DATA MINING TECHNIQUES

FOR MARKETING, SALES, AND CUSTOMER RELATIONSHIP MANAGEMENT

THIRD EDITION

BY GORDON S. LINOFF AND
MICHAEL J. A. BERRY

Contents

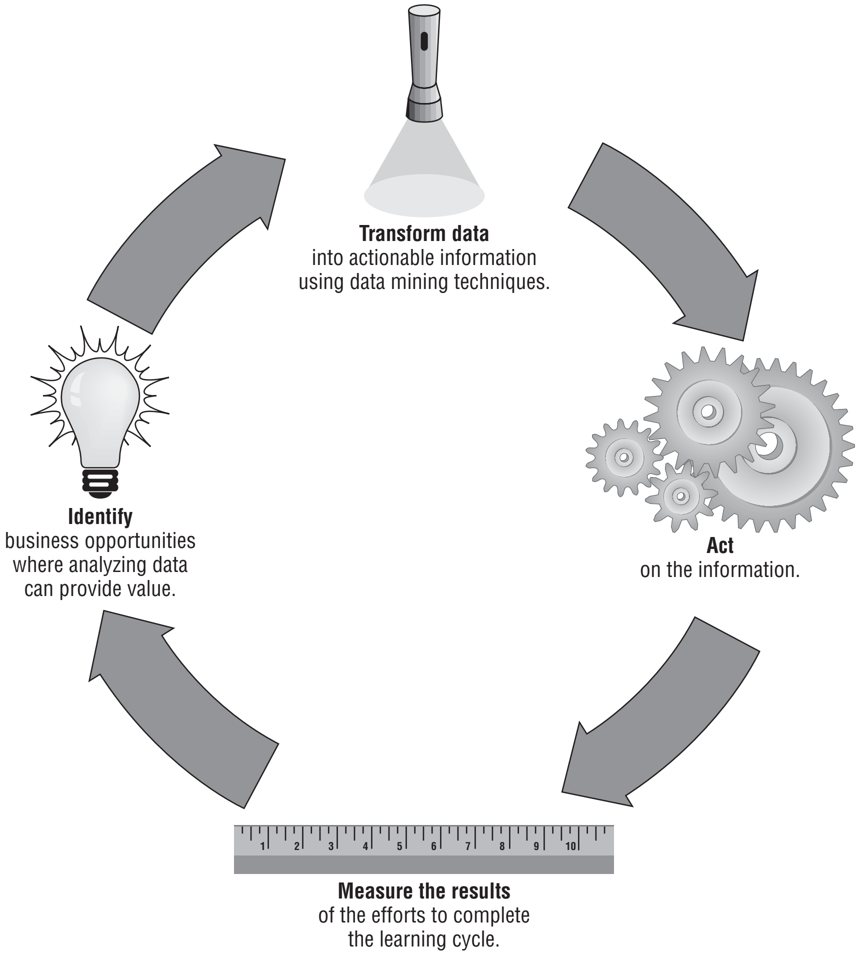


Figure 1-1: The virtuous cycle of data mining focuses on business results, rather than just exploiting advanced techniques.

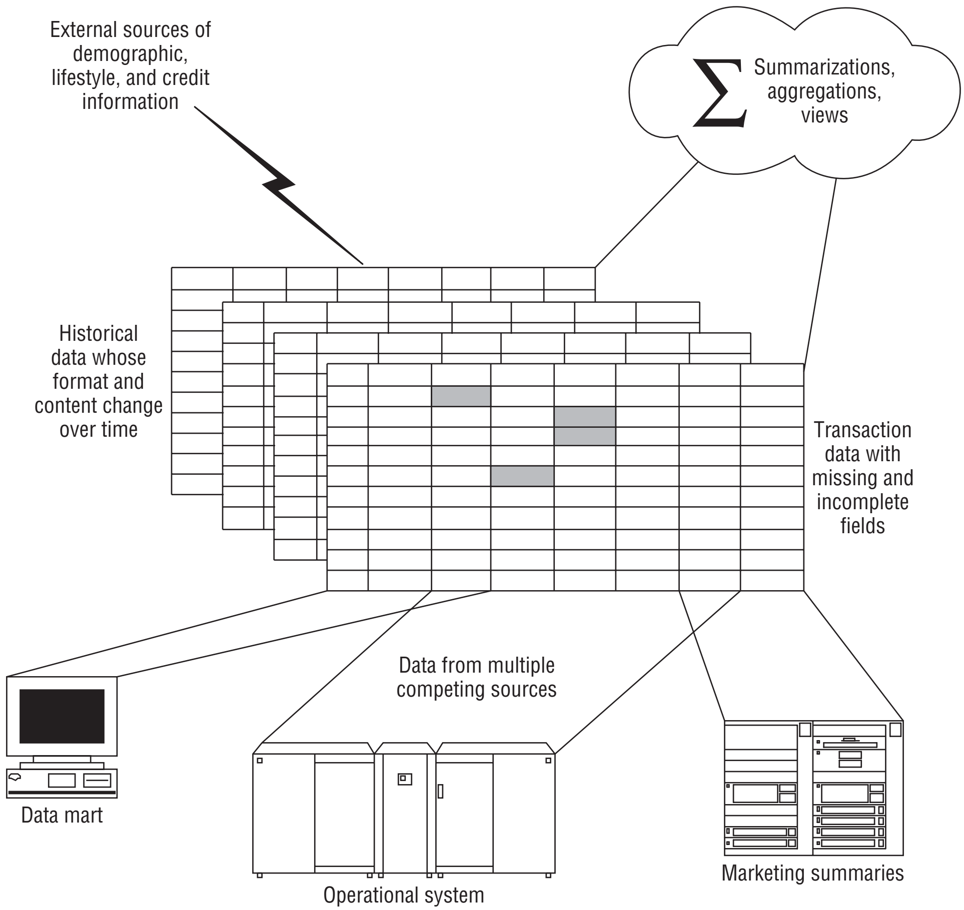
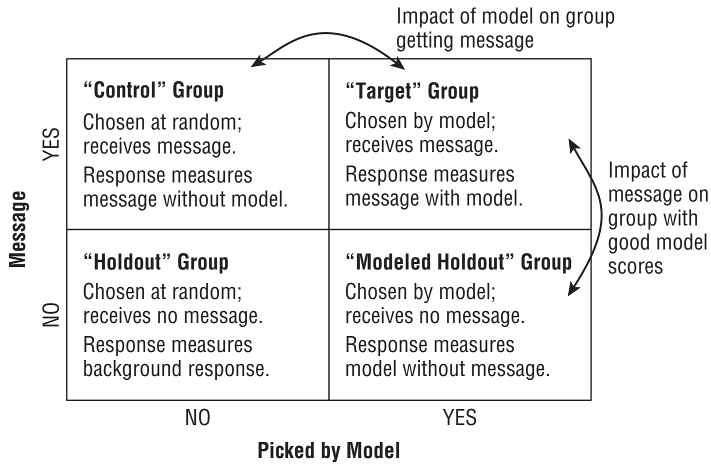
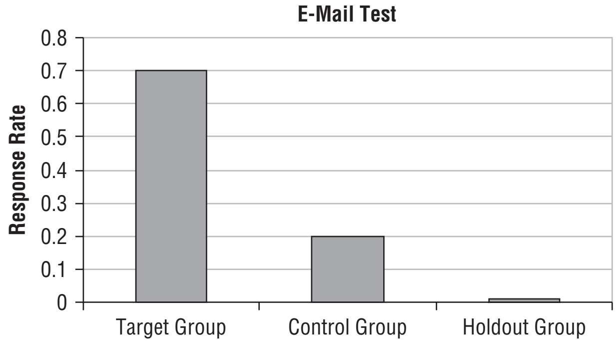


Figure 1-2: Data is never clean. It comes in many forms, from many sources both internal and external.



These four groups are used for measuring the effectiveness of both the message and the modeling effort.

4 Groups in Marketing Research



This chart readily shows the difference in response to determine whether the treatment works and whether the modeling works.

Bank Chart

Table 1-1: Data Mining Differs from Typical Operational Business Processes

TYPICAL OPERATIONAL SYSTEM	DATA MINING SYSTEM
Operations and reports on historical data	Analysis on historical data often applied to most current data to determine future actions
Predictable and periodic flow of work, typically tied to calendar	Unpredictable flow of work depending on business and marketing needs
Focus on individual items, one at a time (the needle in the haystack)	Focusing on larger groups at one time, trying to make sense of the haystack
Limited use of enterprise-wide data	The more data, the better the results (generally)
Focus on line of business (such as account, region, product code, minutes of use, and so on), not on customer	Focus on actionable entity, product, customer, sales region
Response times often measured in seconds/milliseconds (for interactive systems) while waiting weeks/month for reports	Iterative processes with response times often measured in minutes or hours
System of record for data	Copy of data
Descriptive and repetitive	Creative

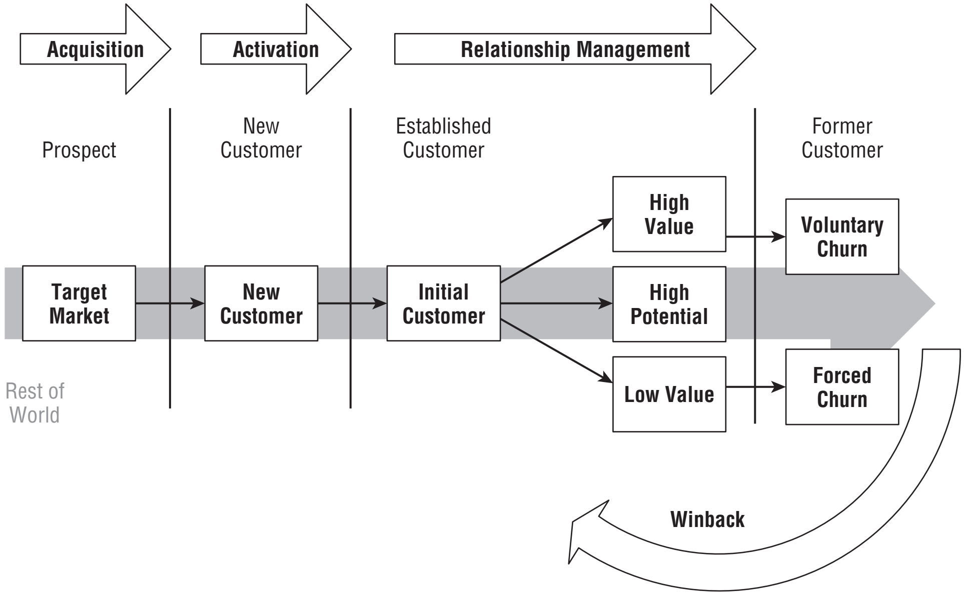


Figure 2-1: The customer lifecycle progresses through different stages.

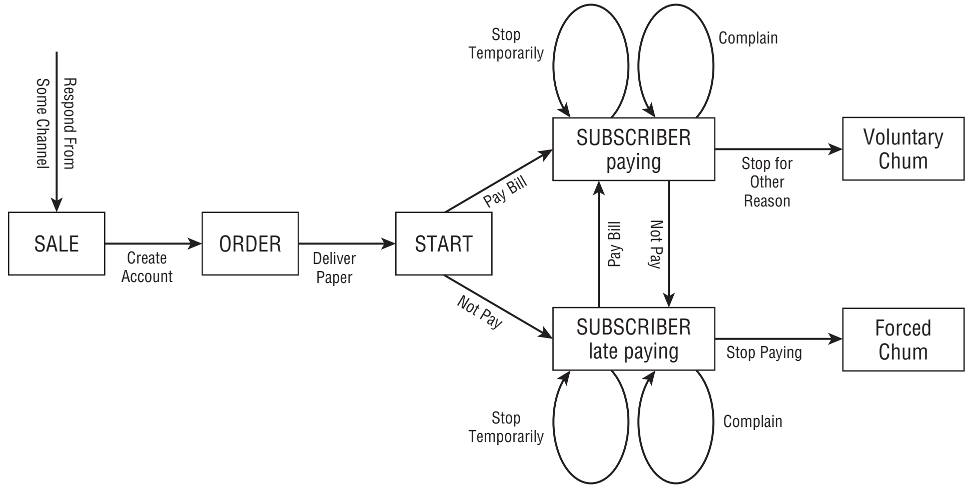


Figure 2-2: (Simplified) customer experience for newspaper subscribers includes several different types of interactions.

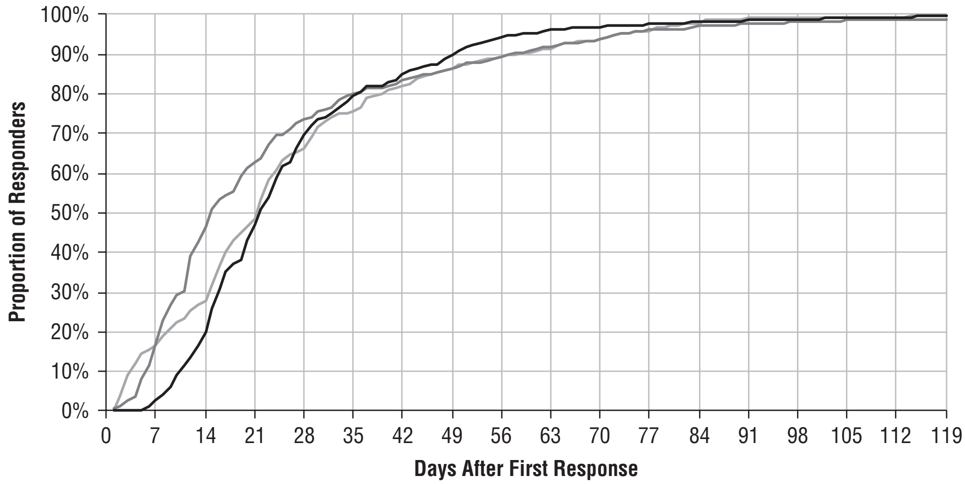


Figure 2-3: These response curves for three direct mail campaigns show that 80 percent of the responses came within five to six weeks.

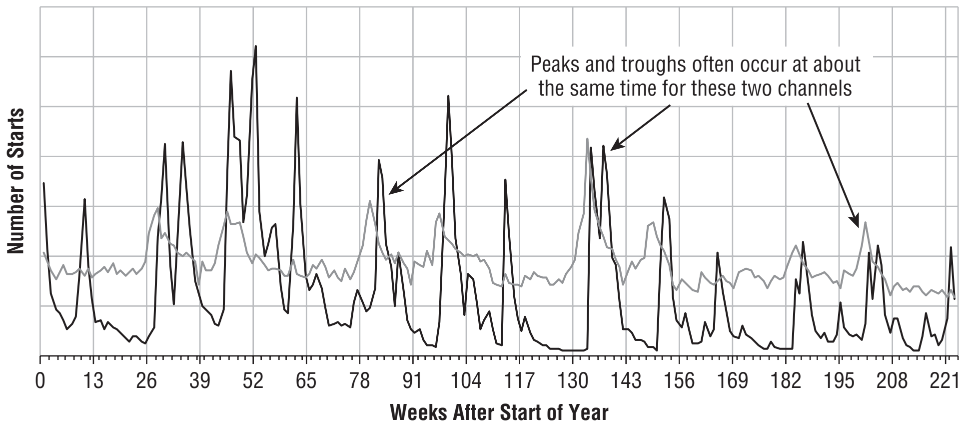


Figure 2-4: The echo effect may artificially under- or overestimate the performance of channels, because customers inspired by one channel may be attributed to another.

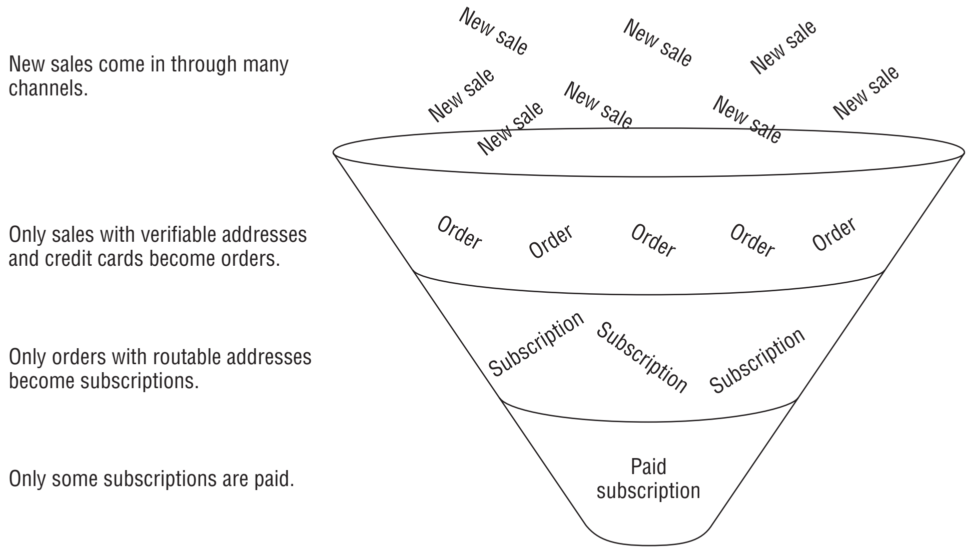


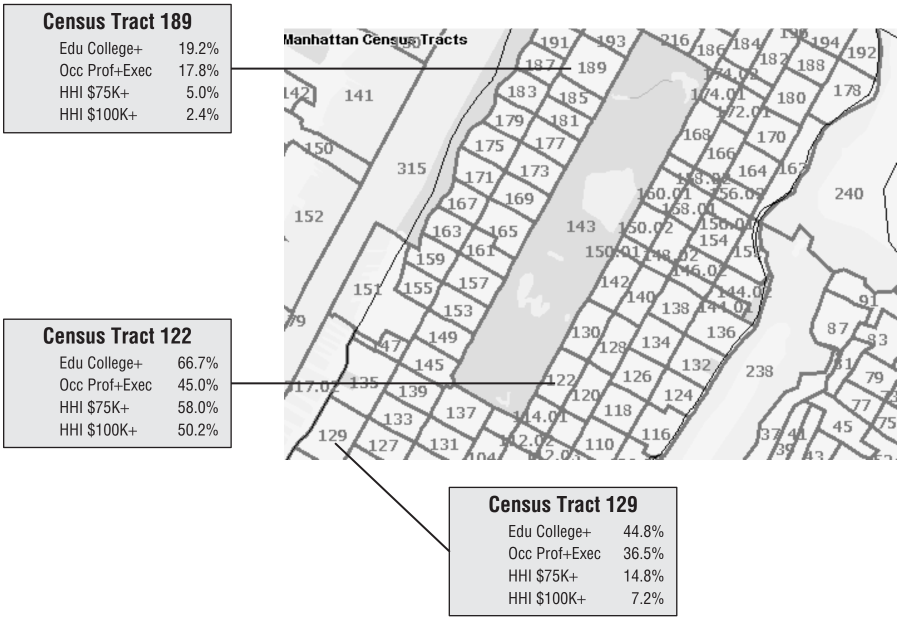
Figure 2-5: The customer activation process funnel eliminates responders at each step of the activation process.

