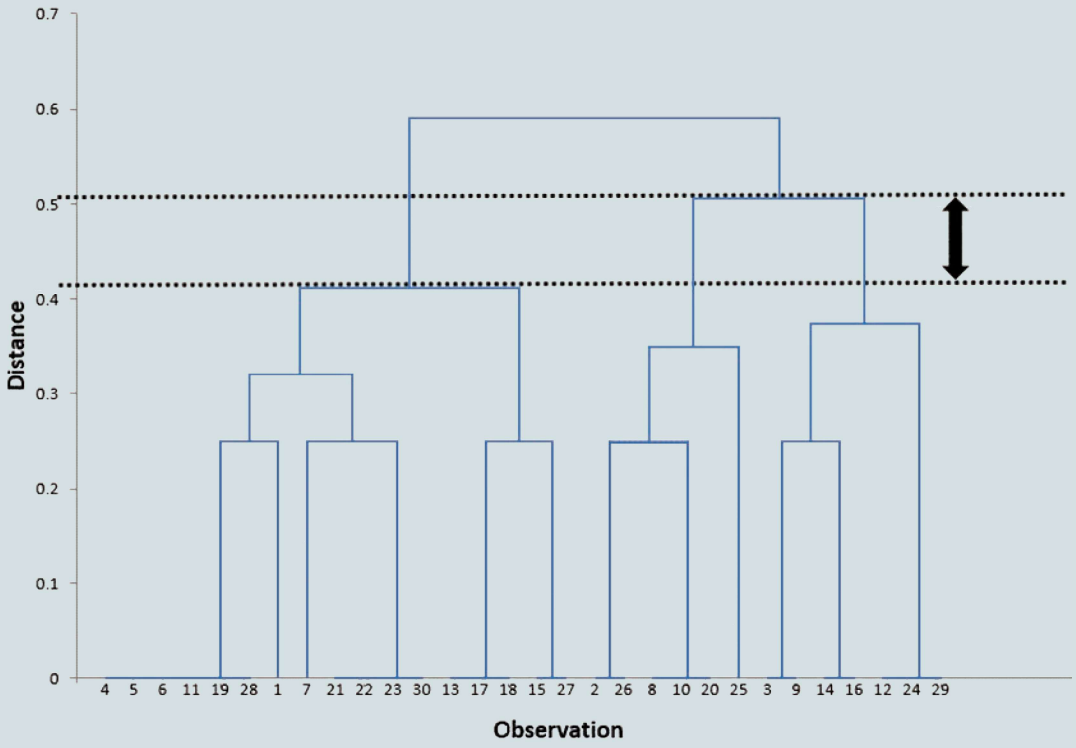
When **McQuitty's method** considers merging two clusters A and B, the dissimilarity of the resulting cluster AB to any other cluster C is calculated as ((dissimilarity between A and C) + (dissimilarity between B and C)) ÷ 2. At each step, this method then merges the pair of clusters that results in the minimal increase in total dissimilarity between the newly merged cluster and all the other clusters.

DATA file
DemoKTC

Returning to our example, KTC is interested in developing customer segments based on gender, marital status, and whether the customer is repaying a car loan and a mortgage. Using data in the file *DemoKTC*, we base the clusters on a collection of 0–1 categorical variables (Female, Married, Loan, and Mortgage). We use the matching coefficient to measure similarity between observations and the group average linkage clustering method to measure similarity between clusters. The choice of the matching coefficient (over Jaccard's coefficent) is reasonable because a pair of customers that both have an entry of zero for any of these four variables implies some degree of similarity. For example, two customers that both have zero entries for Mortgage means that neither has significant debt associated with a mortgage.

Figure 4.3 depicts a **dendrogram** to visually summarize the output from a hierarchical clustering using the matching coefficient to measure similarity between observations and the group average linkage clustering method to measure similarity between clusters. A dendrogram is a chart that depicts the set of nested clusters resulting at each step of aggregation. The horizontal axis of the dendrogram lists the observation indexes. The vertical axis of the dendrogram represents the dissimilarity (distance) resulting from a merger of two different groups of observations. Each blue horizontal line in the dendrogram represents a merger of two (or more) clusters, where the observations composing the merged clusters are connected to the blue horizontal line with a blue vertical line.

FIGURE 4.3 Dendrogram for KTC Using Matching Coefficients and Group Average Linkage



For example, the blue horizontal line connecting observations 4, 5, 6, 11, 19, and 28 conveys that these six observations are grouped together and the resulting cluster has a dissimilarity measure of 0. A dissimilarity of 0 results from this merger because these six observations have identical values for the Female, Married, Loan, and Mortgage variables. In this case, each of these six observations corresponds to a married female with no car loan and no mortgage. Following the blue vertical line up from the cluster of {4, 5, 6, 11, 19, 28}, another blue horizontal line connects this cluster with the cluster consisting solely of Observation 1. Thus, the cluster {4, 5, 6, 11, 19, 28} and cluster {1} are merged resulting in a dissimilarity of 0.25. The dissimilarity of 0.25 results from this merger because Observation 1 differs in one out of the four categorical variable values; Observation 1 is an *unmarried* female with no car loan and no mortgage.

To interpret a dendrogram at a specific level of aggregation, it is helpful to visualize a horizontal line such as one of the black dashed lines we have drawn across Figure 4.3. The bottom horizontal black dashed line intersects with the vertical branches in the dendrogram three times; each intersection corresponds to a cluster containing the observations connected by the vertical branch that is intersected. The composition of these three clusters is as follows:

Cluster 1: {4, 5, 6, 11, 19, 28, 1, 7, 21, 22, 23, 30, 13, 17, 18, 15, 27}
= mix of males and females, 15 out of 17 married, no car loans, 5 out of 17 with mortgages

Cluster 2: {2, 26, 8, 10, 20, 25}
= all males with car loans, 5 out of 6 married, 2 out of 6 with mortgages

Cluster 3: {3, 9, 14, 16, 12, 24, 29}
= all females with car loans, 4 out of 7 married, 5 out of 7 with mortgages

These clusters segment KTC's customers into three groups that could possibly indicate varying levels of responsibility—an important factor to consider when providing financial advice.

The nested construction of the hierarchical clusters allows KTC to identify different numbers of clusters and assess (often qualitatively) the implications. By sliding a horizontal line up or down the vertical axis of a dendrogram and observing the intersection of the horizontal line with the vertical dendrogram branches, an analyst can extract varying numbers of clusters. Note that sliding up to the position of the top horizontal black line in Figure 4.3 results in merging Cluster 2 with Cluster 3 into a single, more dissimilar, cluster. The vertical distance between the points of agglomeration is the "cost" of merging clusters in terms of decreased homogeneity within clusters. Thus, vertically elongated portions of the dendrogram represent mergers of more dissimilar clusters, and vertically compact portions of the dendrogram represent mergers of more similar clusters. A cluster's durability (or strength) can be measured by the difference between the distance value at which a cluster is originally formed and the distance value at which it is merged with another cluster. Figure 4.3 shows that the cluster consisting of {12, 24, 29} (single females with car loans and mortgages) is a very durable cluster in this example because the vertical line for this cluster is very long before it is merged with another cluster.

k-Means Clustering

In *k*-means clustering, the analyst must specify the number of clusters, *k*. If the number of clusters, *k*, is not clearly established by the context of the business problem, the *k*-means clustering algorithm can be repeated for several values of *k*. Given a value of *k*, the *k*-means algorithm randomly assigns each observation to one of the *k* clusters. After all observations have been assigned to a cluster, the resulting cluster centroids are calculated (these cluster centroids are the "means" of *k*-means clustering). Using the updated cluster centroids, all observations are reassigned to the cluster with the closest centroid (where Euclidean distance is the standard metric). The algorithm repeats this process (calculate cluster centroid, assign each observation to the cluster with nearest centroid) until there is no change in the clusters or a specified maximum number of iterations is reached.