Table 2-1: Calculating Fitness Scores for Individuals by Comparing Them along Each Demographic Measure

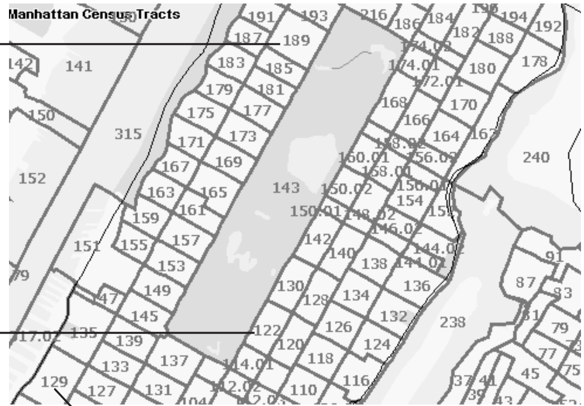
	READER-SHIP	YES SCORE	NO SCORE	AMY	BOB	AMY SCORE	BOB SCORE
College educated	58%	0.58	0.42	YES	NO	0.58	0.42
Prof or exec	46%	0.46	0.54	YES	NO	0.46	0.54
Income >$75K	21%	0.21	0.79	YES	NO	0.21	0.79
Income >$100K	7%	0.07	0.93	NO	NO	0.93	0.93
Total						2.18	2.68

Table 2-2: Calculating Scores by Taking the Proportions in the Population into Account

	YES			NO		
	READERSHIP	U.S. POP.	INDEX	READERSHIP	U.S. POP.	INDEX
College educated	58%	20.3%	2.86	42%	79.7%	0.53
Professional or executive	46%	19.2%	2.40	54%	80.8%	0.67
Income >$75K	21%	9.5%	2.21	79%	90.5%	0.87
Income >$100K	7%	2.4%	2.92	93%	97.6%	0.95



Census Tracts



Tract 189	Goal	Tract	Fitness
Edu College+	19.2%	61.3%	0.31
Occ Prof+Exec	17.8%	45.5%	0.39
HHI $75K+	5.0%	22.6%	0.22
HHI $100K+	2.4%	7.4%	0.32
Overall Advertising Fitness			**0.31**

Tract 122	Goal	Tract	Fitness
Edu College+	66.7%	61.3%	1.00
Occ Prof+Exec	45.0%	45.5%	0.99
HHI $75K+	58.0%	22.6%	1.00
HHI $100K+	50.2%	7.4%	1.00
Overall Advertising Fitness			**1.00**

Tract 129	Goal	Tract	Fitness
Edu College+	44.8%	61.3%	0.73
Occ Prof+Exec	36.5%	45.5%	0.80
HHI $75K+	14.8%	22.6%	0.65
HHI $100K+	7.2%	7.4%	0.97
Overall Advertising Fitness			**0.79**

Figure 2-6: Example of calculating readership fitness for three census tracts in Manhattan.

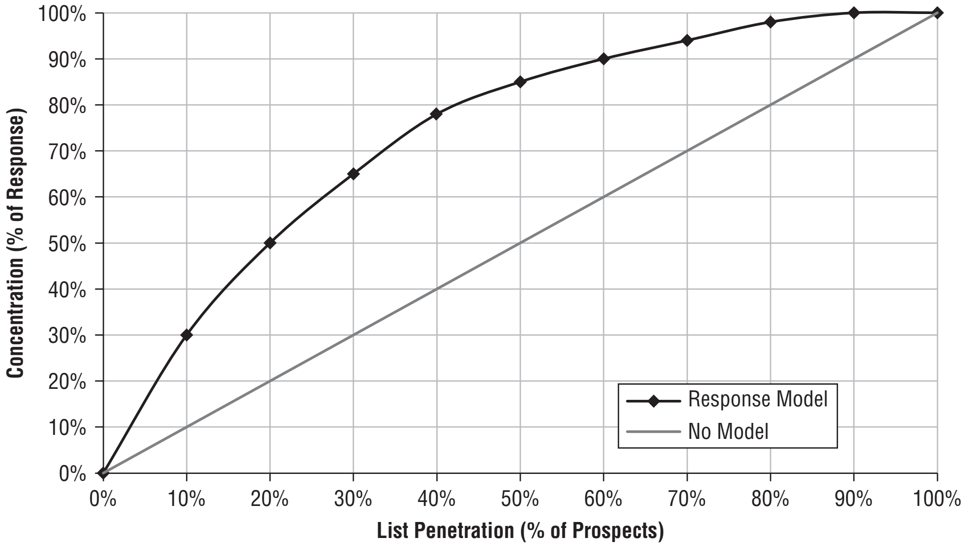


Figure 2-7: A cumulative gains or concentration chart shows the benefit of using a model.

Table 2-3: Profit/LossMatrixfortheSimplifyingAssumptionsCorporation

MAILED	RESPONDED	
	YES	**NO**
YES	$44	–$1
NO	$0	$0

Table 2-4: Lift and Cumulative Gains by Decile

PENETRATION	GAINS	CUMULATIVE GAINS	LIFT
0%	0%	0%	0.000
10%	30%	30%	3.000
20%	20%	50%	2.500
30%	15%	65%	2.167
40%	13%	78%	1.950
50%	7%	85%	1.700
60%	5%	90%	1.500
70%	4%	94%	1.343
80%	4%	96%	1.225
90%	2%	100%	1.111
100%	0%	100%	1.000

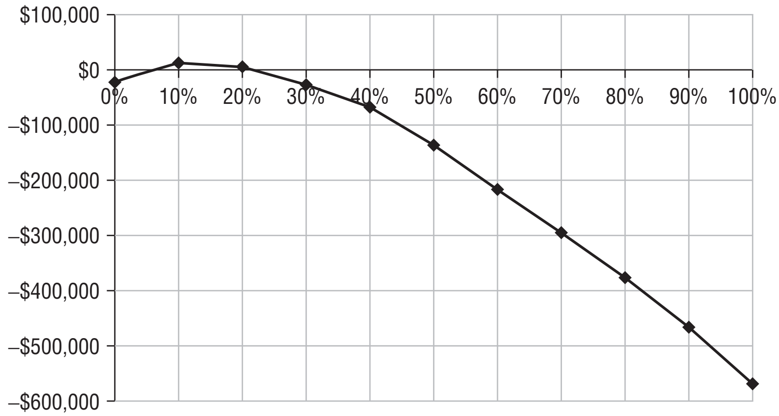


Figure 2-8: Campaign profitability as a function of penetration.

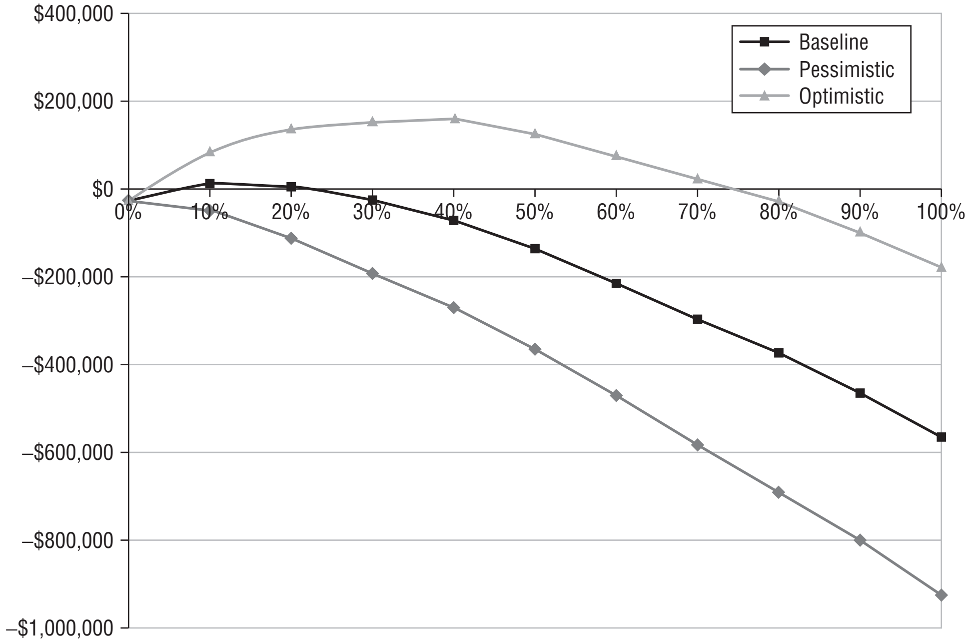


Figure 2-9: A 20 percent variation in response rate, cost, and revenue per responder has a large effect on the profitability of a campaign.

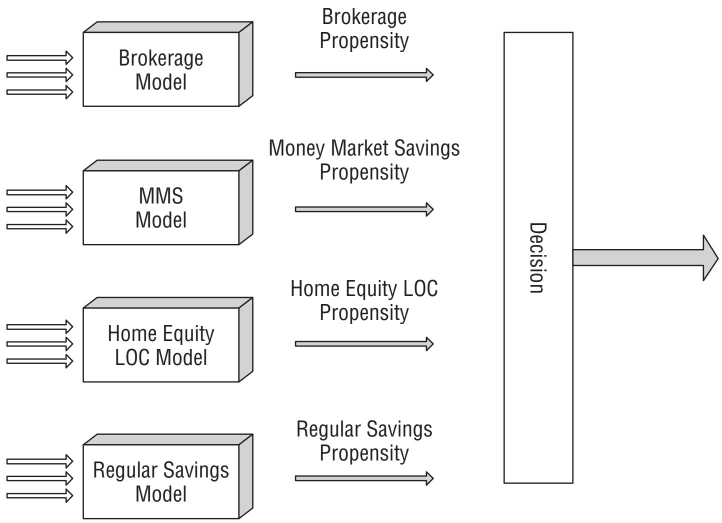


Figure 2-10: Comparing scores from multiple models to decide which offers will be shown to customers.

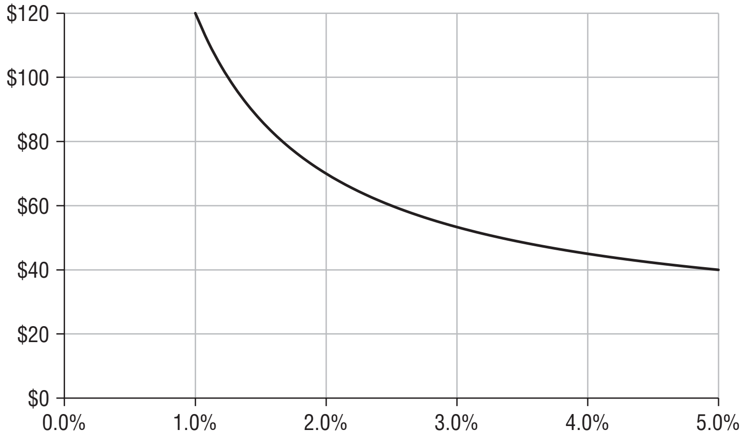


Figure 2-11: As the response rate to an acquisition campaign goes down, the cost per customer acquired goes up.

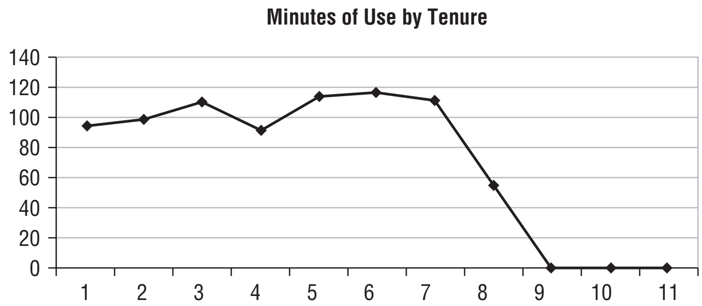


Figure 3-1: Does declining usage in month 8 predict attrition in month 9?

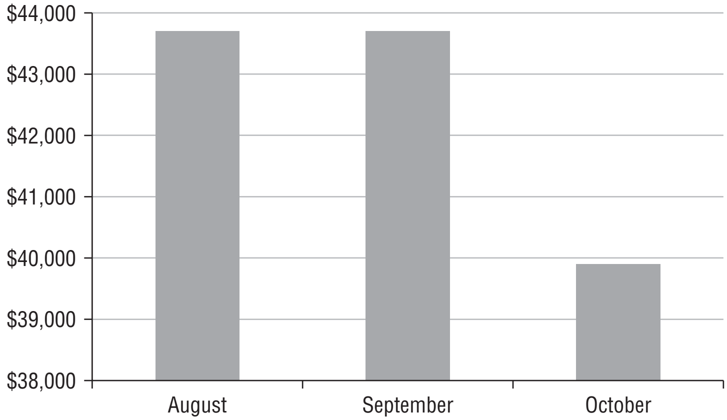


Figure 3-2: Did sales really drop off in October?

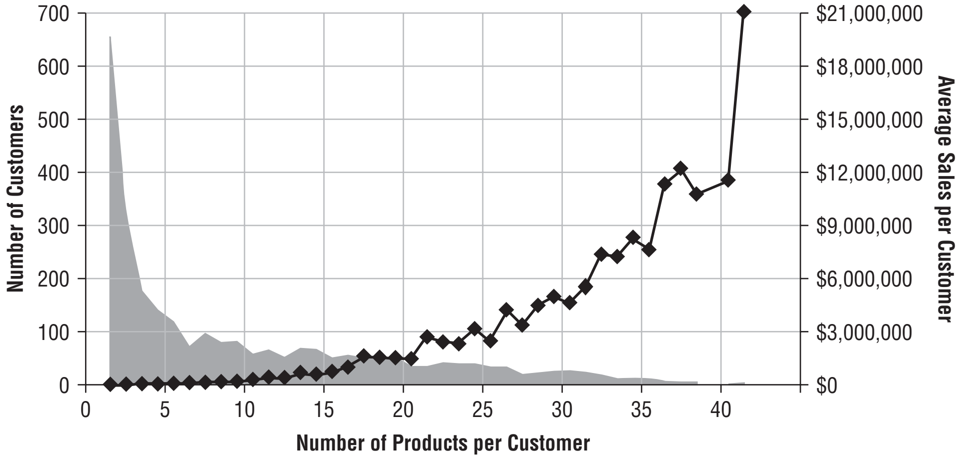


Figure 3-3: Customers who buy more product types spend more money.

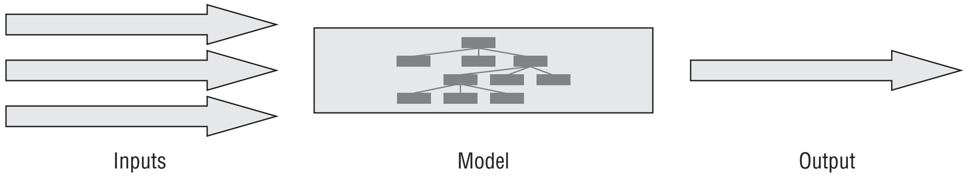


Figure 3-4: Models take an input and produce an output.

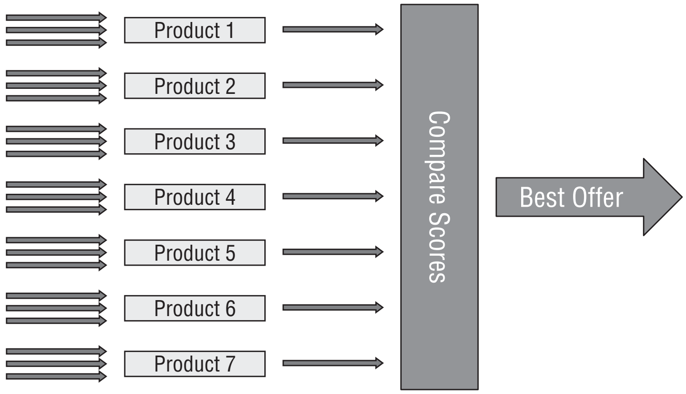


Figure 3-5: Individual propensity scores for each product are compared to determine the best offer.

CustomerID	Response	Contribution	X1	X2	X3
292129	0		A	39,220	1
292130	0		A	39,749	1
292134	0		C	40,052	1
197549	0		A	39,485	1
292137	0		A	39,749	1
291800	0		A	39,610	1
292138	0		A	39,749	0
332806	0		A	39,860	0
292140	0		A	39,686	1
347807	1	$40	C	40,139	0
292141	0		A	39,749	1
292143	1	$30	C	40,027	0
409542	0		A	40,050	0
292848	0		C	40,012	1
292850	0		C	40,151	1
292851	0		A	39,750	0
292852	0		C	39,997	1
292853	0		A	39,750	1
292857	0		A	39,750	1
292859	1	$30	A	39,994	1
292860	0		A	39,750	0
292861	0		A	39,750	0
292862	1	$30	C	39,859	0
292863	0		C	39,877	1
292864	1	$40	C	40,071	1
292868	0		A	39,750	0
403246	0		A	40,035	0
292869	1	$30	D	40,132	0
292870	0		C	39,788	0
292871	0		A	39,750	1
292872	0		A	39,750	1
292873	0		C	39,997	1
292874	1	$40	C	40,150	1
292878	0		A	39,750	1
292879	1	$40	C	40,132	0
292880	1	$30	C	39,859	1
292881	0		C	39,879	0
24583	0		A	38,966	0
292884	0		A	39,750	1
126612	1	$40	A	40,016	0
292886	0		A	39,288	1
292887	0		A	39,750	1
292888	1	$40	A	40,113	0
292889	0		C	39,795	0
390095	0		A	40,000	1
292893	0		A	39,462	1
292894	0		A	40,118	1
292964	0		D	40,138	0
292897	1	$30	C	39,859	1
292900	0		A	39,750	1
292901	0		C	39,808	1
292902	1	$30	C	39,859	0
292905	0		A	39,750	1
292908	0		A	39,750	0
292909	0		A	39,750	1
292911	0		A	39,750	1
292913	0		C	39,798	1
292914	1	$30	D	40,132	0
292915	0		A	39,750	0
292916	0		C	39,812	0
292917	0		A	39,750	0
292919	0		A	39,750	1
292920	0		D	40,114	0

Response model based on all rows of training data:

$$P(\text{response}) = f(X_1, X_2, X_3)$$

Contribution model based on responders:

$$E(\$|\text{response}) = g(X_1, X_2, X_3)$$

Both models are applied to all rows of a table describing potential contributors. The expected contribution is the product of the two model results:

$$E(\$) = E*P$$

A two-stage model for the expected value of a contribution

Two-Stage Model

Table 3-1: What Techniques for Which Tasks?