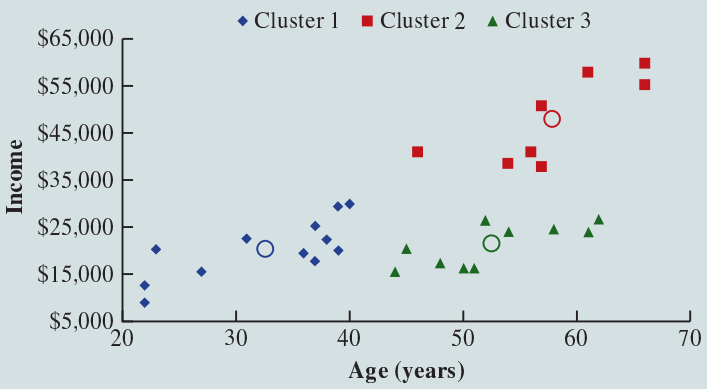
DATA file
DemoKTC

As an unsupervised learning technique, cluster analysis is not guided by any explicit measure of accuracy, and thus the notion of a "good" clustering is subjective and is dependent on what the analyst hopes the cluster analysis will uncover. Regardless, one can measure the strength of a cluster by comparing the average distance in a cluster to the distance between cluster centroids. One rule of thumb is that the ratio of between-cluster distance (as measured by the distance between cluster centroids) to average within-cluster distance should exceed 1.0 for useful clusters.

A wide disparity in cluster strength across a set of clusters may make it possible to find a better clustering of the data by removing all members of the strong clusters and then continuing the clustering process on the remaining observations.

To illustrate *k*-means clustering, we consider a 3-means clustering of a small sample of KTC's customer data in the file *DemoKTC*. Figure 4.4 shows three clusters based on

FIGURE 4.4 Clustering Observations by Age and Income Using *k*-Means Clustering with $k = 3$



Cluster centroids are depicted by circles in Figure 4.4.

Although Figure 4.4 is plotted in the original scale of the variables, the clustering was based on the variables after standardizing (normalizing) their values.

Tables 4.2 and 4.3 are expressed in terms of standardized coordinates in order to eliminate any distortion resulting from differences in the scale of the input variables.

TABLE 4.2 Average Distances Within Clusters

	No. of Observations	Average Distance Between Observations in Cluster
Cluster 1	12	0.622
Cluster 2	8	0.739
Cluster 3	10	0.520

TABLE 4.3 Distances Between Cluster Centroids

	Cluster 1	Cluster 2	Cluster 3
Cluster 1	0	2.784	1.529
Cluster 2	2.784	0	1.964
Cluster 3	1.529	1.964	0

customer income and age. Cluster 1 is characterized by relatively younger, lower-income customers (Cluster 1's centroid is at [33, $20,364]). Cluster 2 is characterized by relatively older, higher-income customers (Cluster 2's centroid is at [58, $47,729]). Cluster 3 is characterized by relatively older, lower-income customers (Cluster 3's centroid is at [53, $21,416]). As visually corroborated by Figure 4.4, Table 4.2 shows that Cluster 2 is the smallest, but most heterogeneous cluster. We also observe that Cluster 1 is the largest cluster and Cluster 3 is the most homogeneous cluster. Table 4.3 displays the distance between each pair of cluster centroids to demonstrate how distinct the clusters are from each other. Cluster 1 and Cluster 2 are the most distinct from each other. To evaluate the strength of the clusters, we compare the average distance within each cluster (Table 4.2) to the average distances between clusters (Table 4.3). For example, although Cluster 2 is the most heterogeneous, with an average distance between observations of 0.739, comparing this to the distance between the Cluster 2 and Cluster 3 centroids (1.964) reveals that on average an observation in Cluster 2 is approximately 2.66 times closer to the Cluster 2 centroid than to the Cluster 3 centroid. In general, the larger the ratio of the distance between a pair of cluster centroids and the average within-cluster distance, the more distinct the clustering is for the observations in the two clusters in the pair. Although qualitative considerations should take priority in evaluating clusters, using the ratios of between-cluster distance and average within-cluster distance provides some guidance in determining k, the number of clusters.

Hierarchical Clustering versus *k*-Means Clustering

If you have a small data set (e.g., fewer than 500 observations) and want to easily examine solutions with increasing numbers of clusters, you may want to use hierarchical clustering. Hierarchical clusters are also convenient if you want to observe how clusters are nested. However, hierarchical clustering can be very sensitive to outliers, and clusters may change dramatically if observations are eliminated from (or added to) the data set. If you know how many clusters you want and you have a larger data set (e.g., more than 500 observations), you may choose to use k-means clustering. Recall that k-means clustering partitions the observations, which is appropriate if you are trying to summarize the data with k "average" observations that describe the data with the minimum amount of error. However, k-means clustering is generally not appropriate for binary or ordinal data, for which an "average" is not meaningful.

NOTES + COMMENTS

Clustering observations based on both numerical and categorical variables (mixed data) can be challenging. Dissimilarity between observations with numerical variables is commonly computed using Euclidean distance. However, Euclidean distance is not well defined for categorical variables as the magnitude of the Euclidean distance measure between two category values will depend on the numerical encoding of the categories. There are elaborate methods beyond the scope of this book to try to address the challenge of clustering mixed data.

Using the methods introduced in this section, there are two alternative approaches to clustering mixed data. The first approach is to decompose the clustering into two steps. The first step applies hierarchical clustering of the observations only on categorical variables using an appropriate measure (matching coefficients or Jaccard's coefficients) to identify a set of "first-step" clusters. The second step is to apply *k*-means clustering (or hierarchical clustering again) separately to each of these "first-step" clusters using only the numerical variables. This decomposition approach is not fail-safe as it fixes clusters with respect to one variable type before clustering with respect to the other variable type, but it does allow the analyst to identify how the observations are similar or different with respect to the two variable types.

A second approach to clustering mixed data is to numerically encode the categorical values (e.g., binary coding, ordinal coding) and then to standardize both the categorical and numerical variable values. To reflect relative importance of the variables, the analyst may experiment with various weightings of the variables and apply hierarchical or k-means clustering. This approach is very experimental and the variable weights are subjective.

4.2 Association Rules

In marketing, analyzing consumer behavior can lead to insights regarding the placement and promotion of products. Specifically, marketers are interested in examining transaction data on customer purchases to identify the products commonly purchased together. Bar-code scanners facilitate the collection of retail transaction data, and membership in a customer's loyalty program can further associate the transaction with a specific customer. In this section, we discuss the development of probabilistic if–then statements, called **association rules**, which convey the likelihood of certain items being purchased together. Although association rules are an important tool in **market basket analysis**, they are also applicable to disciplines other than marketing. For example, association rules can assist medical researchers in understanding which treatments have been commonly prescribed to certain patient symptoms (and the resulting effects).

Hy-Vee grocery store would like to gain insight into its customers' purchase patterns to possibly improve its in-aisle product placement and cross-product promotions. Table 4.4 contains a small sample of data in which each transaction comprises the items purchased by a shopper in a single visit to a Hy-Vee. An example of an association rule from this data would be "if {bread, jelly}, then {peanut butter}," meaning that "if a transaction includes bread and jelly, then it also includes peanut butter." The collection of items (or item set) corresponding to the *if* portion of the rule, {bread, jelly}, is called the **antecedent**. The item set corresponding to the *then* portion of the rule, {peanut butter}, is called the **consequent**.

Typically, only association rules for which the consequent consists of a single item are considered because these are more actionable. Although the number of possible association rules can be overwhelming, we typically investigate only association rules that involve antecedent and consequent item sets that occur together frequently. To formalize the notion of "frequent," we define the **support count** of an item set as the number of transactions in the data that include that item set. In Table 4.4, the support count of {bread, jelly} is 4. The potential impact of an association rule is often governed by the number of transactions it may affect, which is measured by computing the support count of the item set consisting of the union of its antecedent and consequent. Investigating the rule "if {bread, jelly}, then {peanut butter}" from Table 4.4, we see the support count of {bread, jelly, peanut butter} is 2. By only considering rules involving item sets with a support above a minimum level, inexplicable rules capturing random noise in the data can generally be avoided. A rule of thumb is to consider only association rules with a support count of at least 20% of the total

Support is also sometimes expressed as the percentage of total transactions containing an item set.