TASK	BEST FIT	ALSO CONSIDER
Classification and prediction	Decision trees, logistic regression, neural networks	Similarity models, table look-up models, nearest neighbor models, naïve Bayesian models
Estimation	Linear regression, neural networks	Regression trees, nearest neighbor models
Binary response	Logistic regression, decision trees	Similarity models, table look-up models, nearest neighbor models, naïve Bayesian models
Finding clusters and patterns	Any of the clustering algorithms	Association rules

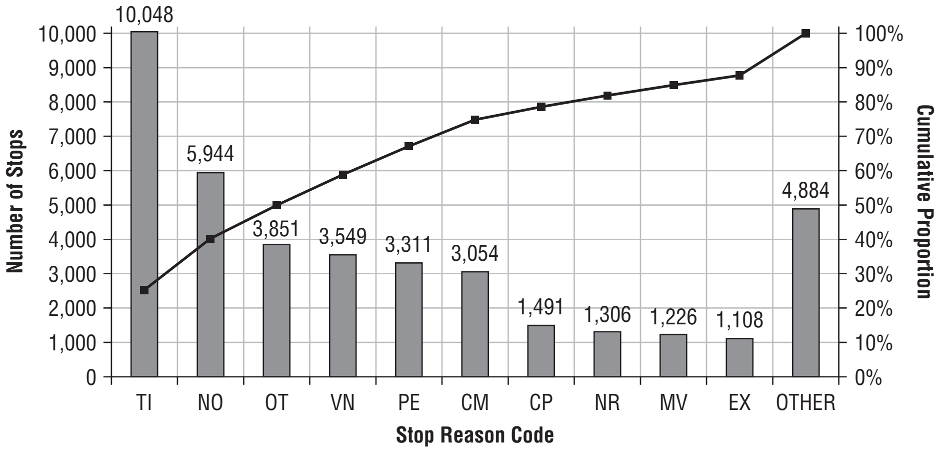


Figure 4-1: This example shows both a histogram (as a vertical bar chart) and cumulative proportion (as a line) on the same chart for stop reasons associated with a particular marketing effort.

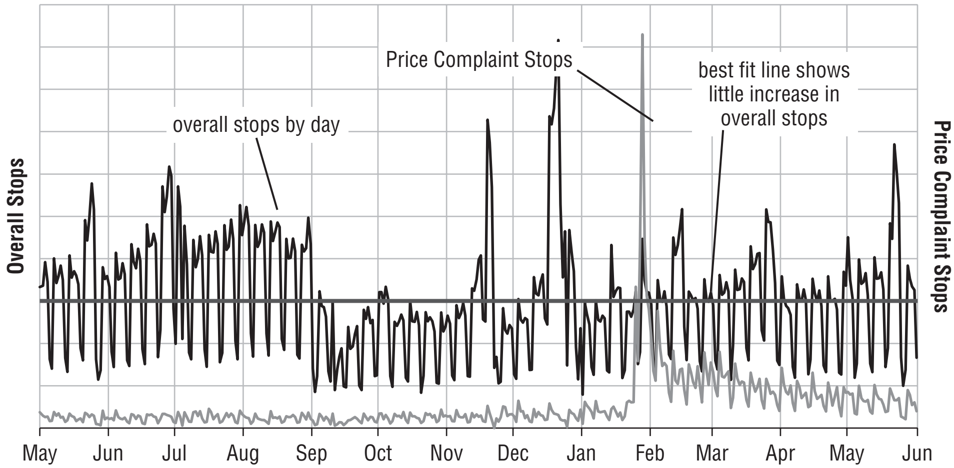


Figure 4-2: This chart shows two time series plotted with different vertical scales. The dark line is for overall stops; the light line for pricing related stops shows the impact of a change in pricing strategy at the end of January.

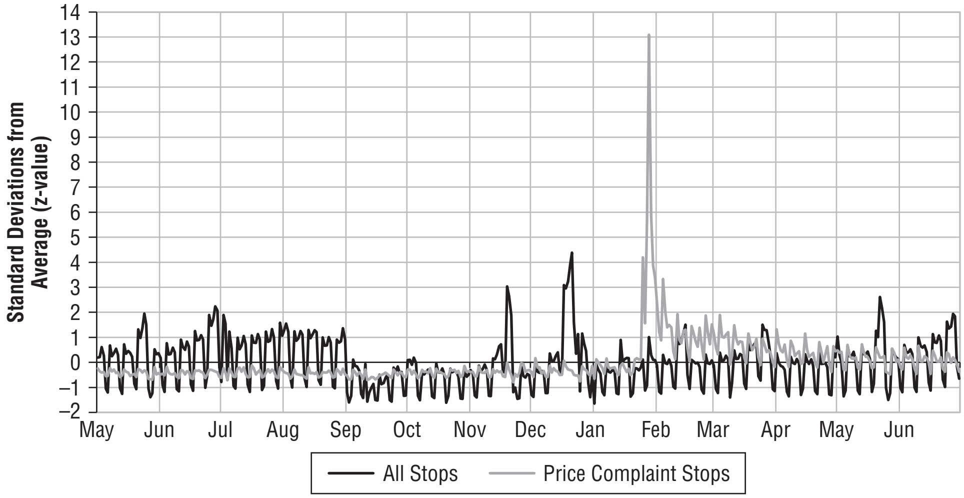
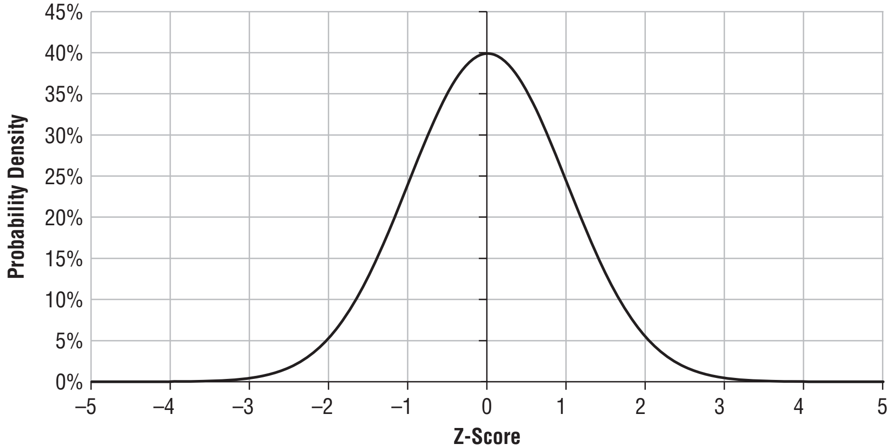


Figure 4-3: Standardized values allow you to compare different groups on the same chart using the same scale; this chart shows overall stops and price increase–related stops.



The probability density function for the normal distribution looks like the familiar bell-shaped curve.

Probability Density Distribution

The (cumulative) distribution function for the normal distribution has an S-shape.

Normal Distribution

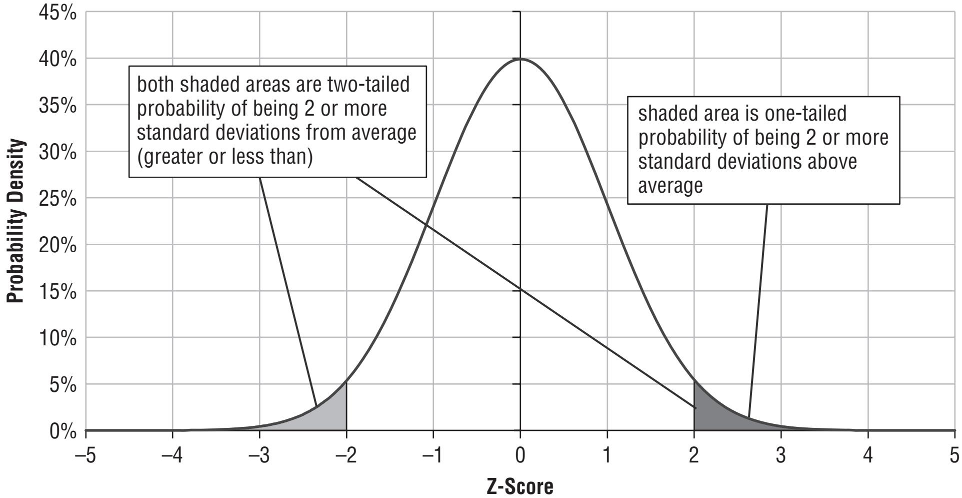


Figure 4-4: The tail of the normal distribution answers the question: "What is the probability of getting a value of *z* or greater?"

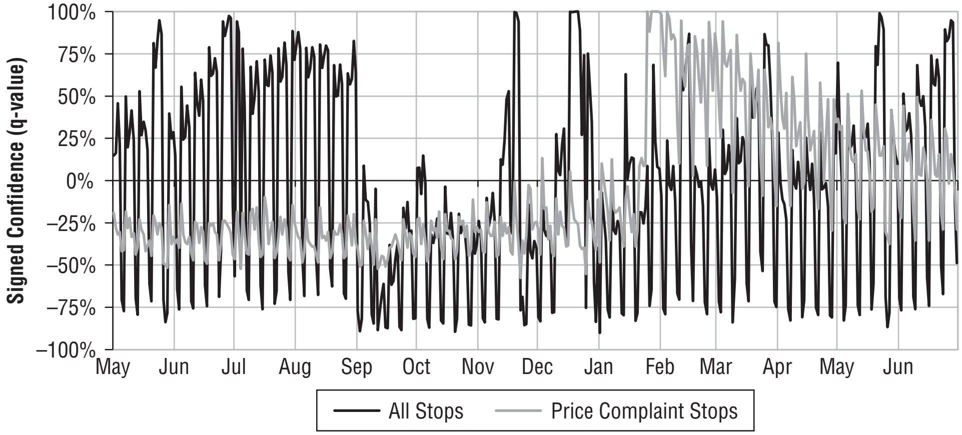


Figure 4-5: Based on the same data from Figures 4-2 and 4-3, this chart shows the signed confidence (q-values) of the observed value based on the average and standard deviation. This sign is positive when the observed value is too high, negative when it is too low.

Table 4-1: Cross-tabulation of Starts by County and Channel

	COUNTS				FREQUENCIES			
COUNTY	TM	DM	OTHER	TOTAL	TM	DM	OTHER	TOTAL
Bronx	3,212	413	2,936	**6,561**	2.5%	0.3%	2.3%	5.1%
Kings	9,773	1,393	11,025	**22,191**	7.7%	1.1%	8.6%	17.4%
Nassau	3,135	1,573	10,367	**15,075**	2.5%	1.2%	8.1%	11.8%
New York	7,194	2,867	28,965	**39,026**	5.6%	2.2%	22.7%	30.6%
Queens	6,266	1,380	10,954	**18,600**	4.9%	1.1%	8.6%	14.6%
Richmond	784	277	1,772	**2,833**	0.6%	0.2%	1.4%	2.2%
Suffolk	2,911	1,042	7,159	**11,112**	2.3%	0.8%	5.6%	8.7%
Westchester	2,711	1,230	8,271	**12,212**	2.1%	1.0%	6.5%	9.6%
Total	**35,986**	**10,175**	**81,449**	**127,610**	28.2%	8.0%	63.8%	100.0%

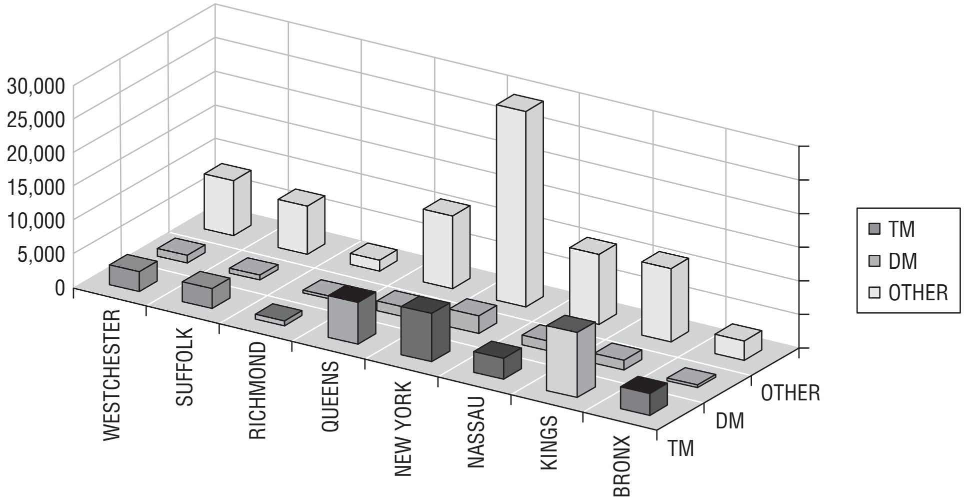


Figure 4-6: A surface plot provides a visual interface for cross-tabulated data.

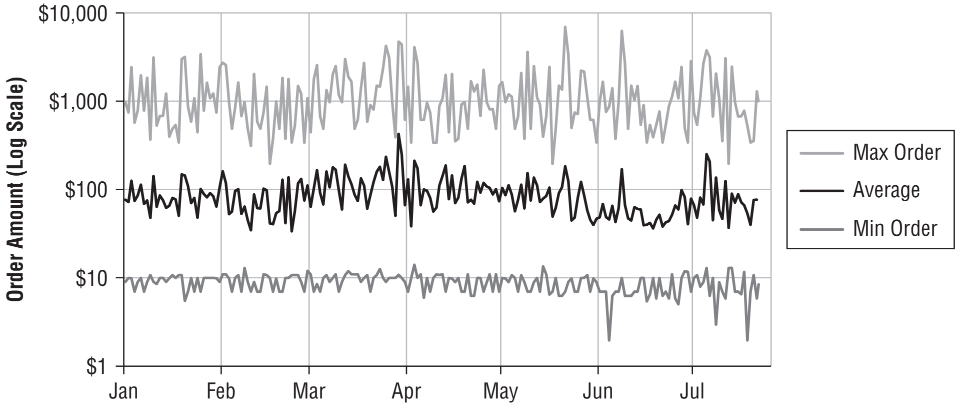


Figure 4-7: A time chart can also be used for continuous values; this one shows the range and average for order amounts each day.

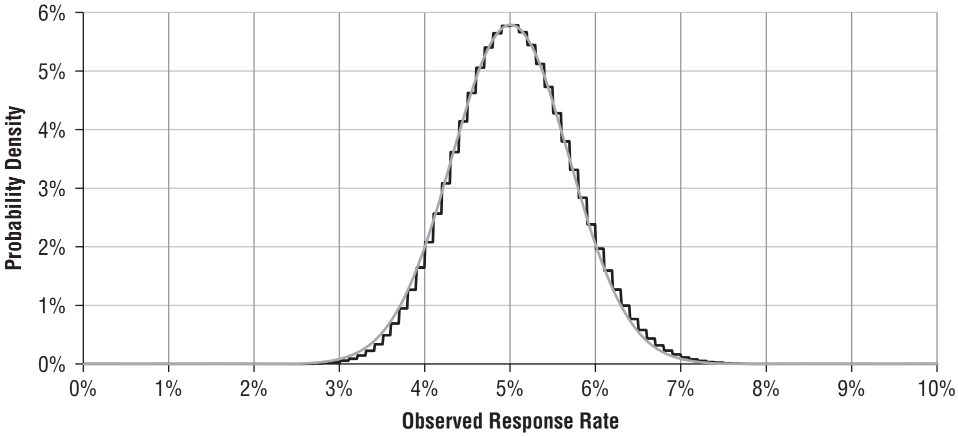


Figure 4-8: Statistics has proven that actual response rate on a population is very close to a normal distribution whose average is the measured response on a sample and whose standard deviation is the standard error of proportion (SEP).

$$SEP = \sqrt{\left(\frac{p * (1 - p)}{N}\right)}$$

Equation 1

Table 4-2: The 95 Percent Confidence Interval Bounds for the Champion Group for Different Response Rates

RESPONSE	SIZE	SEP	95% CONF	95% CONF * SEP	LOWER	UPPER
4.5%	900,000	0.0219%	1.96	0.0219%*1.96=0.0429%	4.46%	4.54%
4.6%	900,000	0.0221%	1.96	0.0221%*1.96=0.0433%	4.56%	4.64%
4.7%	900,000	0.0223%	1.96	0.0223%*1.96=0.0437%	4.66%	4.74%
4.8%	900,000	0.0225%	1.96	0.0225%*1.96=0.0441%	4.76%	4.84%
4.9%	900,000	0.0228%	1.96	0.0228%*1.96=0.0447%	4.86%	4.94%
5.0%	900,000	0.0230%	1.96	0.0230%*1.96=0.0451%	4.95%	5.05%
5.1%	900,000	0.0232%	1.96	0.0232%*1.96=0.0455%	5.05%	5.15%
5.2%	900,000	0.0234%	1.96	0.0234%*1.96=0.0459%	5.15%	5.25%
5.3%	900,000	0.0236%	1.96	0.0236%*1.96=0.0463%	5.25%	5.35%
5.4%	900,000	0.0238%	1.96	0.0238%*1.96=0.0466%	5.35%	5.45%
5.5%	900,000	0.0240%	1.96	0.0240%*1.96=0.0470%	5.45%	5.55%

Response rates vary from 4.5% to 5.5%. The bounds for the 95% confidence level are calculated using1.96 standard deviations from the average.

$$\text{SEDP} = \sqrt{\left(\frac{p_1 * (1 - p_1)}{N_1} + \frac{p_2 * (1 - p_2)}{N_2}\right)}$$

Equation 2

Table 4-3: Z-scores and P-values for Difference Between Champion and Challenger Response Rates, with 900,000 contacts for Champion and 100,000 for Challenger

RESPONSE			DIFFERENCE OF PROPORTIONS		
CHAMPION	CHALLENGER	DIFFERENCE	SEDP	Z-VALUE	P-VALUE
5.0%	4.5%	0.5%	0.07%	6.9	0.0%
5.0%	4.6%	0.4%	0.07%	5.5	0.0%
5.0%	4.7%	0.3%	0.07%	4.1	0.0%
5.0%	4.8%	0.2%	0.07%	2.8	0.6%
5.0%	4.9%	0.1%	0.07%	1.4	16.8%
5.0%	5.0%	0.0%	0.07%	0.0	100.0%
5.0%	5.1%	-0.1%	0.07%	-1.4	16.9%
5.0%	5.2%	-0.2%	0.07%	-2.7	0.6%
5.0%	5.3%	-0.3%	0.07%	-4.1	0.0%
5.0%	5.4%	-0.4%	0.07%	-5.5	0.0%
5.0%	5.5%	-0.5%	0.07%	-6.9	0.0%

Table 4-4: The 95 Percent Confidence Interval for Difference Sizes of the Challenger Group