The data in Table 4.4 are in item list format; that is, each transaction row corresponds to a list of item names. Alternatively, the data can be represented in binary matrix format, in which each row is a transaction record and the columns correspond to each distinct item. A third approach is to store the data in stacked form in which each row is an ordered pair; the first entry is the transaction number and the second entry is the item.

TABLE 4.4 Shopping-Cart Transactions

Transaction	Shopping Cart
1	bread, peanut butter, milk, fruit, jelly
2	bread, jelly, soda, potato chips, milk, fruit, vegetables, peanut butter
3	whipped cream, fruit, chocolate sauce, beer
4	steak, jelly, soda, potato chips, bread, fruit
5	jelly, soda, peanut butter, milk, fruit
6	jelly, soda, potato chips, milk, bread, fruit
7	fruit, soda, potato chips, milk
8	fruit, soda, peanut butter, milk
9	fruit, cheese, yogurt
10	yogurt, vegetables, beer

number of transactions. If an item set is particularly valuable and represents a lucrative opportunity, then the minimum support count used to filter the rules is often lowered.

To help identify reliable association rules, we define the measure of **confidence** of a rule, which is computed as

CONFIDENCE

$$\frac{\text{support of \{antecedent and consequent\}}}{\text{support of antecedent}}$$

Conditional probability is discussed in more detail in Chapter 5.

This measure of confidence can be viewed as the conditional probability of the consequent item set occurring given that the antecedent item set occurs. A high value of confidence suggests a rule in which the consequent is frequently true when the antecedent is true, but a high value of confidence can be misleading. For example, if the support of the consequent is high—that is, the item set corresponding to the *then* part is very frequent—then the confidence of the association rule could be high even if there is little or no association between the items. In Table 4.4, the rule "if {cheese}, then {fruit}" has a confidence of 1.0 (or 100%). This is misleading because {fruit} is a frequent item; the confidence of *almost any* rule with {fruit} as the consequent will have high confidence. Therefore, to evaluate the efficiency of a rule, we compute the **lift ratio** of the rule by accounting for the frequency of the consequent:

LIFT RATIO

$$\frac{\text{confidence}}{\text{support of consequent/total number of transactions}}$$

Recall that confidence, the numerator of the lift ratio, can be thought of as the probability of the consequent item set given the antecedent item set occurs. The denominator of the lift ratio is the probability of a randomly selected transaction containing the consequent set. Thus, the lift ratio represents how effective an association rule is at identifying transactions in which the consequent item set occurs versus a randomly selected transaction. A lift ratio greater than one suggests that there is some usefulness to the rule and that it is better at identifying cases when the consequent occurs than having no rule at all. In other words, a lift ratio greater than one suggests that the level of association between the antecedent and consequent is higher than would be expected if these item sets were independent.

For the data in Table 4.4, the rule "if {bread, jelly}, then {peanut butter}" has confidence = 2/4 = 0.5 and lift ratio = 0.5/(4/10) = 1.25. In other words, identifying a customer who purchased both bread and jelly as one who also purchased peanut butter is 25% better than just guessing that a random customer purchased peanut butter.

The utility of a rule depends on both its support and its lift ratio. Although a high lift ratio suggests that the rule is very efficient at finding when the consequent occurs, if it has a very low support, the rule may not be as useful as another rule that has a lower lift ratio but affects a large number of transactions (as demonstrated by a high support). However, an association rule with a high lift ratio and low support may still be useful if the consequent represents a very valuable opportunity.

DATA file

HyVeeDemoBinary
HyVeeDemoStacked

Based on the data in Table 4.4, Table 4.5 shows the list of association rules that achieve a lift ratio of at least 1.39 while satisfying a minimum support of 4 transactions (out of 10) and a minimum confidence of 50%. The top rules in Table 4.5 suggest that bread, fruit, and jelly are commonly associated items. For example, the fourth rule listed in Table 4.5 states, "If Fruit and Jelly are purchased, then Bread is also purchased." Perhaps Hy-Vee could consider a promotion and/or product placement to leverage this perceived relationship.

Evaluating Association Rules

Although explicit measures such as support, confidence, and lift ratio can help filter association rules, an association rule is ultimately judged on how actionable it is and how well

TABLE 4.5 Association Rules for Hy-Vee

Antecedent (A)	Consequent (C)	Support for A	Support for C	Support for A & C	Confidence (%)	Lift Ratio
Bread	Fruit, Jelly	4	5	4	100.0	2.00
Bread	Jelly	4	5	4	100.0	2.00
Bread, Fruit	Jelly	4	5	4	100.0	2.00
Fruit, Jelly	Bread	5	4	4	80.0	2.00
Jelly	Bread	5	4	4	80.0	2.00
Jelly	Bread, Fruit	5	4	4	80.0	2.00
Fruit, Potato Chips	Soda	4	6	4	100.0	1.67
Peanut Butter	Milk	4	4	6	100.0	1.67
Peanut Butter	Milk, Fruit	4	6	4	100.0	1.67
Peanut Butter, Fruit	Milk	4	6	4	100.0	1.67
Potato Chips	Fruit, Soda	4	6	4	100.0	1.67
Potato Chips	Soda	4	6	4	100.0	1.67
Fruit, Soda	Potato Chips	6	4	4	66.7	1.67
Milk	Peanut Butter	6	4	4	66.7	1.67
Milk	Peanut Butter, Fruit	6	4	4	66.7	1.67
Milk, Fruit	Peanut Butter	6	4	4	66.7	1.67
Soda	Fruit, Potato Chips	6	4	4	66.7	1.67
Soda	Potato Chips	6	4	4	66.7	1.67
Fruit, Soda	Milk	6	6	5	83.3	1.39
Milk	Fruit, Soda	6	6	5	83.3	1.39
Milk	Soda	6	6	5	83.3	1.39
Milk, Fruit	Soda	6	6	5	83.3	1.39
Soda	Milk	6	6	5	83.3	1.39
Soda	Milk, Fruit	6	6	5	83.3	1.39

it explains the relationship between item sets. For example, suppose Walmart mined its transactional data to uncover strong evidence of the association rule, "If a customer purchases a Barbie doll, then a customer also purchases a candy bar." Walmart could leverage this relationship in product placement decisions as well as in advertisements and promotions, perhaps by placing a high-margin candy-bar display near the Barbie dolls. However, we must be aware that association rule analysis often results in obvious relationships such as "If a customer purchases hamburger patties, then a customer also purchases hamburger buns," which may be true but provide no new insight. Association rules with a weak support measure often are inexplicable. For an association rule to be useful, it must be well supported *and* explain an important previously unknown relationship. The support of an association rule can generally be improved by basing it on less specific antecedent and consequent item sets. Unfortunately, association rules based on less specific item sets tend to yield less insight. Adjusting the data by aggregating items into more general categories (or splitting items into more specific categories) so that items occur in roughly the same number of transactions often yields better association rules.

4.3 Text Mining

Every day, nearly 500 million tweets are published on the on-line social network service Twitter. Many of these tweets contain important clues about how Twitter users value a company's products and services. Some tweets might sing the praises of a product; others might complain about low-quality service. Furthermore, Twitter users vary greatly in the number of followers (some have thousands of followers and others just a few) and therefore these users have varying degrees of influence. Data-savvy companies can use social media data to improve their products and services. On-line reviews on web sites such as Amazon and Yelp provide data on how customers feel about products and services.

However, the data in these examples are not numerical. The data are text: words, phrases, sentences, and paragraphs. Text, like numerical data, may contain information that can help solve problems and lead to better decisions. **Text mining** is the process of extracting useful information from text data. In this section, we discuss text mining, how it is different from data mining of numerical data, and how it can be useful for decision making.

Text data is often referred to as **unstructured data** because in its raw form, it cannot be stored in a traditional structured database (rows and columns). Audio and video data are also examples of unstructured data. Data mining with text data is more challenging than data mining with traditional numerical data, because it requires more preprocessing to convert the text to a format amenable for analysis. However, once the text data has been converted to numerical data, the analytical methods used for descriptive text mining are the same as those used for numerical data discussed earlier in this chapter. We begin with a small example which illustrates how text data can be converted to numerical data and then analyzed. Then we will provide more in-depth discussion of text-mining concepts and preprocessing procedures.