RESPONSE	SIZE	SEP	95% CONF	LOWER	UPPER	WIDTH
5.0%	1,000	0.6892%	1.96	3.65%	6.35%	2.70%
5.0%	5,000	0.3082%	1.96	4.40%	5.60%	1.21%
5.0%	10,000	0.2179%	1.96	4.57%	5.43%	0.85%
5.0%	20,000	0.1541%	1.96	4.70%	5.30%	0.60%
5.0%	40,000	0.1090%	1.96	4.79%	5.21%	0.43%
5.0%	60,000	0.0890%	1.96	4.83%	5.17%	0.35%
5.0%	80,000	0.0771%	1.96	4.85%	5.15%	0.30%
5.0%	100,000	0.0689%	1.96	4.86%	5.14%	0.27%
5.0%	120,000	0.0629%	1.96	4.88%	5.12%	0.25%
5.0%	140,000	0.0582%	1.96	4.89%	5.11%	0.23%
5.0%	160,000	0.0545%	1.96	4.89%	5.11%	0.21%
5.0%	180,000	0.0514%	1.96	4.90%	5.10%	0.20%
5.0%	200,000	0.0487%	1.96	4.90%	5.10%	0.19%
5.0%	500,000	0.0308%	1.96	4.94%	5.06%	0.12%
5.0%	1,000,000	0.0218%	1.96	4.96%	5.04%	0.09%

$$\frac{0.2\%}{1.96} = \sqrt{\left(\frac{p * (1 - p)}{N} + \frac{(p + d) * (1 - p - d)}{N}\right)}$$

$$0.102\% = \sqrt{\left(\frac{5\% * 95\%}{N} + \frac{5.2\% * (94.8\%)}{N}\right)} = \sqrt{\left(\frac{0.0963}{N}\right)}$$

$$N = \frac{((5\% * 95\%) + (5.2\% * 94.8\%))}{(0.00102)^2}$$

$$= \frac{0.096796}{(0.00102)^2} = 92{,}963$$

Equations 3, 4, and 5

Table 4-5: The Champion-Challenger Data Laid Out for the Chi-Square Test

GROUP	RESPONDERS	NON-RESPONDERS	TOTAL	RESPONSE RATE
Champion	43,200	856,800	900,000	4.80%
Challenger	5,000	95,000	100,000	5.00%
Total	48,200	951,800	1,000,000	4.82%

Table 4-6: Calculating the Expected Values and Deviations from Expected for the Data in Table 4-5

	ACTUAL RESPONSE			EXPECTED RESPONSE		DEVIATION	
	YES	NO	TOTAL	YES	NO	YES	NO
Champion	43,200	856,800	900,000	43,380	856,620	–180	180
Challenger	5,000	95,000	100,000	4,820	95,180	180	–180
Total	48,200	951,800	1,000,000	48,200	951,800		
Overall Proportion	4.82%	95.18%					

$$\text{Chi-square}(x) = \frac{(x - \text{expected}(x))^2}{\text{expected}(x)}$$

Equation 7

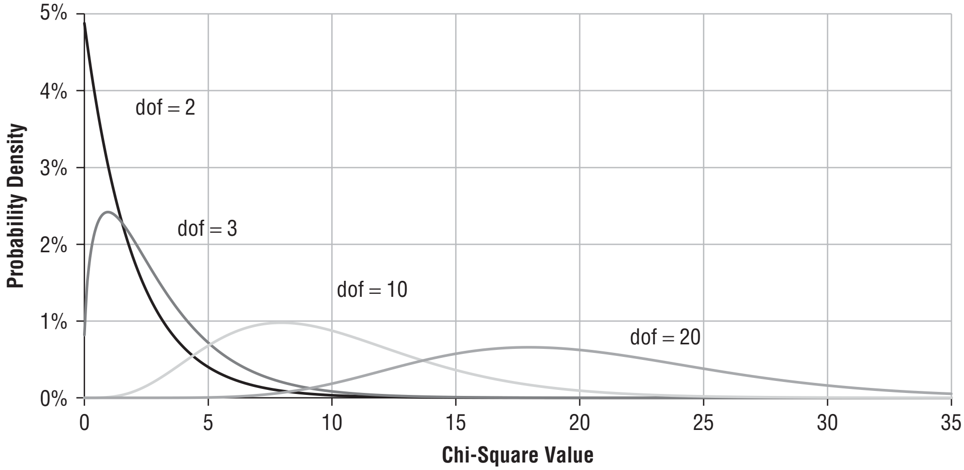


Figure 4-9: The chi-square distribution depends on the degrees of freedom. In general, though, it starts low, peaks early, and gradually descends.

Table 4-7: Chi-Square Calculation for Difference of Proportions Example in Table 4-4

CHALLENGER		CHAMPION			CHALLENGER EXP.		CHAMPION EXP.	
RESP	NON-RESP	RESP	NON-RESP	OVERALL RESP	RESP	NON-RESP	RESP	NON-RESP
5,000	95,000	40,500	859,500	4.55%	4,550	95,450	40,950	859,050
5,000	95,000	41,400	858,600	4.64%	4,640	95,360	41,760	858,240
5,000	95,000	42,300	857,700	4.73%	4,730	95,270	42,570	857,430
5,000	95,000	43,200	856,800	4.82%	4,820	95,180	43,380	856,620
5,000	95,000	44,100	855,900	4.91%	4,910	95,090	44,190	855,810
5,000	95,000	45,000	855,000	5.00%	5,000	95,000	45,000	855,000
5,000	95,000	45,900	854,100	5.09%	5,090	94,910	45,810	854,190
5,000	95,000	46,800	853,200	5.18%	5,180	94,820	46,620	853,380
5,000	95,000	47,700	852,300	5.27%	5,270	94,730	47,430	852,570
5,000	95,000	48,600	851,400	5.36%	5,360	94,640	48,240	851,760
5,000	95,000	49,500	850,500	5.45%	5,450	94,550	49,050	850,950

CHALLENGER		CHALLENGER CHI-SQUARE		CHAMPION CHI-SQUARE		CHI-SQUARE		DIFF. PROP.
RESP	NON-RESP	RESP	NON RESP	RESP	NON RESP	VALUE	P-VALUE	P-VALUE
5,000	95,000	44.51	2.12	4.95	0.24	51.81	0.00%	0.00%
5,000	95,000	27.93	1.36	3.10	0.15	32.54	0.00%	0.00%
5,000	95,000	15.41	0.77	1.71	0.09	17.97	0.00%	0.00%
5,000	95,000	6.72	0.34	0.75	0.04	7.85	0.51%	0.58%
5,000	95,000	1.65	0.09	0.18	0.01	1.93	16.50%	16.83%
5,000	95,000	0.00	0.00	0.00	0.00	0.00	100.00%	100.00%
5,000	95,000	1.59	0.09	0.18	0.01	1.86	17.23%	16. 91%
5,000	95,000	6.25	0.34	0.69	0.04	7.33	0.68%	0.60%
5,000	95,000	13.83	0.77	1.54	0.09	16.23	0.01%	0.00%
5,000	95,000	24.18	1.37	2.69	0.15	28.39	0.00%	0.00%
5,000	95,000	37.16	2.14	4.13	0.24	43.66	0.00%	0.00%

Table 4-8: Chi-Square Calculation for Counties and Channels Example

COUNTY	EXPECTED			DEVIATION			CHI-SQUARE		
	TM	DM	OTHER	TM	DM	OTHER	TM	DM	OTHER
Bronx	1,850.2	523.1	4,187.7	1,362	–110	–1,252	1,002.3	23.2	374.1
Kings	6,257.9	1,769.4	14,163.7	3,515	–376	–3,139	1,974.5	80.1	695.6
Nassau	4,251.1	1,202.0	9,621.8	–1,116	371	745	293.0	114.5	57.7
New York	11,005.3	3,111.7	24,908.9	–3,811	–245	4,056	1,319.9	19.2	660.5
Queens	5,245.2	1,483.1	11,871.7	1,021	–103	–918	198.7	7.2	70.9
Richmond	798.9	225.9	1,808.2	–15	51	–36	0.3	11.6	0.7
Suffolk	3,133.6	886.0	7,092.4	–223	156	67	15.8	27.5	0.6
Westchester	3,443.8	973.7	7,794.5	–733	256	477	155.9	67.4	29.1

Table 4-9: Chi-Square Calculation for Bronx and TM

	EXPECTED		DEVIATION		CHI-SQUARE	
COUNTY	TM	NOT_TM	TM	NOT_TM	TM	NOT_TM
Bronx	1,850.2	4,710.8	1,361.8	–1,361.8	1,002.3	393.7
Not Bronx	34,135.8	86,913.2	–1,361.8	1,361.8	54.3	21.3

Table 4-10: Estimated P-Value for Each Combination of County and Channel, without Correcting for Number of Comparisons

COUNTY	TM	DM	OTHER
Bronx	0.00%	0.00%	0.00%
Kings	0.00%	0.00%	0.00%
Nassau	0.00%	0.00%	0.00%
New York	0.00%	0.00%	0.00%
Queens	0.00%	0.74%	0.00%
Richmond	59.79%	0.07%	39.45%
Suffolk	0.01%	0.00%	42.91%
Westchester	0.00%	0.00%	0.00%

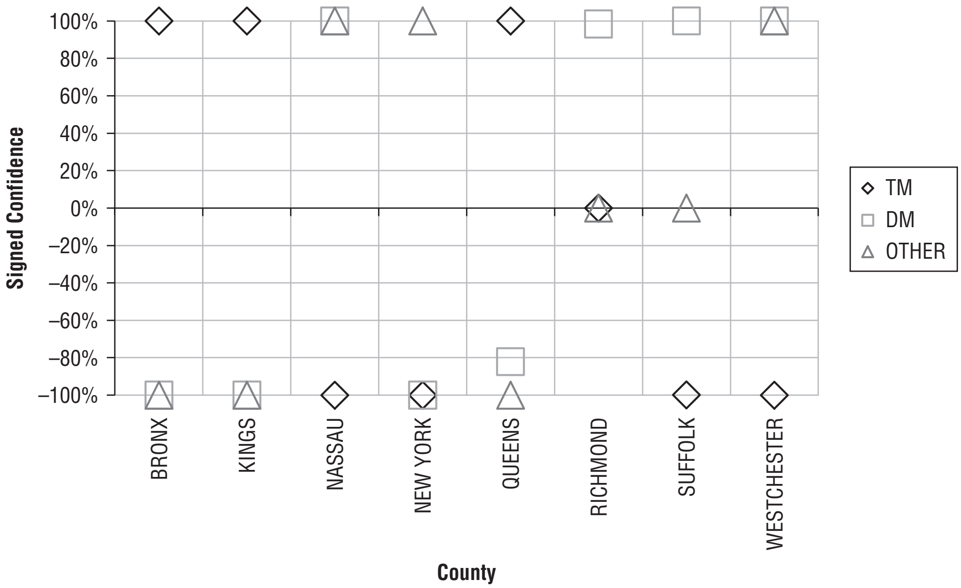


Figure 4-10: This chart shows the signed confidence values for each county and region combination; the preponderance of values near 100% and –100% indicate that observed differences are statistically significant.

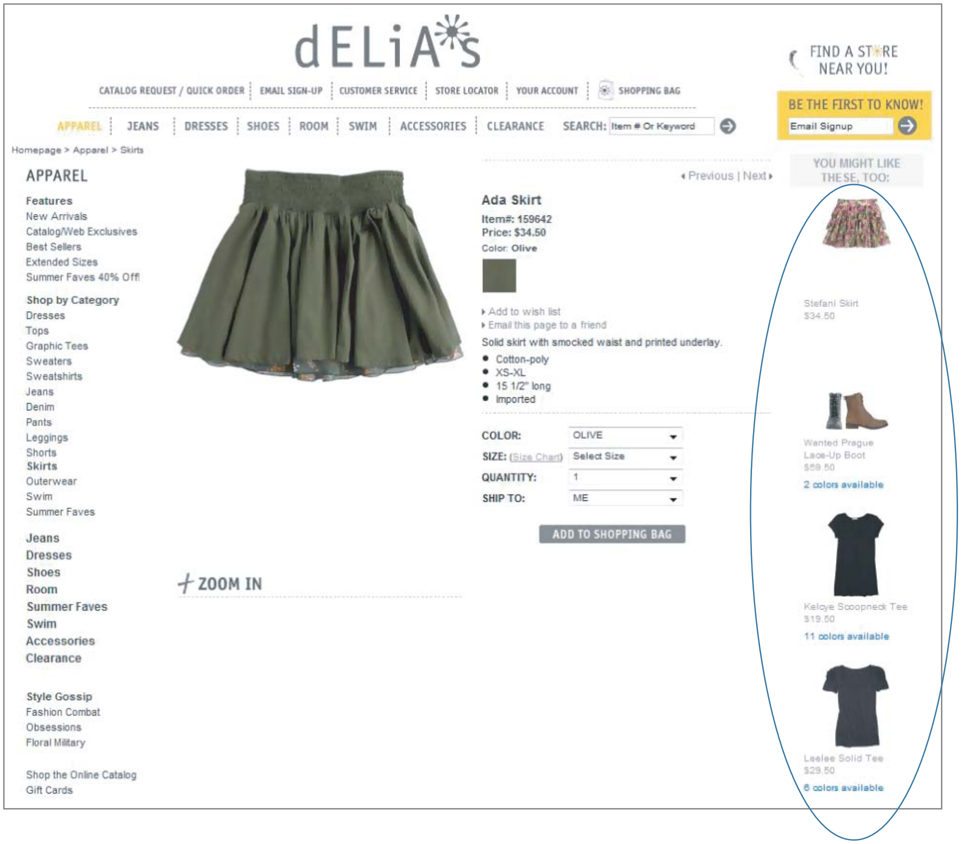


Figure 4-11: This screen shot shows an example of a site using MyBuys recommendations.

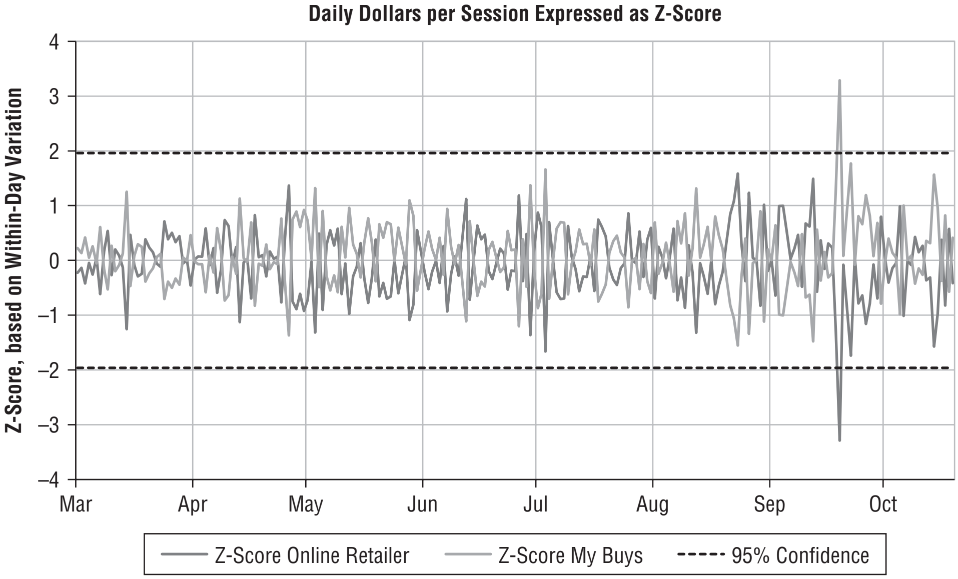


Figure 4-12: The daily revenue for both sides of the A/B tests is usually within the 95% confidence bounds and does not obviously favor one side of the test over the other.

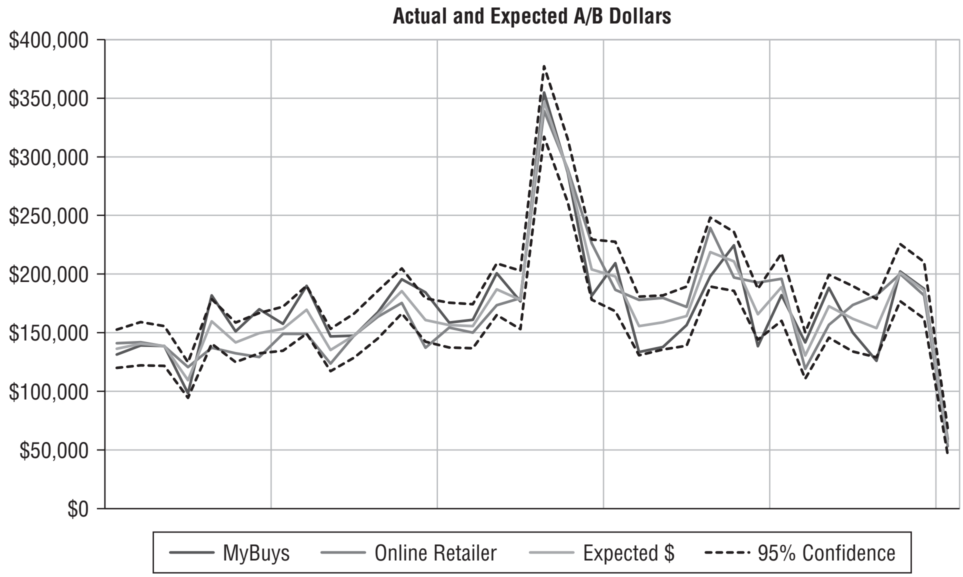


Figure 4-13: Using dollar amounts provides more information about what is happening over time, in terms of sales. The data here is similar to Figure 4-12, but for a shorter time frame.

$$\text{SEM} = \frac{\text{standard deviation}}{\sqrt{\text{sample size}}}$$

Equation 8

Figure 4-14: A time series of product sales and inventory illustrates the problem of censored data.

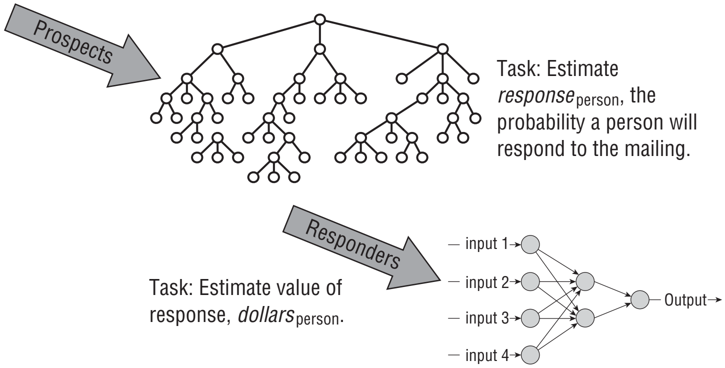


Figure 5-1: This is an example of a two-step model for estimating response amounts. The first model predicts response; the second estimates the amount of the response. The product is the expected response amount.