Voice of the Customer at Triad Airline

Triad Airlines is a regional commuter airline. Through its voice of the customer program, Triad solicits feedback from its customers through a follow-up e-mail the day after the customer has completed a flight. The e-mail survey asks the customer to rate various aspects of the flight and asks the respondent to type comments into a dialog box in the e-mail.

In addition to the quantitative feedback from the ratings, the comments entered by the respondents need to be analyzed so that Triad can better understand its customers' specific concerns and respond in an appropriate manner. We will use a small training sample of these concerns to illustrate how descriptive text mining can be used in this business context. In general, a collection of text documents to be analyzed is called a **corpus**. In the Triad Airline example, our corpus consists of 10 documents, where each document contains concerns made by a customer.

DATA file
Triad

TABLE 4.6 Ten Respondents' Concerns for Triad Airlines

Concerns
The wi-fi service was horrible. It was slow and cut off several times.
My seat was uncomfortable.
My flight was delayed 2 hours for no apparent reason.
My seat would not recline.
The man at the ticket counter was rude. Service was horrible.
The flight attendant was rude. Service was bad.
My flight was delayed with no explanation.
My drink spilled when the guy in front of me reclined his seat.
My flight was canceled.
The arm rest of my seat was nasty.

Triad's management would like to categorize these customer concerns into groups whose members share similar characteristics so that a solution team can be assigned to each group of concerns.

To be analyzed, text data needs to be converted to structured data (rows and columns of numerical data) so that the tools of descriptive statistics, data visualization and data mining can be applied. We can think of converting a group of documents into a matrix of rows and columns where the rows correspond to a document and the columns correspond to a particular word. In Triad's case, a document is a single respondent's comment. A **presence/absence or binary term-document matrix** is a matrix with the rows representing documents and the columns representing words, and the entries in the columns indicating either the presence or the absence of a particular word in a particular document (1 = present and 0 = not present).

Creating the list of terms to use in the presence/absence matrix can be a complicated matter. Too many terms results in a matrix with many columns, which may be difficult to manage and could yield meaningless results. Too few terms may miss important relationships. Often, term frequency along with the problem context are used as a guide. We discuss this in more detail in the next section. In Triad's case, management used word frequency and the context of having a goal of satisfied customers to come up with the following list of terms they feel are relevant for categorizing the respondent's comments: delayed, flight, horrible, recline, rude, seat, and service.

As shown in Table 4.7, these seven terms correspond to the columns of the presence/absence term-document matrix and the rows correspond to the 10 documents. Each matrix entry indicates whether or not a column's term appears in the document corresponding to the row. For example, a one entry in the first row and third column means that the term "horrible" appears in Document 1. A zero entry in the third row and fourth column means that the term "recline" does not appear in Document 3.

Having converted the text to numerical data, we can apply clustering. In this case, because we have binary presence-absence data, we apply hierarchical clustering. Observing that the absence of a term in two different documents does not imply similarity between the documents, we select Jaccard's coefficient as the similarity measure. To measure similarity between clusters, we use complete linkage. At the level of three clusters, hierarchical clustering results in the following groups of documents:

Cluster 1: {1, 5, 6} = documents discussing service issues
Cluster 2: {2, 4, 8, 10} = documents discussing seat issues
Cluster 3: {3, 7, 9} = documents discussing schedule issues

With these three clusters defined, management can assign an expert team to each of these clusters to directly address the concerns of its customers.

TABLE 4.7 The Presence/Absence Term-Document Matrix for Triad Airlines

	Term						
Document	Delayed	Flight	Horrible	Recline	Rude	Seat	Service
1	0	0	1	0	0	0	1
2	0	0	0	0	0	1	0
3	1	1	0	0	0	0	0
4	0	0	0	1	0	1	0
5	0	0	1	0	1	0	1
6	0	1	0	0	1	0	1
7	1	1	0	0	0	0	0
8	0	0	0	1	0	1	0
9	0	1	0	0	0	0	0
10	0	0	0	0	0	1	0

Preprocessing Text Data for Analysis

In general, the text-mining process converts unstructured text into numerical data and applies quantitative techniques. For the Triad example, we converted the text documents into a term-document matrix and then applied hierarchical clustering to gain insight on the different types of comments (and their frequencies). In this section, we present a more detailed discussion of terminology and methods used in preprocessing text data into numerical data for analysis.

Converting documents to a term-document matrix is not a simple task. Obviously, which terms become the headers of the columns of the term-document matrix can greatly impact the analysis. **Tokenization** is the process of dividing text into separate terms, referred to as tokens. The process of identifying tokens is not straightforward. First, symbols and punctuations must be removed from the document and all letters should be converted to lowercase. For example, "Awesome!", "awesome," and "#Awesome" should all be converted to "awesome." Likewise, different forms of the same word, such as "stacking", "stacked," and "stack" probably should not be considered as distinct terms. **Stemming**, the process of converting a word to its stem or root word, would drop the "ing" and "ed" and place only "stack" in the list of words to be tracked.

The goal of preprocessing is to generate a list of most-relevant terms that is sufficiently small so as to lend itself to analysis. In addition to stemming, frequency can be used to eliminate words from consideration as tokens. For example, if a term occurs very frequently in every document in the corpus, then it probably will not be very useful and can be eliminated from consideration; "the" is an example of frequent, uninformative term. Similarly, low-frequency words probably will not be very useful as tokens. Another technique for reducing the consideration set for tokens is to consolidate a set of words that are synonyms. For example, "courteous," "cordial," and "polite" might be best represented as a single token, "polite."

In addition to automated stemming and text reduction via frequency and synonyms, most text-mining software gives the user the ability to manually specify terms to include or exclude as tokens. Also, the use of slang, humor, and sarcasm can cause interpretation problems and might require more sophisticated data cleansing and subjective intervention on the part of the analyst to avoid misinterpretation.

Data preprocessing parses the original text data down to the set of tokens deemed relevant for the topic being studied. Based on these tokens, a presence/absence term-document matrix as in Table 4.7 can be generated.

When the documents in a corpus contain many more words than the brief comments in the Triad Airline example, and when the *frequency* of word occurrence is important to the context

of the business problem, preprocessing can be used to develop a **frequency term-document matrix**. A frequency term-document matrix is a matrix whose rows represent documents and columns represent tokens, and the entries in the matrix are the frequency of occurrence of each token in each document. We illustrate this in the following example.

Movie Reviews

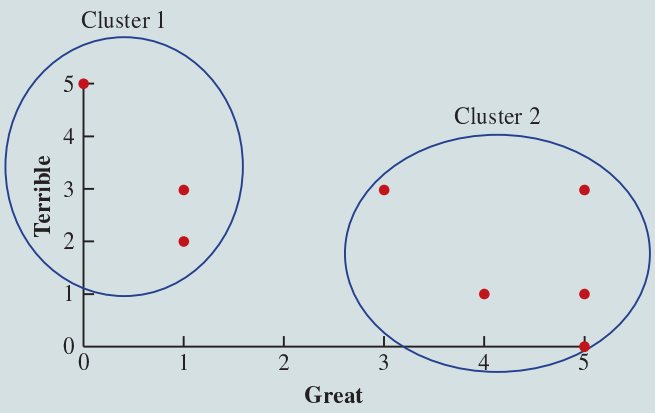
A new action film has been released and we now have a sample of 10 reviews from movie critics. Using preprocessing techniques, including text reduction by synonyms, we have reduced the number of tokens to only two: "great" and "terrible." Table 4.8 displays the corresponding frequency term-document matrix. As Table 4.8 shows, the token "great" appears four times in Document 7. Reviewing the entire table, we observe that five is the maximum frequency of a token in a document and zero is the minimum frequency.

To demonstrate the analysis of a frequency term-document matrix with descriptive data mining, we apply k-means clustering with $k = 2$ to the frequency term-document matrix to obtain the two clusters in Figure 4.5. Cluster 1 contains reviews that tend to be negative and Cluster 2 contains reviews that tend to be positive. We note that the Observation (3, 3) corresponds to the balanced review of Document 4; based on this small corpus, the balanced review is more similar to the positive reviews than the negative reviews, suggesting that the negative reviews may tend to be more extreme.

TABLE 4.8 The Frequency Term-Document Matrix for Movie Reviews

	Term	
Document	Great	Terrible
1	5	0
2	5	1
3	5	1
4	3	3
5	5	1
6	0	5
7	4	1
8	5	3
9	1	3
10	1	2

FIGURE 4.5 Two Clusters Using k-Means Clustering on Movie Reviews



NOTES + COMMENTS

1. The term-document matrix is also sometimes referred to as a document-term matrix.
2. In addition to the binary term-document matrix and frequency term-document matrix, there are more complex types of term-document matrices that can be used to preprocess unstructured text data. These methods utilize frequency measures other than simple counts, and include logarithmic-scaled frequency, inverse document frequency, and term frequency-inverse document frequency (TF-IDF), which is the term frequency multiplied by the inverse of the document frequency.
3. The process of converting words to all lowercase is often referred to as term normalization.
4. The process of clustering/categorizing comments or reviews as positive, negative, or neutral is known as sentiment analysis.

SUMMARY

We have introduced the descriptive data-mining methods and related concepts. After introducing how to measure the similarity of individual observations, we presented two different methods for grouping observations based on the similarity of their respective variable values: hierarchical clustering and *k*-means clustering. Hierarchical clustering begins with each observation in its own cluster and iteratively aggregates clusters using a specified linkage method. We described several of these hierarchical clustering methods and discussed their features. In *k*-means clustering, the analyst specifies *k*, the number of clusters, and then observations are placed into these clusters in an attempt to minimize the dissimilarity within the clusters. We concluded our discussion of clustering with a comparison of hierarchical clustering and *k*-means clustering.

We introduced association rules and explained their use for identifying patterns across transactions, particularly in retail data. We defined the concepts of support count, confidence, and lift ratio, and described their utility in gleaning actionable insight from association rules.

Finally, we discussed the text-mining process. Text is first preprocessed by deriving a smaller set of tokens from the larger set of words contained in a collection of documents. Then the tokenized text data is converted into a presence/absence term-document matrix or a frequency term-document matrix. We then demonstrated the application of hierarchical clustering on a binary term-document matrix and *k*-means clustering on a frequency term-document matrix to glean insight from the underlying text data.

GLOSSARY

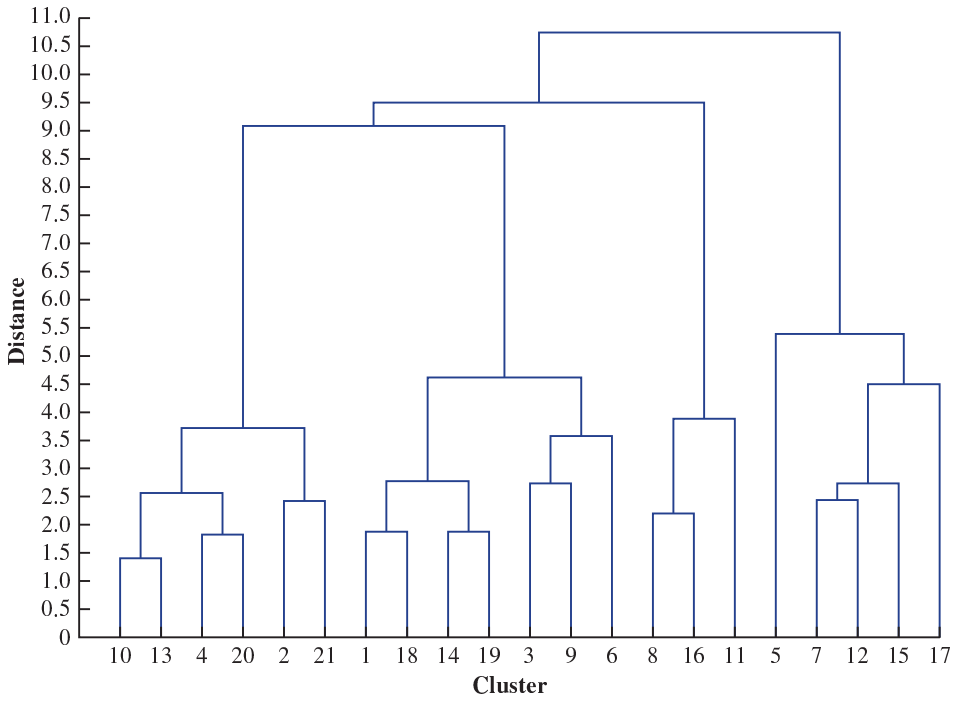
Antecedent The item set corresponding to the *if* portion of an if–then association rule.
Association rule An if–then statement describing the relationship between item sets.
Binary term-document matrix A matrix with the rows representing documents and the columns representing words, and the entries in the columns indicating either the presence or absence of a particular word in a particular document (1 = present and 0 = not present).
Centroid linkage Method of calculating dissimilarity between clusters by considering the two centroids of the respective clusters.
Complete linkage Measure of calculating dissimilarity between clusters by considering only the two most dissimilar observations between the two clusters.
Confidence The conditional probability that the consequent of an association rule occurs given the antecedent occurs.
Consequent The item set corresponding to the *then* portion of an if–then association rule.
Corpus A collection of documents to be analyzed.
Dendrogram A tree diagram used to illustrate the sequence of nested clusters produced by hierarchical clustering.

Euclidean distance Geometric measure of dissimilarity between observations based on the Pythagorean theorem.
Frequency term-document matrix A matrix whose rows represent documents and columns represent tokens (terms), and the entries in the matrix are the frequency of occurrence of each token (term) in each document.
Group average linkage Measure of calculating dissimilarity between clusters by considering the distance between each pair of observations between two clusters.
Hierarchical clustering Process of agglomerating observations into a series of nested groups based on a measure of similarity.
Jaccard's coefficient Measure of similarity between observations consisting solely of binary categorical variables that considers only matches of nonzero entries.
***k*-means clustering** Process of organizing observations into one of k groups based on a measure of similarity (typically Euclidean distance).
Lift ratio The ratio of the performance of a data mining model measured against the performance of a random choice. In the context of association rules, the lift ratio is the ratio of the probability of the consequent occurring in a transaction that satisfies the antecedent versus the probability that the consequent occurs in a randomly selected transaction.
Market basket analysis Analysis of items frequently co-occurring in transactions (such as purchases).
Market segmentation The partitioning of customers into groups that share common characteristics so that a business may target customers within a group with a tailored marketing strategy.
Matching coefficient Measure of similarity between observations based on the number of matching values of categorical variables.
McQuitty's method Measure that computes the dissimilarity introduced by merging clusters A and B by, for each other cluster C, averaging the distance between A and C and the distance between B and C and the summing these average distances.
Median linkage Method that computes the similarity between two clusters as the median of the similarities between each pair of observations in the two clusters.
Observation (record) A set of observed values of variables associated with a single entity, often displayed as a row in a spreadsheet or database.
Presence /absence document-term matrix A matrix with the rows representing documents and the columns representing words, and the entries in the columns indicating either the presence or the absence of a particular word in a particular document (1 = present and 0 = not present).
Single linkage Measure of calculating dissimilarity between clusters by considering only the two most similar observations between the two clusters.
Stemming The process of converting a word to its stem or root word.
Support count The number of times that a collection of items occurs together in a transaction data set.
Text mining The process of extracting useful information from text data.
Tokenization The process of dividing text into separate terms, referred to as tokens.
Unsupervised learning Category of data-mining techniques in which an algorithm explains relationships without an outcome variable to guide the process.
Unstructured data Data, such as text, audio, or video, that cannot be stored in a traditional structured database.
Ward's method Procedure that partitions observations in a manner to obtain clusters with the least amount of information loss due to the aggregation.