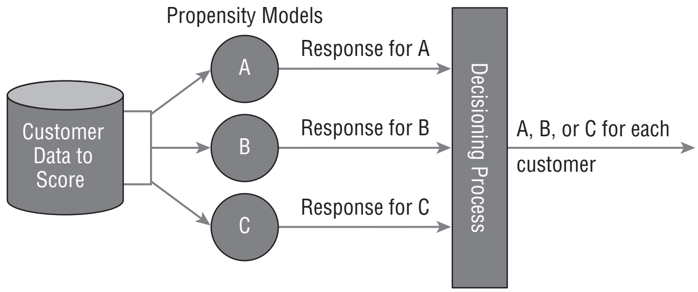
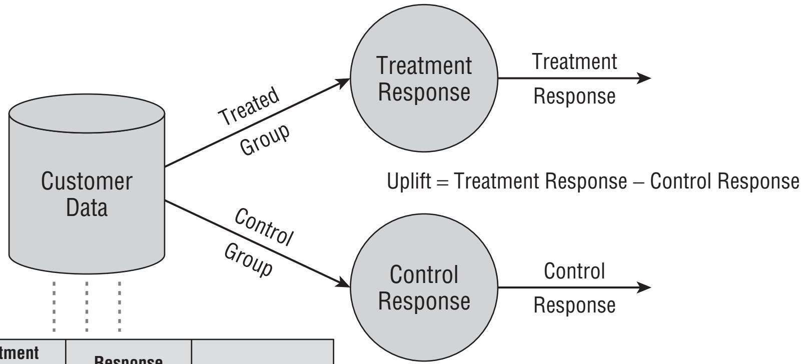


Figure 5-2: A cross-sell model for a handful of options consists of a separate model for each option along with a decision function for choosing the optimal option.



Customer ID	Treatment Group	Response	...
1	Treated	Y	
2	Treated	N	
3	Control	Y	
4	Treated	N	
...	...	...	

Figure 5-3: An incremental response model can be approximated using two different models — one to estimate the response with no intervention and the other to estimate the response with the intervention.

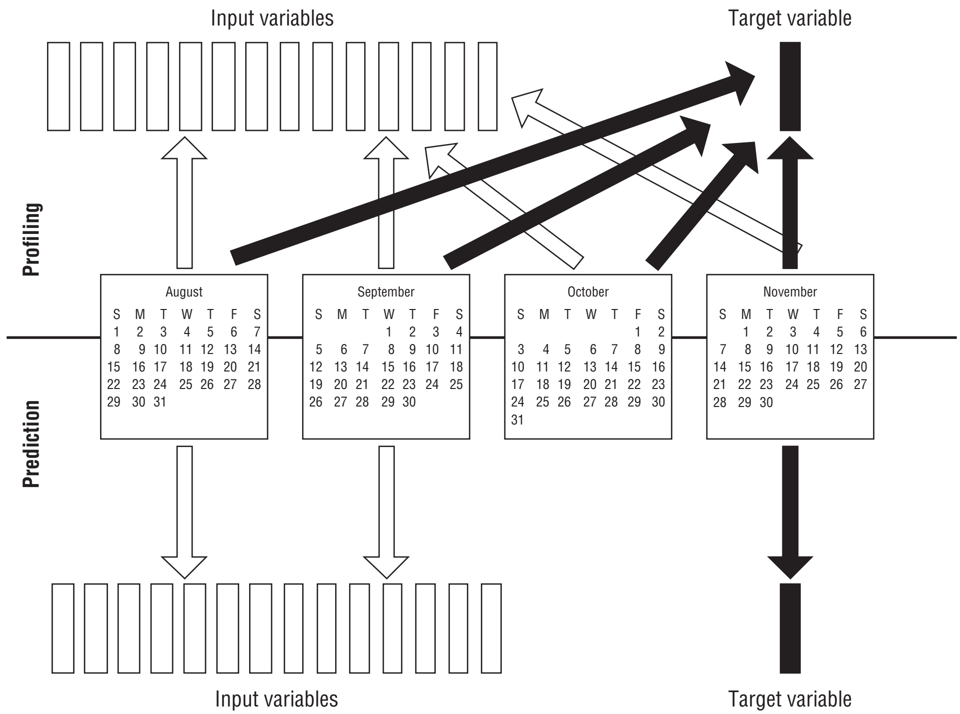


Figure 5-4: Profiling models and prediction models differ only in the temporal relationship of the target variable to the input variables.

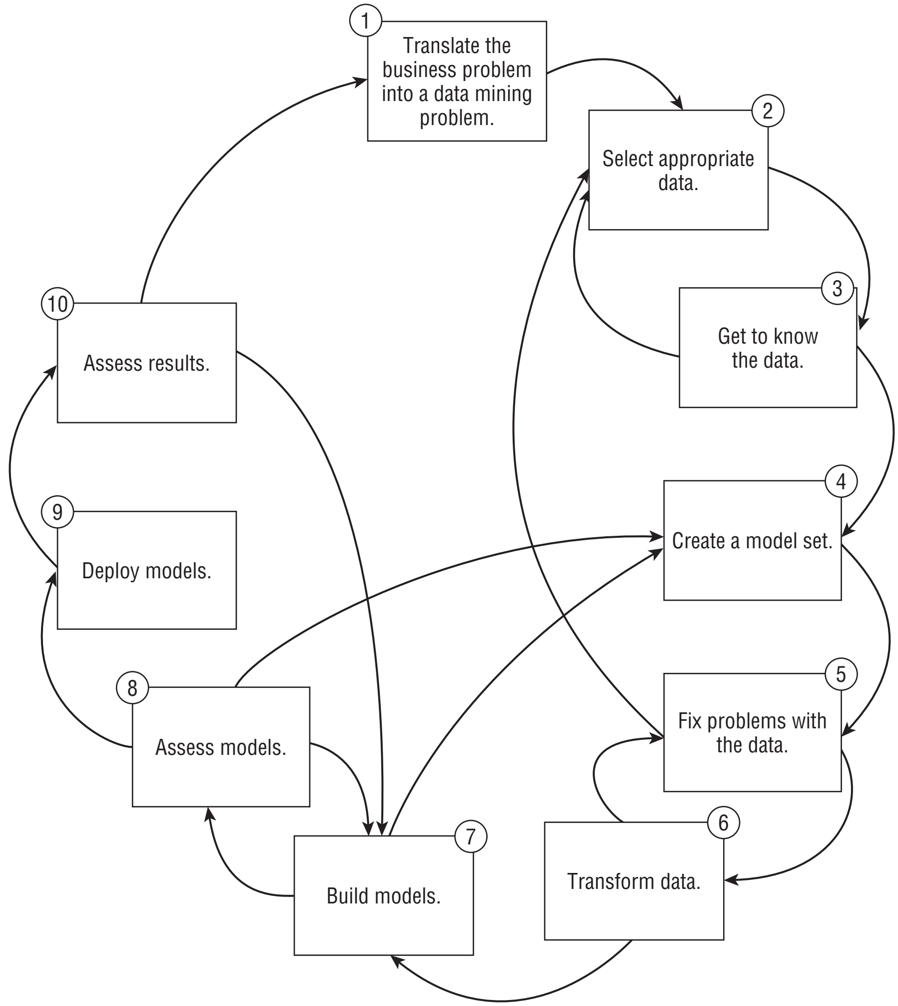
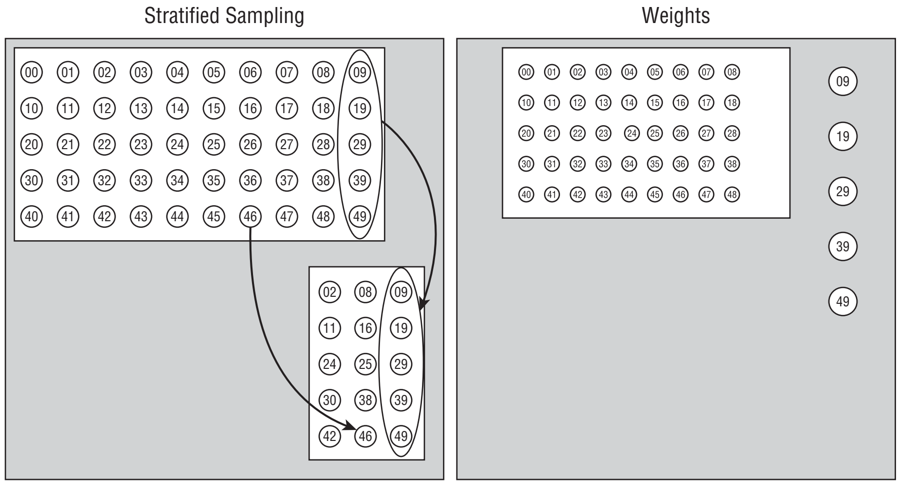


Figure 5-5: Directed data mining is not a linear process.



When an outcome is rare, there are two ways to create a balanced sample.

Two Ways to Create a Balanced Sample

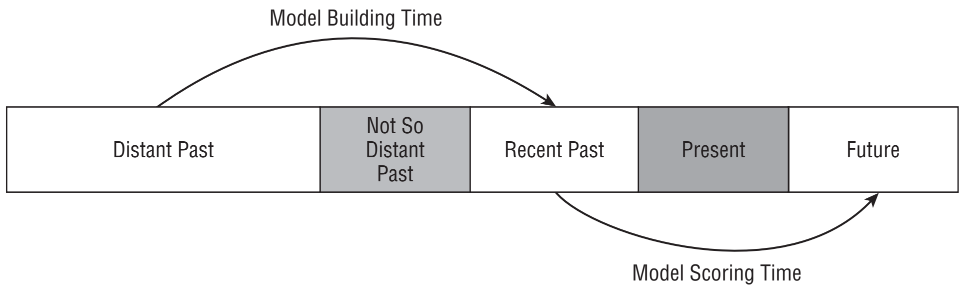


Figure 5-6: Data from the past mimics data from the past, present, and future.

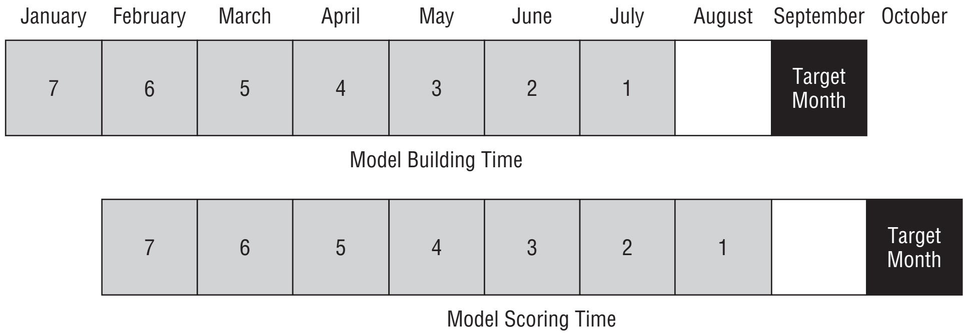


Figure 5-7: Time when the model is built compared to time when the model is used.

Predicted	Actual	
	YES	NO
YES	1,000	200
NO	600	900

Predicted	Actual	
	YES	NO
YES	# CORRECT POSITIVE	# FALSE POSITIVE
NO	# FALSE NEGATIVE	# CORRECT NEGATIVE

Figure 5-8: A confusion matrix cross-tabulates predicted outcomes with actual outcomes.

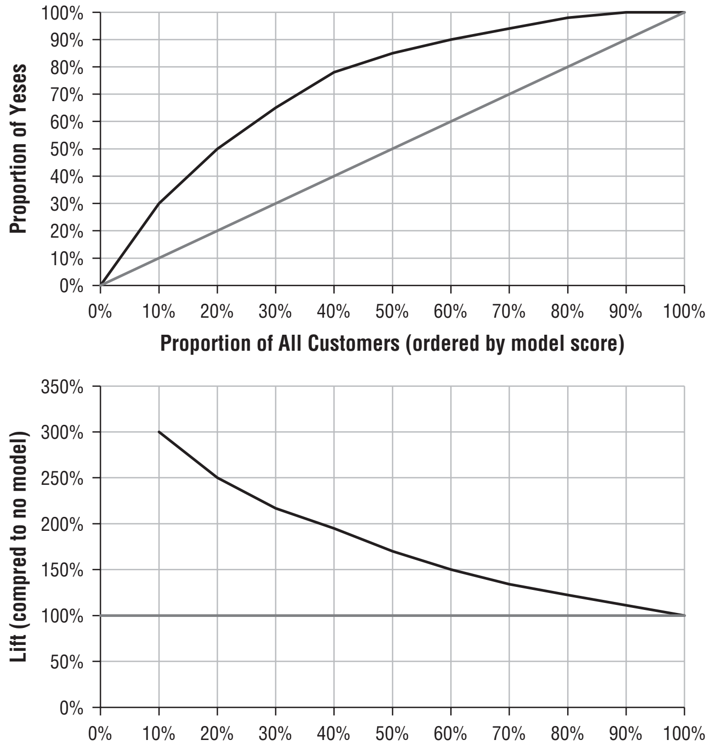


Figure 5-9: The top part of this chart shows the cumulative gains for a binary response model. The lower chart shows the lift (cumulative ratio by decile). A lift chart starts high and descends to 1.

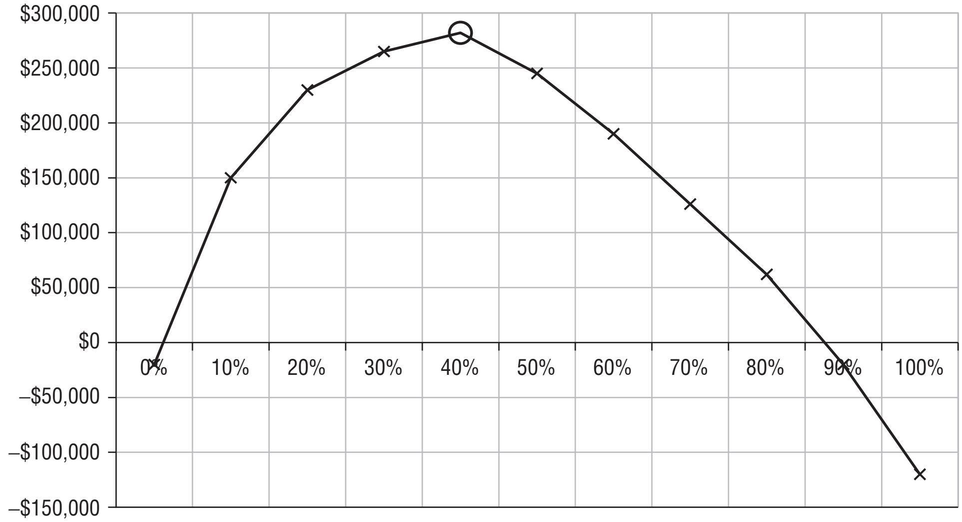


Figure 5-10: A profitability curve translates model results into dollars and cents, making it possible to optimize the model based on financial gain. In this case, maximum profitability occurs when contacting the top 40% of people chosen by the model.

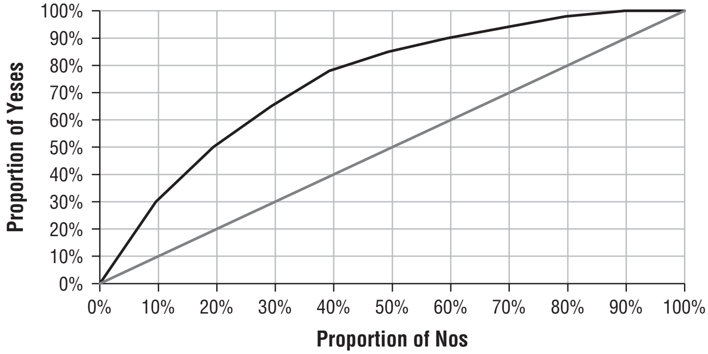


Figure 5-11: An ROC chart looks very similar to a cumulative gains chart, but the horizontal axis is the proportion of false positives, rather than the proportion of the overall population.

Table 5-1: Errors cancel each other out (the sum of the error column is zero)

TRUE VALUE	ESTIMATED VALUE	ERROR
127	132	–5
78	76	2
120	122	–2
130	129	1
95	91	4

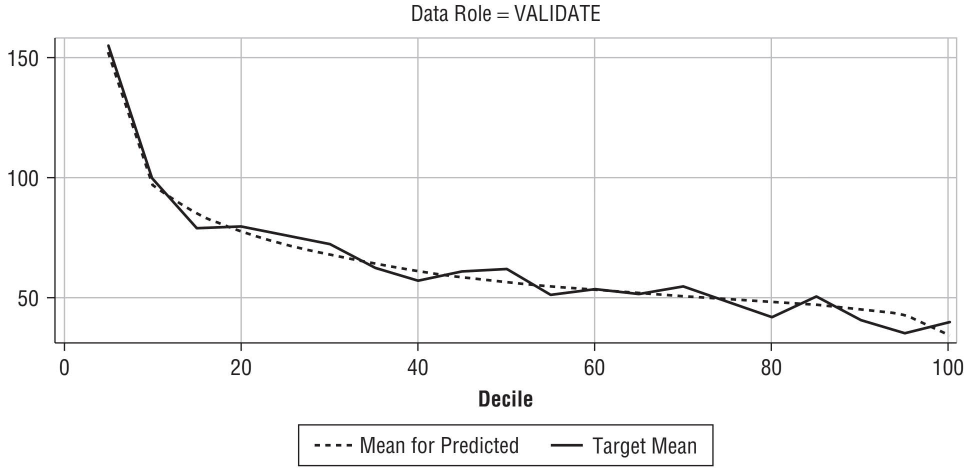


Figure 5-12: This example of a Score Ranking Chart from SAS Enterprise Miner compares the average values of the target variable with the average value of the prediction, by decile (or other grouping).

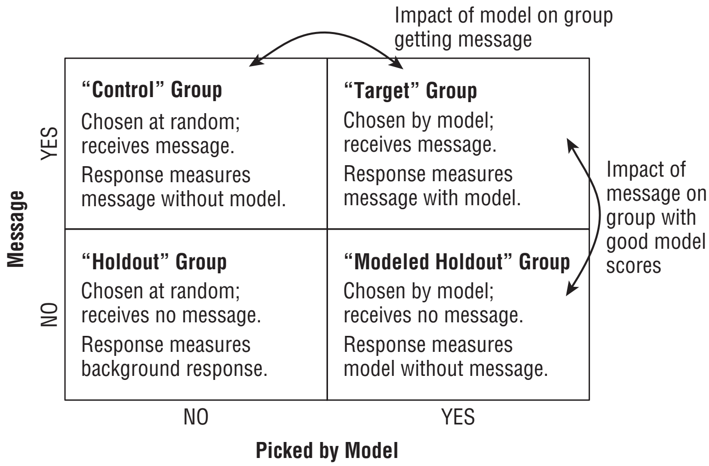


Figure 5-13: When you deploy a campaign, four different treatment groups exist. Comparisons between the groups yield different insights.