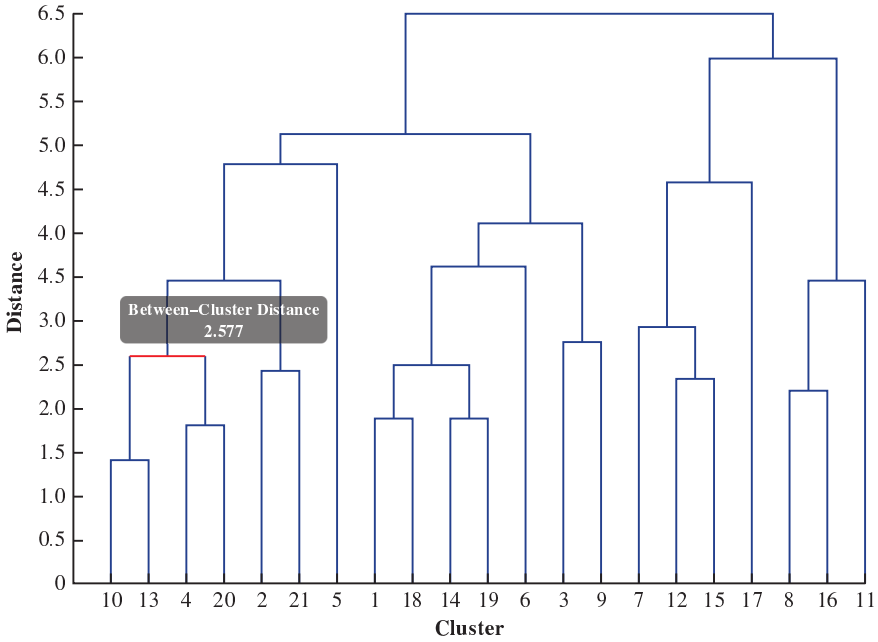
PROBLEMS

1. The regulation of electric and gas utilities is an important public policy question affecting consumer's choice and cost of energy provider. To inform deliberation on public policy, data on eight numerical variables have been collected for a group of energy

companies. To summarize the data, hierarchical clustering has been executed using Euclidean distance as the similarity measure and Ward's method as the clustering method. Based on the following dendrogram, what is the most appropriate number of clusters to organize these utility companies?



2. In an effort to inform political leaders and economists discussing the deregulation of electric and gas utilities, data on eight numerical variables from utility companies have been grouped using hierarchical clustering based on Euclidean distance as the similarity measure and complete linkage as the clustering method.
   a. Based on the following dendrogram, what is the most appropriate number of clusters to organize these utility companies?



b. Using the following data on the Observations 10, 13, 4, and 20, confirm that the complete linkage distance between the cluster containing {10, 13} and the cluster containing {4, 20} is 2.577 units as displayed in the dendrogram.

	Observation			
	10	**13**	**4**	**20**
Income/Debt	0.032	0.195	−0.510	0.466
Return	0.741	0.875	0.207	0.474
Cost	0.700	0.748	−0.004	−0.490
Load	−0.892	−0.735	−0.219	0.655
Peak	−0.173	1.013	−0.943	0.083
Sales	−0.693	−0.489	−0.702	−0.458
PercentNuclear	1.620	2.275	1.328	1.733
TotalFuelCosts	−0.863	−1.035	−0.724	−0.721

3. Amanda Boleyn, an entrepreneur who recently sold her start-up for a multi-million-dollar sum, is looking for alternate investments for her newfound fortune. She is considering an investment in wine, similar to how some people invest in rare coins and fine art. To educate herself on the properties of fine wine, she has collected data on 13 different characteristics of 178 wines. Amanda has applied *k*-means clustering to this data for $k = 2$, 3, and 4 and provided the summaries for each set of resulting clusters. Which value of k is the most appropriate to categorize these wines? Justify your choice with calculations.

Inter-Cluster Distances

	Cluster 1	Cluster 2
Cluster 1	0	3.829
Cluster 2	3.829	0

Within-Cluster Summary

	Size	Average Distance
Cluster 1	94	3.080
Cluster 2	84	2.746
Total	178	2.922

Inter-Cluster Distances

	Cluster 1	Cluster 2	Cluster 3
Cluster 1	0	5.005	3.576
Cluster 2	5.005	0	3.951
Cluster 3	3.576	3.951	0

Within-Cluster Summary

	Size	Average Distance
Cluster 1	63	2.357
Cluster 2	51	2.438
Cluster 3	64	2.765
Total	178	2.527

Inter-Cluster Distances

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Cluster 1	0	2.991	2.576	4.785
Cluster 2	2.991	0	3.951	5.105
Cluster 3	2.576	3.951	0	3.808
Cluster 4	4.785	5.105	3.808	0

Within-Cluster Summary

	Size	Average Distance
Cluster 1	21	2.738
Cluster 2	55	2.285
Cluster 3	51	2.559
Cluster 4	51	2.438
Total	178	2.461

4. Jay Gatsby categorizes wines into one of three clusters. The centroids of these clusters, describing the average characteristics of a wine in each cluster, are listed in the following table.

Characteristic	Cluster 1	Cluster 2	Cluster 3
Alcohol	0.819	0.164	−0.937
MalicAcid	−0.329	0.869	−0.368
Ash	0.248	0.186	−0.393
Alcalinity	−0.677	0.523	0.249
Magnesium	0.643	−0.075	−0.573
Phenols	0.825	0.977	−0.034
Flavanoids	0.896	−1.212	0.083
Nonflavanoids	−0.595	0.724	0.009
Proanthocyanins	0.619	−0.778	0.010
ColorIntensity	0.135	0.939	−0.881
Hue	0.497	−1.162	0.437
Dilution	0.744	−1.289	0.295
Proline	1.117	−0.406	−0.776

Jay has recently discovered a new wine from the Piedmont region of Italy with the following characteristics. In which cluster of wines should he place this new wine? Justify your choice with appropriate calculations.

Characteristic	
Alcohol	−1.023
MalicAcid	−0.480
Ash	0.049
Alcalinity	0.600
Magnesium	−1.242
Phenols	1.094
Flavanoids	0.001
Nonflavanoids	0.548
Proanthocyanins	−0.229
ColorIntensity	−0.797
Hue	0.711
Dilution	−0.425
Proline	0.010

5. Leggere, an internet book retailer, is interested in better understanding the purchase decisions of its customers. For a set of 2,000 customer transactions, it has categorized the individual book purchases comprising those transactions into one or more of the following categories: Novels, Willa Bean series, Cooking Books, Bob Villa Do-It-Yourself, Youth Fantasy, Art Books, Biography, Cooking Books by Mossimo Bottura, Harry Potter series, Florence Art Books, and Titian Art Books. Leggere has conducted association rules analysis on this data set and would like to analyze the output. Based on a minimum support of 200 transactions and a minimum confidence of 50%, the table below shows the top 10 rules with respect to lift ratio.
   a. Explain why the top rule "If customer buys a Bottura cooking book, then they buy a cooking book," is not helpful even though it has the largest lift and 100% confidence.
   b. Explain how the confidence of 52.99% and lift ratio of 2.20 was computed for the rule "If a customer buys a cooking book and a biography book, then they buy an art book." Interpret these quantities.
   c. Based on these top 10 rules, what general insight can Leggere gain on the purchase habits of these customers?
   d. What will be the effect on the rules generated if Leggere decreases the minimum support and reruns the association rules analysis?
   e. What will be the effect on the rules generated if Leggere decreases the minimum confidence and reruns the association rules analysis?

Antecedent	Consequent	Support for A	Support for C	Support for A & C	Confidence	Ratio
BotturaCooking	Cooking	227	862	227	100.00	2.32
Cooking, BobVilla	Art	379	482	205	54.09	2.24
Cooking, Art	Biography	334	554	204	61.08	2.20
Cooking, Biography	Art	385	482	204	52.99	2.20
Youth Fantasy	Novels, Cooking	446	512	245	54.93	2.15
Cooking, Art	BobVilla	334	583	205	61.38	2.11
Cooking, BobVilla	Biography	379	554	218	57.52	2.08
Biography	Novels, Cooking	554	512	293	52.89	2.07
Novels, Cooking	Biography	512	554	293	57.23	2.07
Art	Novels, Cooking	482	512	249	51.66	2.02

DATA *file*
FBS

6. The Football Bowl Subdivision (FBS) level of the National Collegiate Athletic Association (NCAA) consists of over 100 schools. Most of these schools belong to one of several conferences, or collections of schools, that compete with each other on a regular basis in collegiate sports. Suppose the NCAA has commissioned a study that will propose the formation of conferences based on the similarities of the constituent schools. The file *FBS* contains data on schools that belong to the Football Bowl Subdivision. Each row in this file contains information on a school. The variables include football stadium capacity, latitude, longitude, athletic department revenue, endowment, and undergraduate enrollment.
   a. Apply k-means clustering with $k = 10$ using football stadium capacity, latitude, longitude, endowment, and enrollment as variables. Normalize the input variables to adjust for the different magnitudes of the variables. Analyze the resultant clusters. What is the smallest cluster? What is the least dense cluster (as measured by the average distance in the cluster)? What makes the least dense cluster so diverse?
   b. What problems do you see with the plan for defining the school membership of the 10 conferences directly with the 10 clusters?
   c. Repeat part (a), but this time do not normalize the values of the input variables. Analyze the resultant clusters. How and why do they differ from those in part (a)? Identify the dominating factor(s) in the formation of these new clusters.

7. Refer to the clustering problem involving the file *FBS* described in Problem 6. Apply hierarchical clustering with 10 clusters using football stadium capacity, latitude, longitude, endowment, and enrollment as variables. Normalize the values of the input variables to adjust for the different magnitudes of the variables. Use Ward's method as the clustering method.
   a. Compute the cluster centers for the clusters created by the hierarchical clustering. (*Hint:* This can be done using a PivotTable in Excel to calculate the average for each variable for the schools in a cluster.)
   b. Identify the cluster with the largest average football stadium capacity. Using all the variables, how would you characterize this cluster?
   c. Examine the smallest cluster. What makes this cluster unique?
8. Refer to the clustering problem involving the file *FBS* described in Problem 6. Apply hierarchical clustering with 10 clusters using latitude and longitude as variables. Normalize the values of the input variables to adjust for the different magnitudes of the variables. Execute the clustering two times—once with single linkage as the clustering method and once with group average linkage as the clustering method. Compute the cluster sizes and the minimum/maximum latitude and longitude for observations in each cluster. (*Hint:* This can be done using a PivotTable in Excel to display the count of schools in each cluster as well as the minimum and maximum of the latitude and longitude within each cluster.) To visualize the clusters, create a scatter plot with longitude as the *x*-variable and latitude as the *y*-variable. Compare the results of the two approaches.
9. Refer to the clustering problem involving the file *FBS* described in Problem 6. Apply hierarchical clustering with 10 clusters using latitude and longitude as variables. Normalize the values of the input variables to adjust for the different magnitudes of the variables. Execute the clustering two times—once with Ward's method as the clustering method and once with group average linkage as the clustering method. Compute the cluster sizes and the minimum/maximum latitude and longitude for observations in each cluster. (*Hint:* This can be done using a PivotTable in Excel to display the count of schools in each cluster as well as the minimum and maximum of the latitude and longitude within each cluster.) To visualize the clusters, create a scatter plot with longitude as the *x*-variable and latitude as the *y*-variable. Compare the results of the two approaches.
10. Refer to the clustering problem involving the file *FBS* described in Problem 6. Apply hierarchical clustering with 10 clusters using latitude and longitude as variables. Normalize the values of the input variables to adjust for the different magnitudes of the variables. Execute the clustering two times—once with complete linkage as the clustering method and once with Ward's method as the clustering method. Compute the cluster sizes and the minimum/maximum latitude and longitude for observations in each cluster. (*Hint:* This can be done using a PivotTable in Excel to display the count of schools in each cluster as well as the minimum and maximum of the latitude and longitude within each cluster.) To visualize the clusters, create a scatter plot with longitude as the *x*-variable and latitude as the *y*-variable. Compare the results of the two approaches.
11. Refer to the clustering problem involving the file *FBS* described in Problem 6. Apply hierarchical clustering with 10 clusters using latitude and longitude as variables. Normalize the values of the input variables to adjust for the different magnitudes of the variables. Execute the clustering two times—once with centroid linkage as the clustering method and once with group average linkage as the clustering method. Compute the cluster sizes and the minimum/maximum latitude and longitude for observations in each cluster. (*Hint:* This can be done using a PivotTable in Excel to display the count of schools in each cluster as well as the minimum and maximum of the latitude and longitude within each cluster.) To visualize the clusters, create a scatter plot with longitude as the *x*-variable and latitude as the *y*-variable. Compare the results of the two approaches.
12. From 1946 to 1990, the Big Ten Conference consisted of the University of Illinois, Indiana University, University of Iowa, University of Michigan, Michigan State University, University of Minnesota, Northwestern University, Ohio State University, Purdue University, and University of Wisconsin. In 1990, the conference added

DATA file
BigTen

Pennsylvania State University. In 2011, the conference added the University of Nebraska. In 2014, the University of Maryland and Rutgers University were added to the conference with speculation of more schools being added in the future. The file *BigTen* contains the similar information as the file *FBS* (see Problem 6 description), except that each variable value for the original 10 schools in the Big Ten conference have been replaced with the respective variable average over these 10 schools.

Apply hierarchical clustering with complete linkage to yield 2 clusters using football stadium capacity, latitude, longitude, endowment, and enrollment as variables. Normalize the values of the input variables to adjust for the different magnitudes of the variables. Which schools does the clustering suggest would have been the most appropriate to be the eleventh school in the Big Ten? The twelfth and thirteenth schools? What is the problem with using this method to identify the fourteenth school to add to the Big Ten?

13. In this problem, we refer to the clustering problem described in Problem 6, but now we remove the observation for Hawai'i and only consider schools in the continental United States; this modified data is contained in the file *ContinentalFBS*. The NCAA has a preference for conferences consisting of similar schools with respect to their endowment, enrollment, and football stadium capacity, but these conferences must be in the same geographic region to reduce traveling costs. Follow the following steps to address this desire. Apply k-means clustering using latitude and longitude as variables with $k = 3$. Normalize the values of the input variables to adjust for the different magnitudes of the variables. Using the cluster assignments, separate the original data in the Data worksheet into three separate data sets—one data set for each of the three "regional" clusters.

DATA file
ContinentalFBS

   a. For Region 1 data set, apply hierarchical clustering with Ward's method to form three clusters using football stadium capacity, endowment, and enrollment as variables. Normalize the input variables. Report the characteristics of each cluster using a PivotTable that includes a count of number of schools in each cluster, the average stadium capacity, the average endowment amount, and the average enrollment for schools in each cluster.
   b. For the Region 2 data set, apply hierarchical clustering with Ward's method to form four clusters using football stadium capacity, endowment, and enrollment as variables. Normalize the input variables. Report the characteristics of each cluster using a PivotTable that includes a count of number of schools in each cluster, the average stadium capacity, the average endowment amount, and the average enrollment for schools in each cluster.
   c. For the Region 3 data set, apply hierarchical clustering with Ward's method to form two clusters using football stadium capacity, endowment, and enrollment as variables. Normalize the input variables. Report the characteristics of each cluster using a PivotTable that includes a count of number of schools in each cluster, the average stadium capacity, the average endowment amount, and the average enrollment for schools in each cluster.
   d. What problems do you see with the plan with defining the school membership of nine conferences directly with the nine total clusters formed from the regions? How could this approach be tweaked to solve this problem?

14. IBM employs a network of expert analytics consultants for various projects. To help it determine how to distribute its bonuses, IBM wants to form groups of employees with similar performance according to key performance metrics. Each observation (corresponding to an employee) in the file *BigBlue* consists of values for: *UsageRate* which corresponds to the proportion of time that the employee has been actively working on high-priority projects, *Recognition* which is the number of projects for which the employee was specifically requested, and *Leader* which is the number of projects on which the employee has served as project leader. Apply k-means clustering with values of $k = 2$ to 7. Normalize the values of the input variables to adjust for the different magnitudes of the variables. How many clusters do you recommend to categorize the employees? Why?