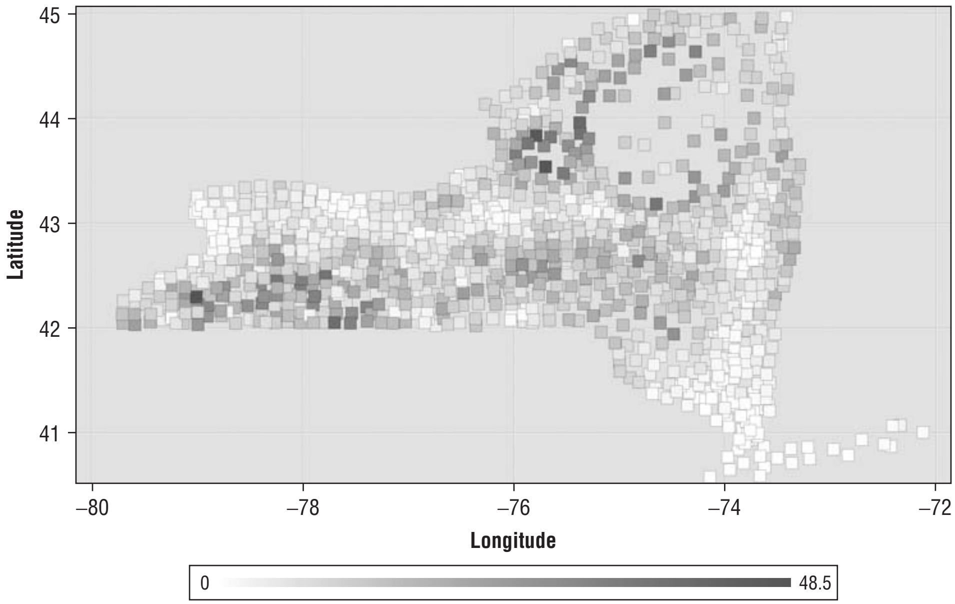


Figure 6-1: This scatter plot is based on the latitude and longitude of towns in New York state; the shading is based on the proportion of the town with wood-burning stoves.

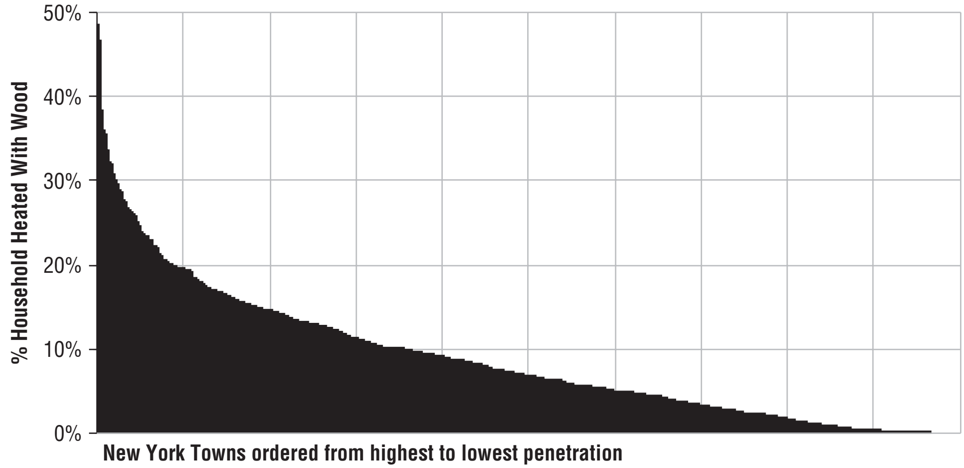


Figure 6-2: The percentage of households in a town heated by wood ranges from near 50 percent to 0.

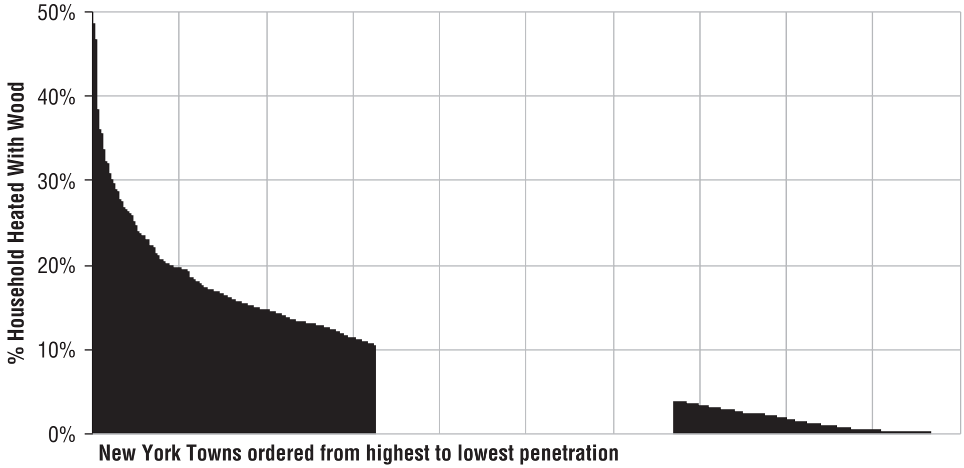


Figure 6-3: Removing towns in the middle of the range sharpens the contrast between high penetration and low penetration.

Table 6-1: Variables with significantly different averages in high- and low-penetration towns.

	WORKING IN AGRICULTURE	MULTI-FAMILY HOMES	MEDIAN HOME VALUE
Low Penetration	1.4%	26.3%	$136,296
High Penetration	6.6%	4.7%	$67,902

Table 6-2: Averages and standard deviations for the selected variables.

	WORKING IN AGRICULTURE	MULTI-FAMILY HOMES	MEDIAN HOME VALUE
Average	3.9%	14.2%	$95,256
Standard Deviation	3.9%	14.8%	$70,754
Ideal	10.0%	0.0%	$60,000

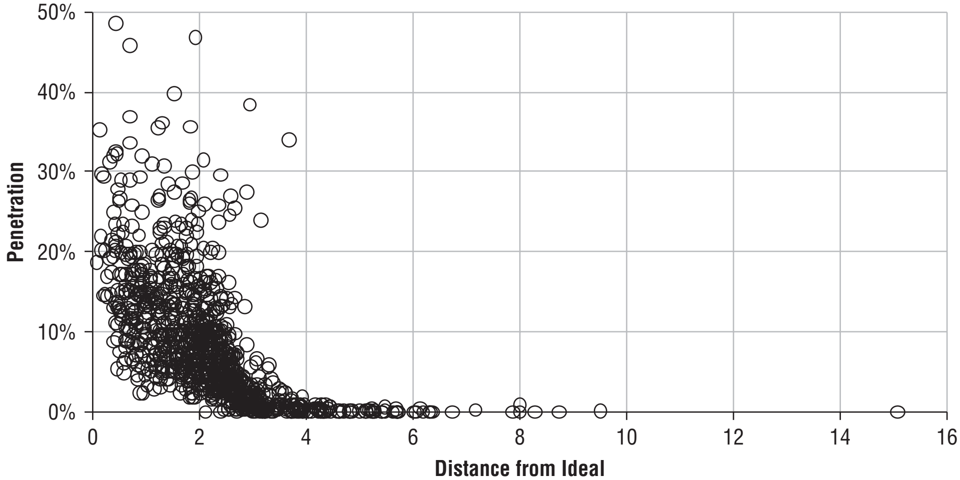


Figure 6-4: As distance from the ideal increases, penetration quickly drops to zero.

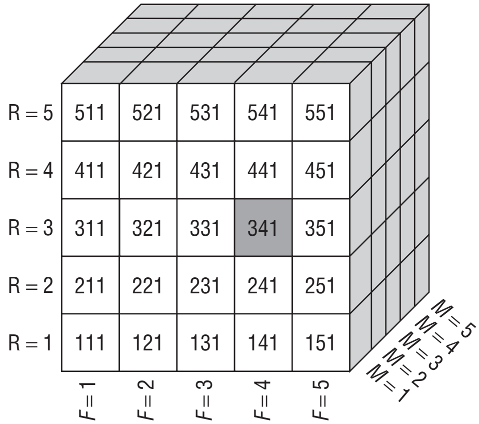


Figure 6-5: Each of the three RFM dimensions has been partitioned into quintiles to form an RFM cube with 125 cells.

$$P(A|B) = P(B|A)\frac{P(A)}{P(B)}$$

Equation 9

$$odds = \frac{probability}{1 - probability} = -1 + \frac{1}{1 - probability}$$

Equation 10

$$probability = 1 - \frac{1}{1 + odds}$$

Equation 11

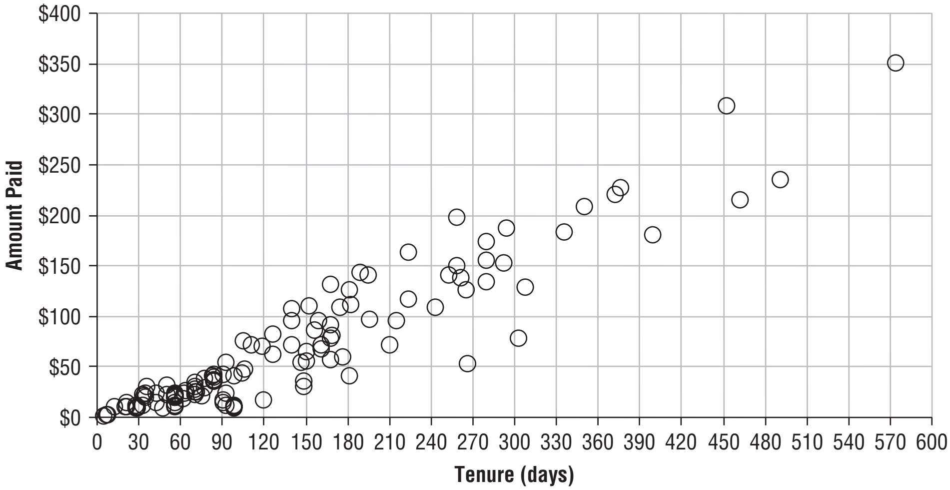


Figure 6-6: The scatter plot shows the relationship between tenure and total amount paid.

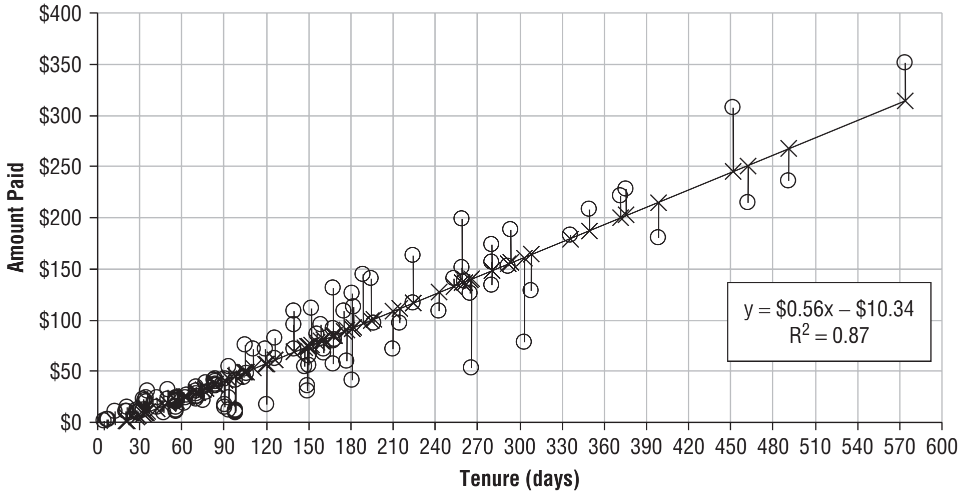


Figure 6-7: The best-fit line minimizes the square of the vertical distance from the observations to the line.

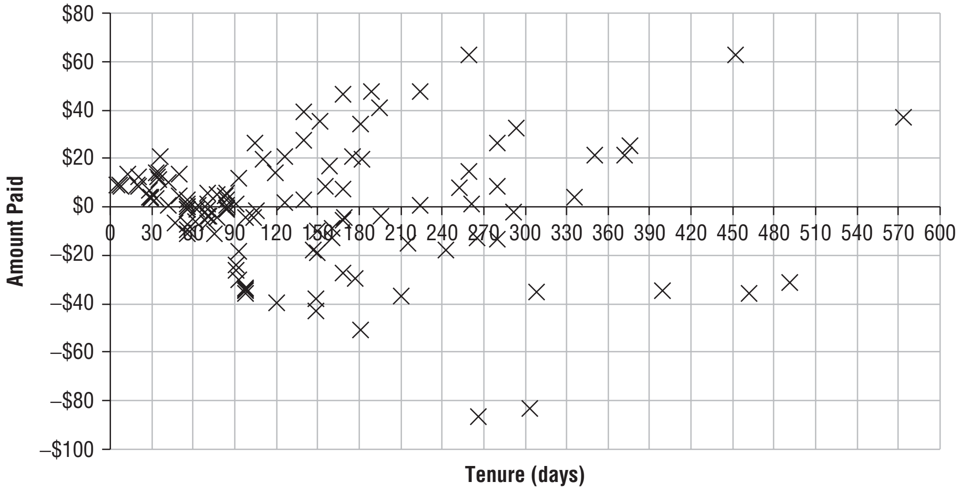


Figure 6-8: There are about as many positive as negative residuals and they do not show any strong patterns.

$$Y = \beta_0 + \beta_1 X_1$$

Equation 12

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

Equation 13

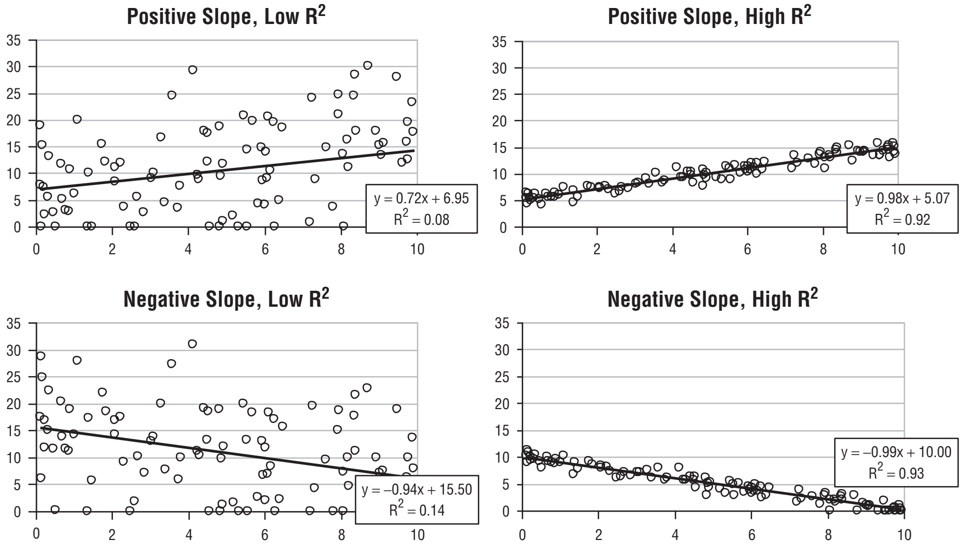


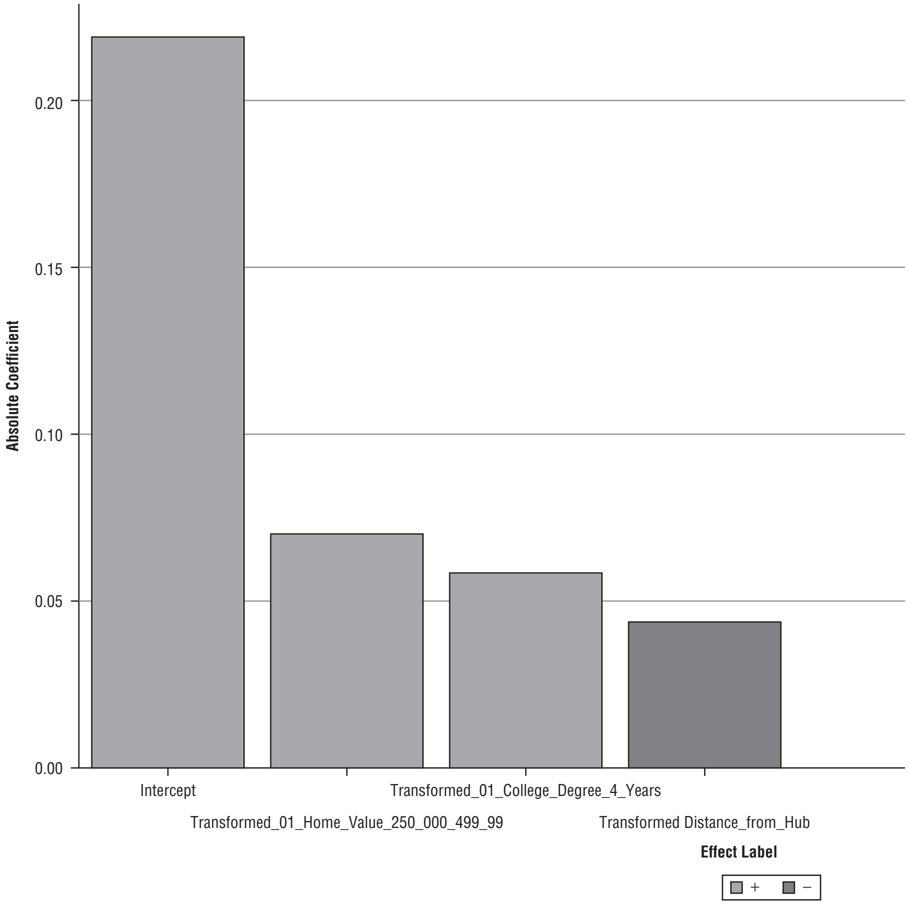
Figure 6-9: R^2 and trend are two ways of characterizing the best-fit line. A high R^2 value implies that the points are very close to the line.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

Equation 14

$$Y = \beta_0 + \beta_1 X_1$$

Equation 15



Parameter	DF	Estimate	Standard Error	t Value	Pr > | t |
Intercept	1	0.2188	0.00475	46.12	< .0001
STD_Distance_from_Hub	1	–0.0437	0.00532	–8.21	< .0001
STD_01_College_Degree_4_Years	1	0.0584	0.00672	8.70	< .0001
STD_01_Home_Value_2500_000_499	1	0.0701	0.00677	10.36	< .0001

Figure 6-10: The height of the bars shows the relative importance of the inputs.

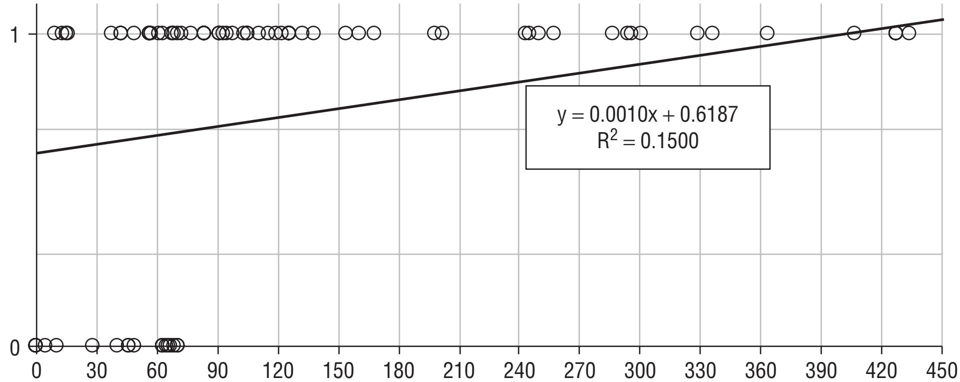


Figure 6-11: A linear regression model does a poor job of modeling the probability that a subscriber has ever paid.