DATA file
BigBlue

15. Apply hierarchical clustering to the data in *DemoKTC* using matching coefficients as the similarity measure and group average linkage as the clustering method to create three clusters based on the Female, Married, Loan, and Mortgage variables. Use a PivotTable to count the total number of customers in each cluster as well as the number of customers who are female, the number of customers who are married, the number of customers with a car loan, and the number of customers with a mortgage in each cluster. How would you characterize each cluster?

DATA file
DemoKTC

16. Apply k-means clustering with values of $k = 2$, 3, 4, and 5 to cluster the data in *DemoKTC* based on the Age, Income, and Children variables. Normalize the values of the input variables to adjust for the different magnitudes of the variables. How many clusters do you recommend? Why?

17. Attracted by the possible returns from a portfolio of movies, hedge funds have invested in the movie industry by financially backing individual films and/or studios. The hedge fund Star Ventures is currently conducting some research involving movies involving Adam Sandler, an American actor, screenwriter, and film producer. As a first step, Star Ventures would like to cluster Adam Sandler movies based on their gross box office returns and movie critic ratings. Using the data in the file *Sandler*, apply k-means clustering with $k = 3$ to characterize three different types of Adam Sandler movies. Base the clusters on the variables Rating and Box. Rating corresponds to movie ratings provided by critics (a higher score represents a movie receiving better reviews). Box represents the gross box office earnings in 2015 dollars. Normalize the values of the input variables to adjust for the different magnitudes of the variables. Report the characteristics of each cluster using a PivotTable that includes a count of movies, the average rating of movies and the average box office earnings of movies in each cluster. How would you characterize the movies in each cluster?

DATA file
Sandler

18. Josephine Mater works for the supply-chain analytics division of Trader Joe's, a national chain of specialty grocery stores. Trader Joe's is considering a redesign of its supply chain. Josephine knows that Trader Joe's uses frequent truck shipments from its distribution centers to its retail stores. To keep costs low, retail stores are typically located near a distribution center. The file *TraderJoes* contains data on the location of Trader Joe's retail stores. Josephine would like to use k-means clustering with $k = 8$ to estimate the preferred locations if Trader Joe's was to establish eight distribution centers to support its retail stores. Normalize the values of the input variables to adjust for the different magnitudes of the variables. If Trader Joe's establishes eight distribution centers, how many retail stores are assigned to each distribution center? What are the drawbacks to using this solution approach to assign retail stores to distribution centers?

DATA file
TraderJoes

19. Apple Inc. tracks online transactions at its iStore and is interested in learning about the purchase patterns of its customers in order to provide recommendations as a customer browses its web site. A sample of the "shopping cart" data resides in the files *AppleCartBinary* and *AppleCartStacked*.

    Use a minimum support of 10% of the total number of transactions and a minimum confidence of 50% to generate a list of association rules.

    a. Interpret what the rule with the largest lift ratio is saying about the relationship between the antecedent item set and consequent item set.
    b. Interpret the confidence of the rule with the largest lift ratio.
    c. Interpret the lift ratio of the rule with the largest lift ratio.
    d. Review the top 15 rules and summarize what the rules suggest.

DATA file
AppleCartBinary
AppleCartStacked

20. Cookie Monster Inc. is a company that specializes in the development of software that tracks web browsing history of individuals. A sample of browser histories is provided in the files *CookieMonsterBinary* and *CookieMonsterStacked* that indicate which websites were visited by which customers.

    Use a minimum support of 4% of the transactions (800 of the 20,000 total transactions) and a minimum confidence of 50% to generate a list of association rules. Review the top 14 rules. What information does this analysis provide Cookie Monster Inc. regarding the online behavior of individuals?

DATA file
CookieMonsterBinary
CookieMonsterStacked

DATA file
GroceryStoreList
GroceryStoreStacked

21. A grocery store introducing items from Italy is interested in analyzing buying trends of these new "international" items, namely prosciutto, Peroni, risotto, and gelato. The files *GroceryStoreList* and *GroceryStoreStacked* provide data on a collection of transactions in item-list format.
    a. Use a minimum support of 100 transactions (10% of the 1,000 total transactions) and a minimum confidence of 50% to generate a list of association rules. How many rules satisfy this criterion?
    b. Use a minimum support of 250 transactions (25% of the 1,000 total transactions) and a minimum confidence of 50% to generate a list of association rules. How many rules satisfy this criterion? Why may the grocery store want to increase the minimum support required for their analysis? What is the risk of increasing the minimum support required?
    c. Using the list of rules from part (b), consider the rule with the largest lift ratio that also involves an Italian item. Interpret what this rule is saying about the relationship between the antecedent item set and consequent item set.
    d. Interpret the confidence of the rule with the largest lift ratio that also involves an Italian item.
    e. Interpret the lift ratio of the rule with the largest lift ratio that also involves an Italian item.
    f. What insight can the grocery store obtain about its purchasers of the Italian fare?

DATA file
AirlineTweets

22. Companies can learn a lot about customer experiences by monitoring the social media web site Twitter. The file *AirlineTweets* contains a sample of 36 tweets of an airline's customers. Normalize the terms by using stemming and generate binary term-document matrix.
    a. What are the five most common terms occurring in these tweets? How often does each term appear?
    b. Apply hierarchical clustering using complete linkage to yield three clusters on the binary term-document matrix using the tokens agent, attend, bag, damag, and rude as variables. How many documents are in each cluster? Give a description of each cluster.
    c. How could management use the results obtained in part (b)?

    Source: Kaggle website

DATA file
YelpItalian

23. The online review service Yelp helps millions of consumers find the goods and services they seek. To help consumers make more-informed choices, Yelp includes over 120 million reviews. The file *YelpItalian* contains a sample of 21 reviews for an Italian restaurant. Normalize the terms by using stemming and a generate binary term-document matrix.
    a. What are the five most common terms in these reviews? How often does each term appear?
    b. Apply hierarchical clustering using complete linkage to yield two clusters from the presence/absence term-document matrix using all five of the most common terms from the reviews. How many documents are in each cluster? Give a description of each cluster.

CASE PROBLEM: KNOW THY CUSTOMER

DATA file
KnowThyCustomer

Know Thy Customer (KTC) is a financial consulting company that provides personalized financial advice to its clients. As a basis for developing this tailored advising, KTC would like to segment its customers into several representative groups based on key characteristics. Peyton Blake, the director of KTC's fledging analytics division, plans to establish the set of representative customer profiles based on 600 customer records in the file *KnowThyCustomer*. Each customer record contains data on age, gender, annual income, marital status, number of children, whether the customer has a car loan, and whether the customer

has a home mortgage. KTC's market research staff has determined that these seven characteristics should form the basis of the customer clustering.

Peyton has invited a summer intern, Danny Riles, into her office so they can discuss how to proceed. As they review the data on the computer screen, Peyton's brow furrows as she realizes that this task may not be trivial. The data contains both categorical variables (Female, Married, Car, and Mortgage) and numerical variables (Age, Income, and Children).

1. Using hierarchical clustering on all seven variables, experiment with using complete linkage and group average linkage as the clustering method. Normalize the values of the input variables. Recommend a set of customer profiles (clusters). Describe these clusters according to their "average" characteristics. Why might hierarchical clustering not be a good method to use for these seven variables?
2. Apply a two-step clustering method:
   a. Use hierarchical clustering with matching coefficients as the similarity measure and group average linkage as the clustering method to produce four clusters using the variables Female, Married, Loan, and Mortgage.
   b. Based on the clusters from part (a), split the original 600 observations into four separate data sets as suggested by the four clusters from part (a). For each of these four data sets, apply k-means clustering with $k = 2$ using Age, Income, and Children as variables. Normalize the values of the input variables. This will generate a total of eight clusters. Describe these eight clusters according to their "average" characteristics. What benefit does this two-step clustering approach have over just using hierarchical clustering on all seven variables as in part (1) or just using k-means clustering on all seven variables? What weakness does it have?