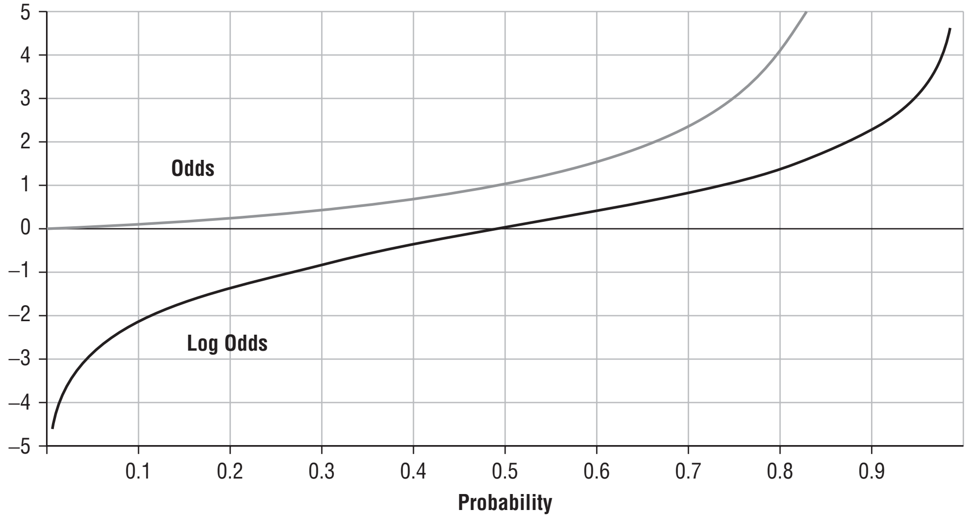


Figure 6-12: A comparison of odds and log odds. The log odds function is symmetrical around 0 and goes from negative infinity to positive infinity.

$$\ln\left(\begin{matrix} p \\ -p \end{matrix}\right) = \beta_0 + \beta_1 X.$$

Equation 16

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}.$$

Equation 17

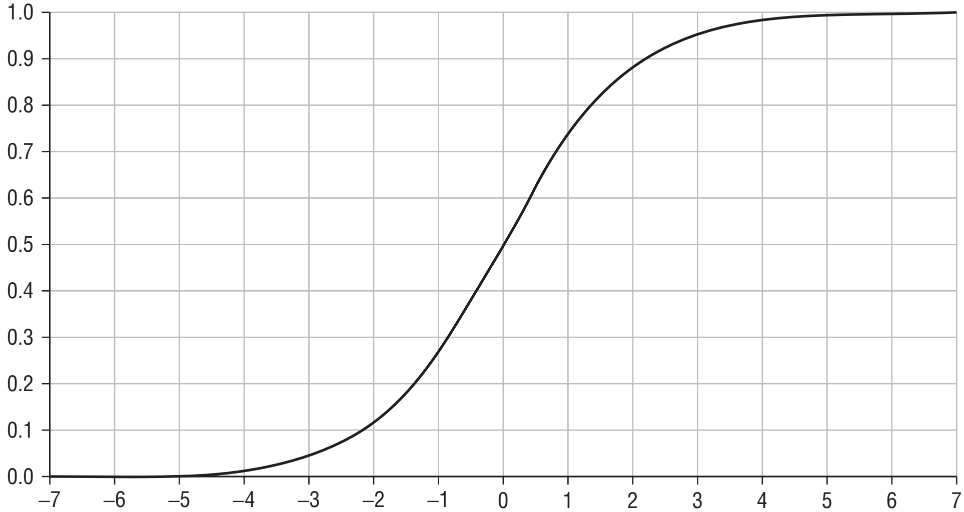


Figure 6-13: The logistic function goes from 0 to 1 just like a probability.

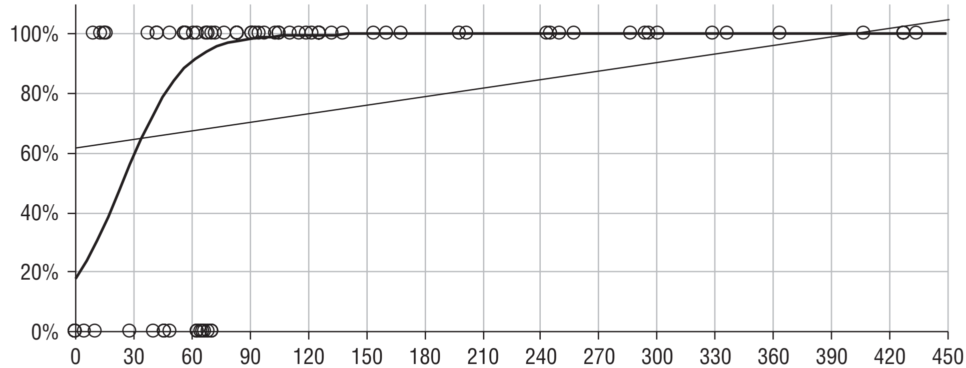


Figure 6-14: Logistic regression does a much better job of estimating the probability that a subscriber has paid.

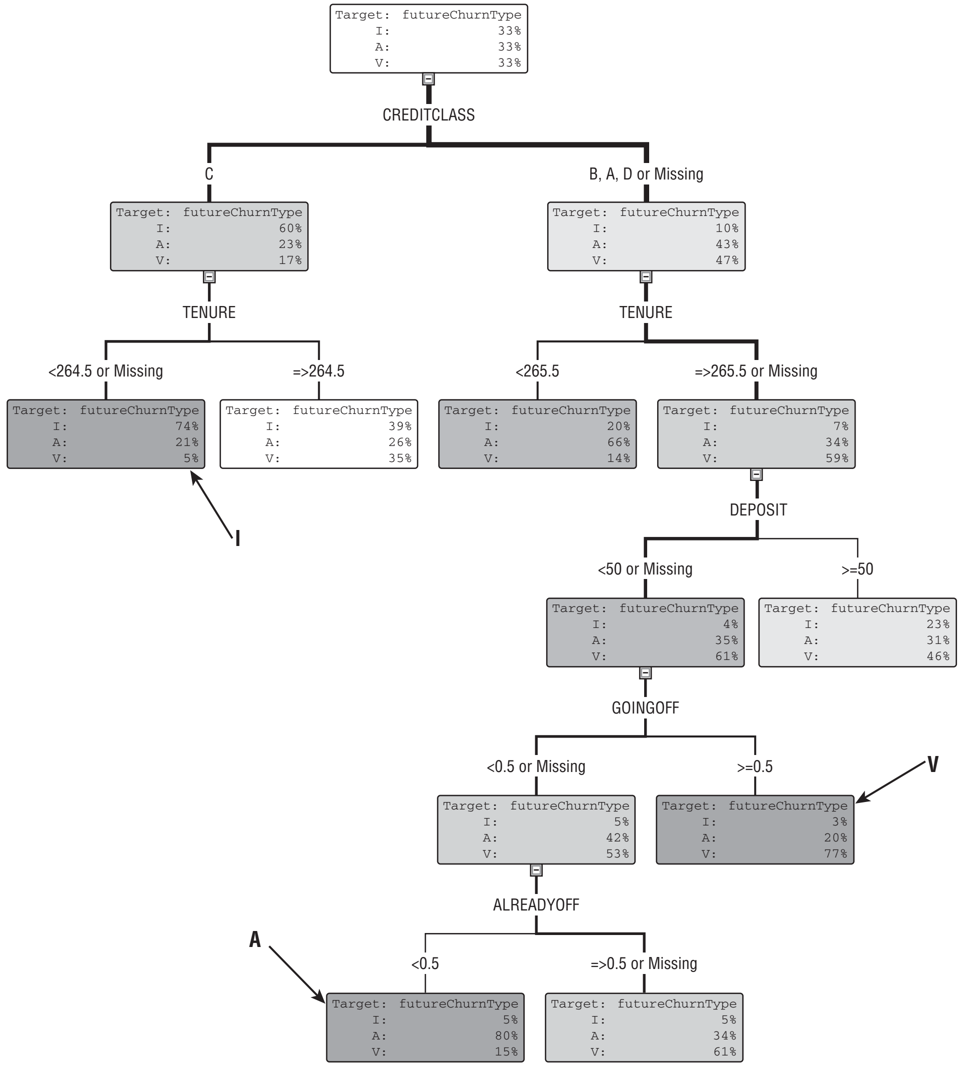


Figure 7-1: A decision tree.

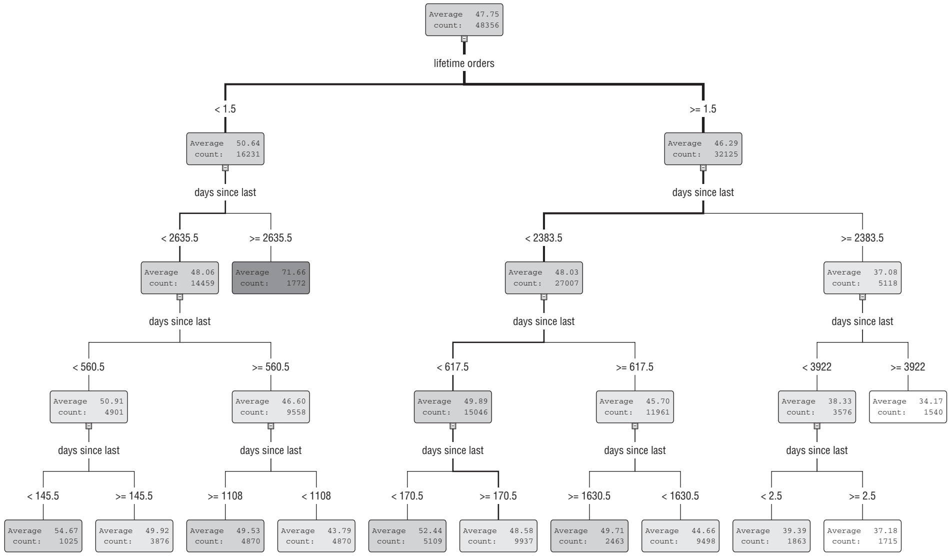


Figure 7-2: A regression tree for average order size as a function of recency and frequency.

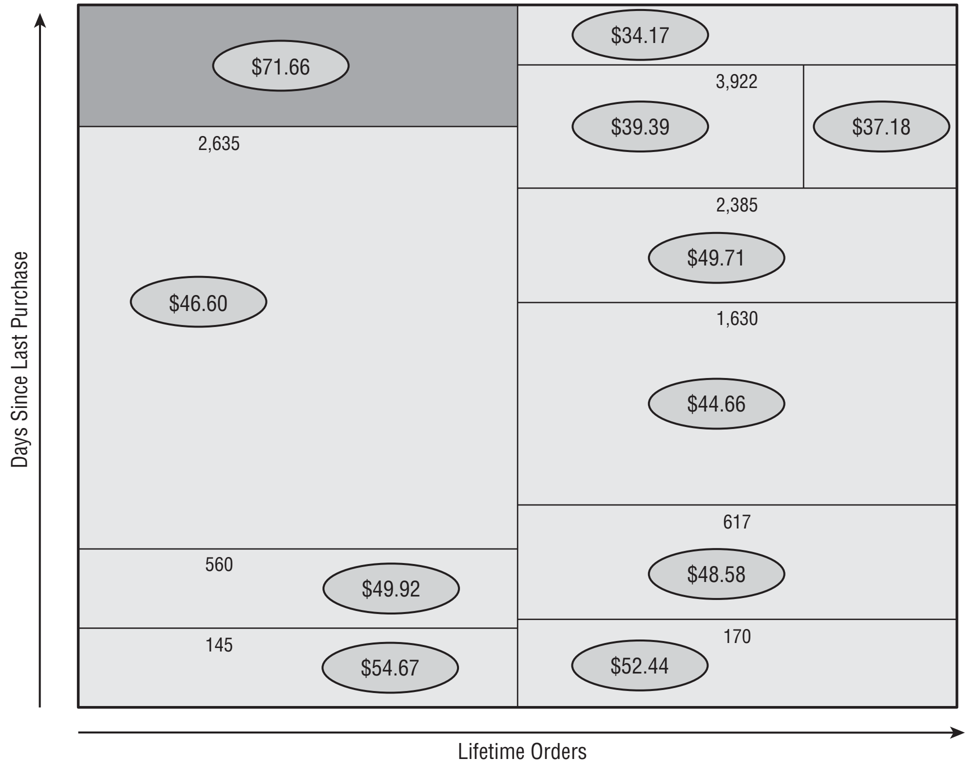


Figure 7-3: The tree puts the records into rectangular boxes.

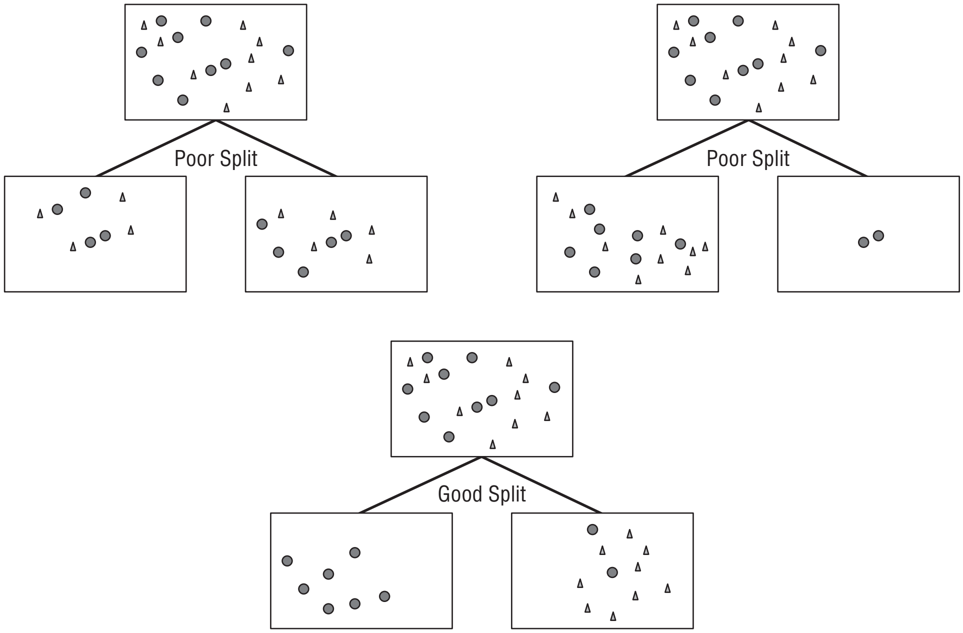


Figure 7-4: A good split increases purity for all the children.

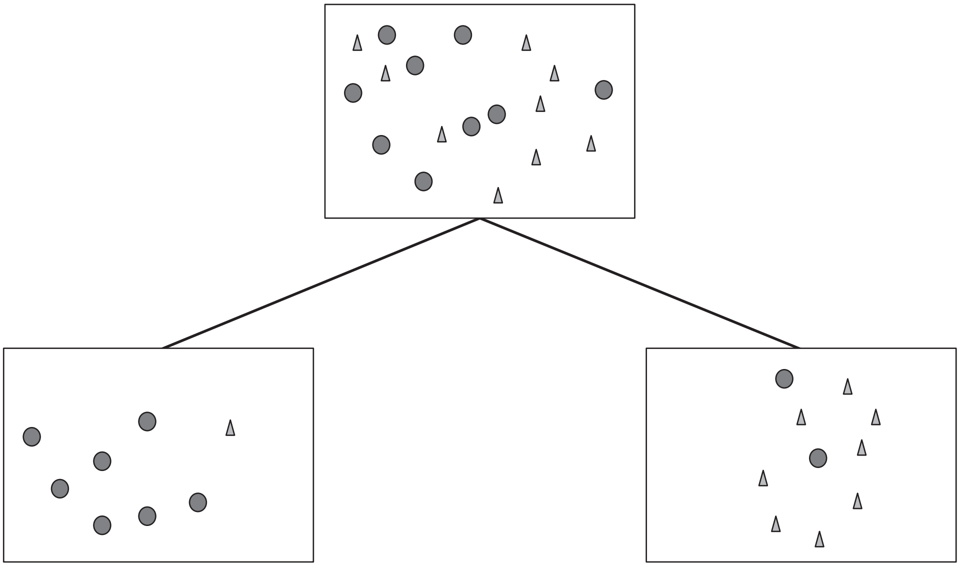


Figure 7-5: A good split on a binary categorical variable increases purity.

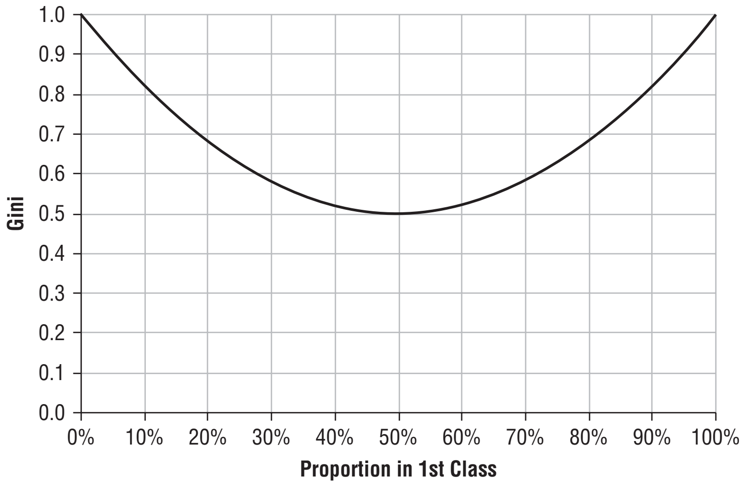


Figure 7-6: For a binary target, the Gini score varies from 0.5 when there is an equal number of each class to 1 when all records are in the same class.

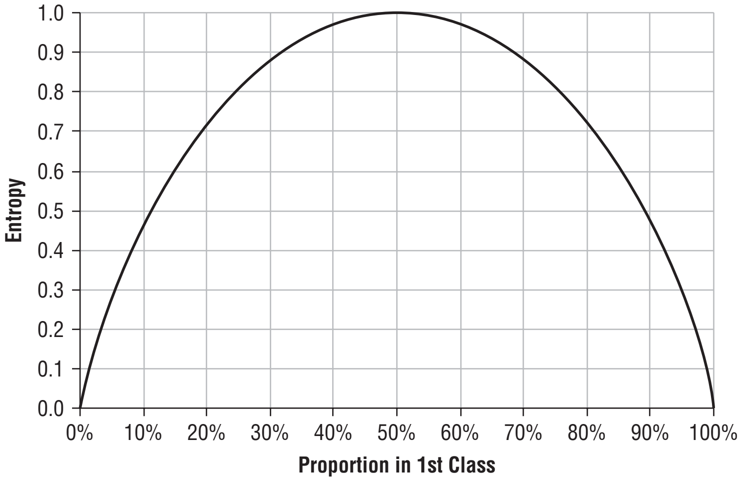


Figure 7-7: Entropy goes from 0 for a pure population to 1 when there is an equal number of each class.

$$-1 * (P(\text{circle})\log_2 \text{P}(\text{circle}) + P(\text{triangle})\log_2 \text{P}(\text{triangle}))$$

Equation 18

$$-1 * (0.875 \log_2(0.875) + 0.125 \log_2(0.125)) = 0.544$$

Equation 19

$$-1 * (0.200 \log_2(0.200) + 0.800 \log_2(0.800)) = 0.722$$

Equation 20

Table 7-1: Contingency Table for Split Evaluation

	RESPONSE = 0	RESPONSE = 1
Left Child	# of 0s on left	# of 1s on left
Right Child	# of 0s on right	# of 1 on right

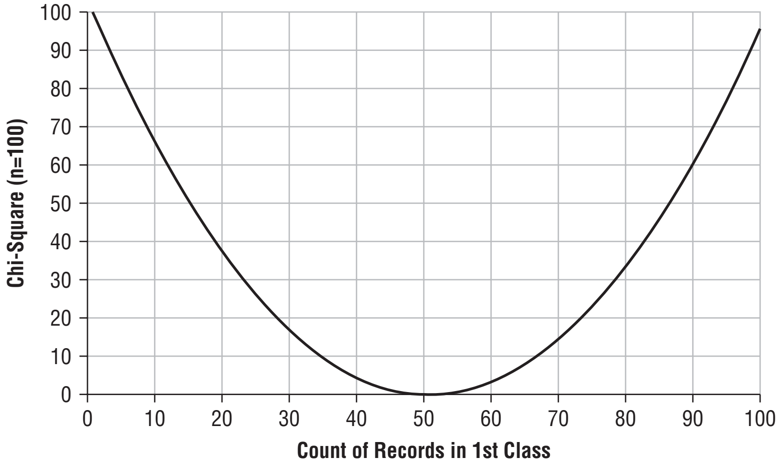
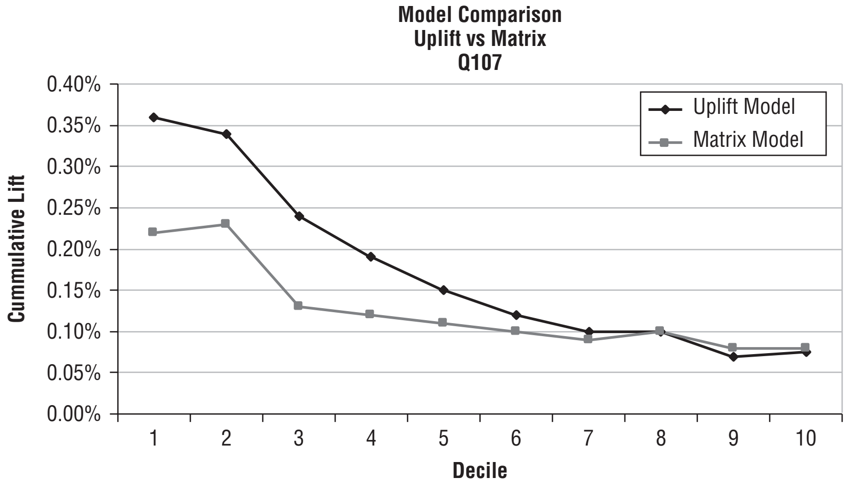


Figure 7-8: Chi-square is 0 when the sample distribution is the same as the population's.



Champion-Challenger comparison.

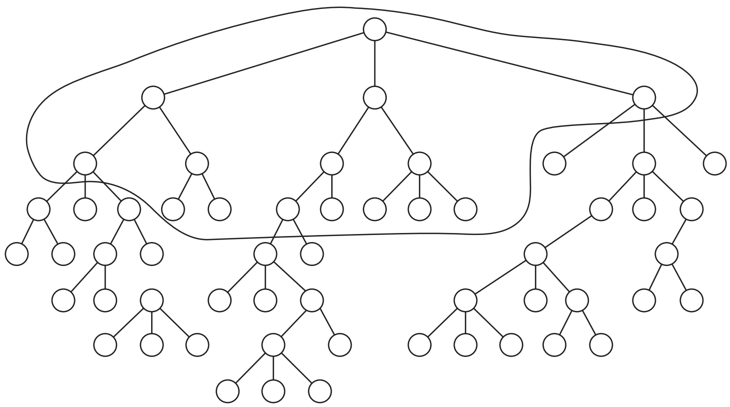


Figure 7-9: Inside a complex tree are simpler, more stable trees.

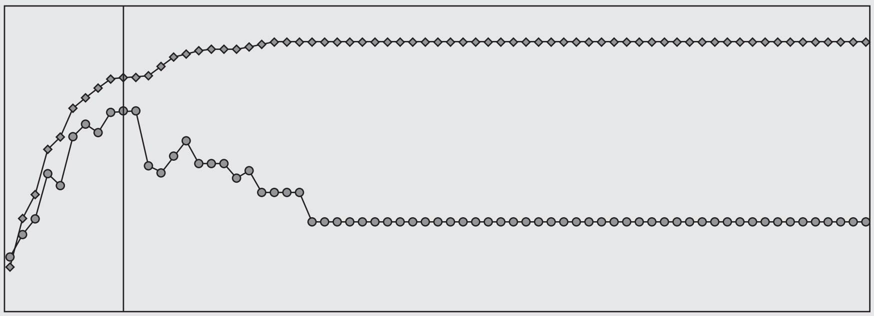
$$AE(T) = E(T) + \alpha leaf\_count(T)$$

Equation 21

COMPARING MISCLASSIFICATION RATES ON TRAINING AND VALIDATION SETS

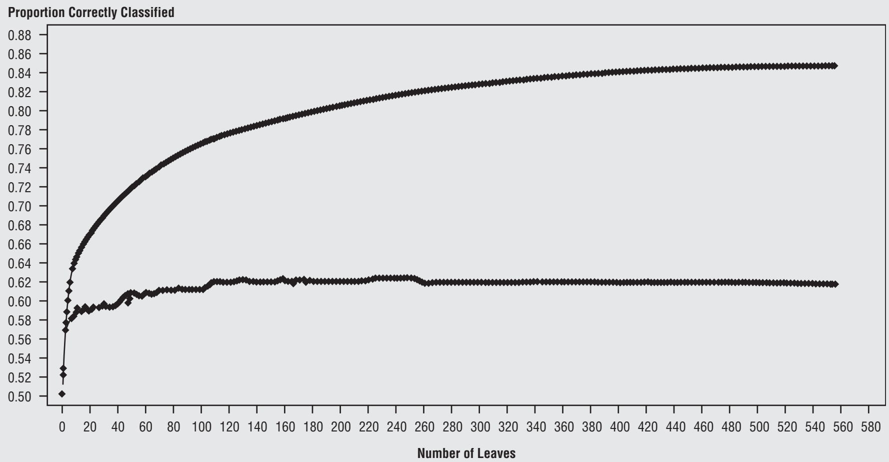
The error rate on the validation set should be larger than the error rate on the training set, because the training set was used to build the rules in the model. A large difference in the misclassification error rate, however, is a symptom of an unstable model. This difference can show up in several ways as shown by the following three graphs. The graphs represent the percent of records correctly classified by the candidate models in a decision tree. Candidate subtrees with fewer nodes are on the left; those with more nodes are on the right.

As expected, the first chart shows the candidate trees performing better and better on the training set as the trees have more and more nodes — the training process stops when the performance no longer improves. On the validation set, however, the candidate trees reach a peak and then the performance starts to decline as the trees get larger. The optimal tree is the one that works best on the validation set, and the choice is easy because the peak is well-defined.



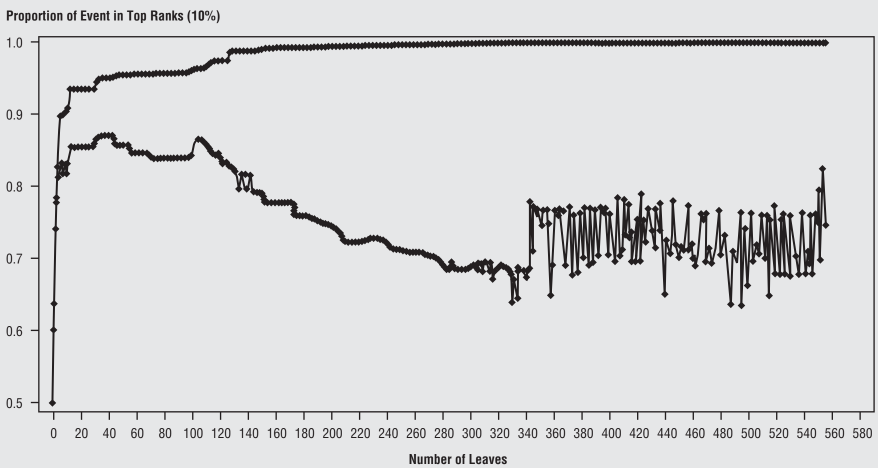
This chart shows a clear inflection point in the graph of the percent correctly classified in the validation set.

Sometimes, though, there is no clear demarcation point. That is, the performance of the candidate models on the validation set never quite reaches a maximum as the trees get larger. In this case, the pruning algorithm chooses the entire tree (the largest possible subtree), as shown.



In this chart, the percent correctly classified in the validation set levels off early and remains far below the percent correctly classified in the training set.

The final example is perhaps the most interesting, because the results on the validation set become unstable as the candidate trees become larger. The cause of the instability is that the leaves are too small. In this tree, there is an example of a leaf that has three records from the training set and all three have a target value of 1 — a perfect leaf. However, in the validation set, the one record that falls there has the value 0. The leaf is 100 percent wrong. As the tree grows more complex, more of these too-small leaves are included, resulting in the instability shown:



In this chart, the percent correctly classified on the validation set decreases with the complexity of the tree and eventually becomes chaotic.

The last two figures are examples of unstable models. The simplest way to avoid instability of this sort is to ensure that leaves are not allowed to become too small.

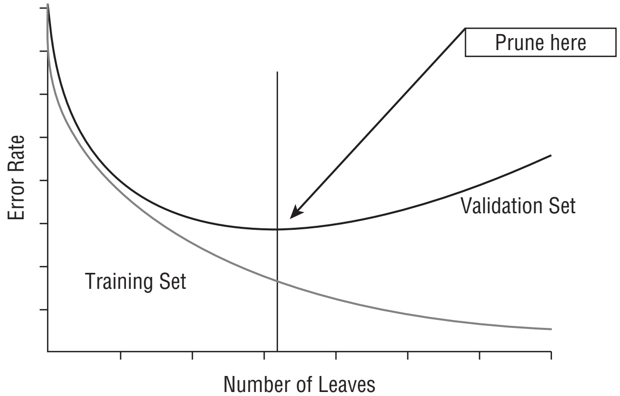


Figure 7-10: Pruning chooses the tree whose miscalculation rate is minimized on the validation set.

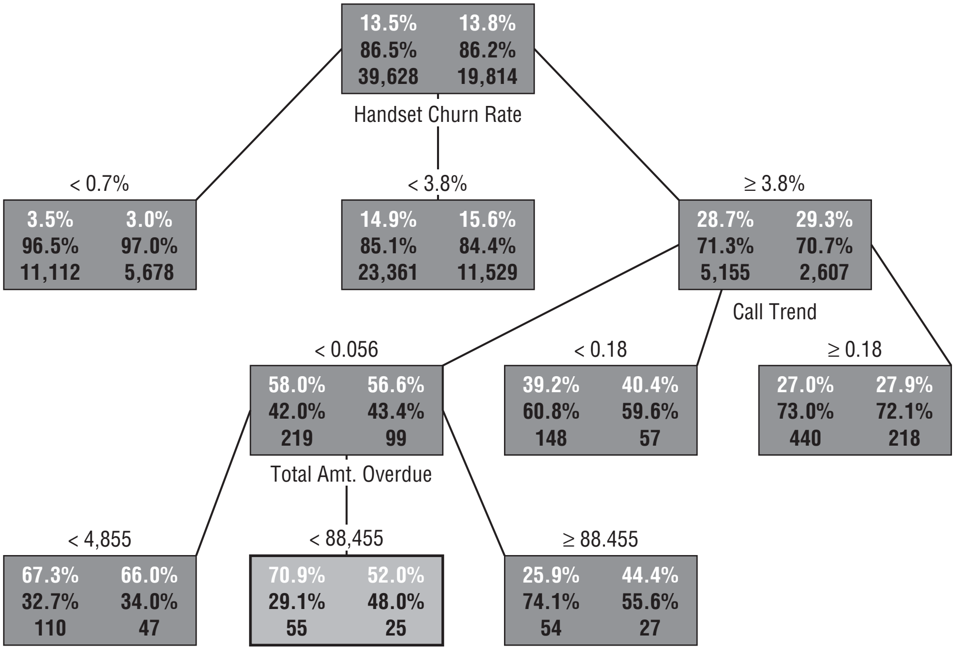


Figure 7-11: An unstable split produces very different distributions on the training and validation sets.

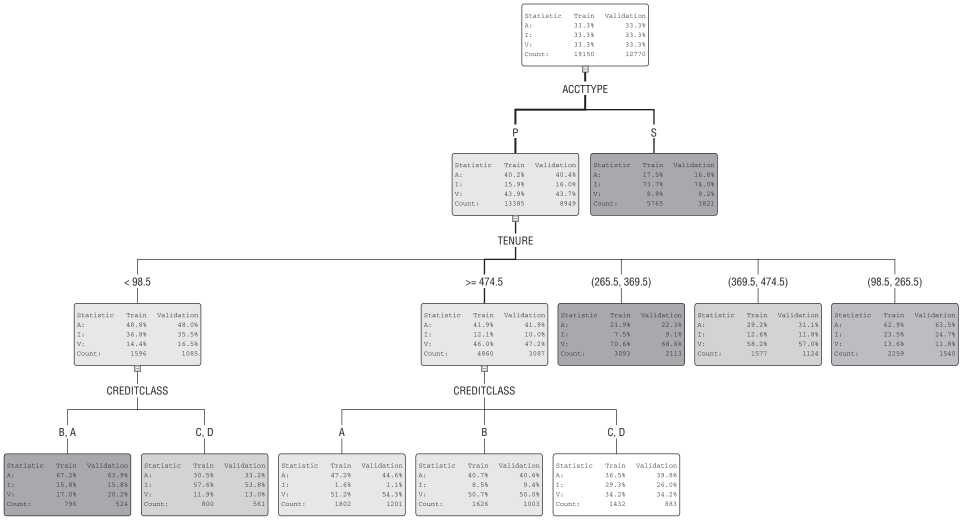


Figure 7-12: This tree with multiway splits does not perform as well as the binary tree in Figure 7-1.

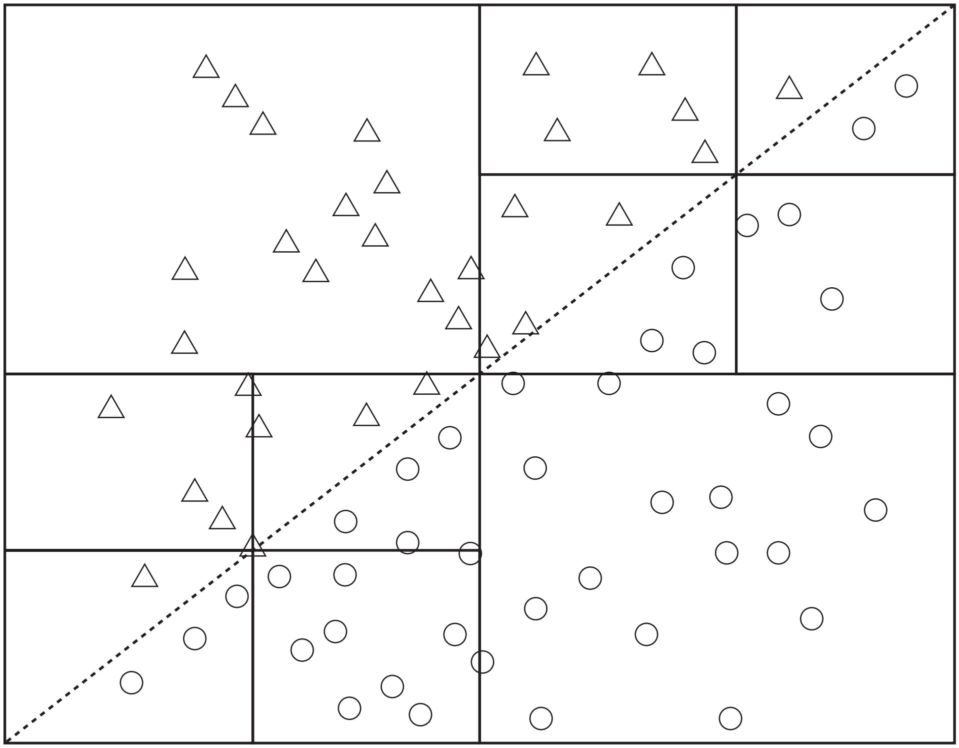
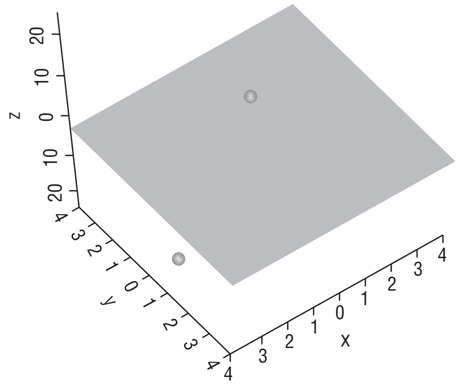
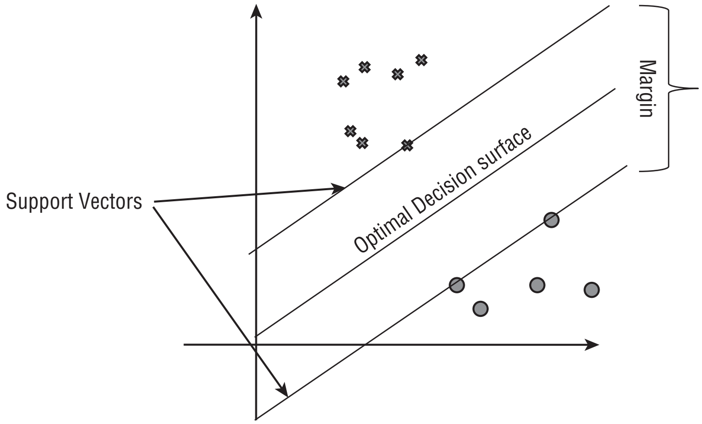


Figure 7-13: The upper-left and lower-right quadrants are easily classified, whereas the other two quadrants must be carved into many small boxes to approximate the boundary between regions.



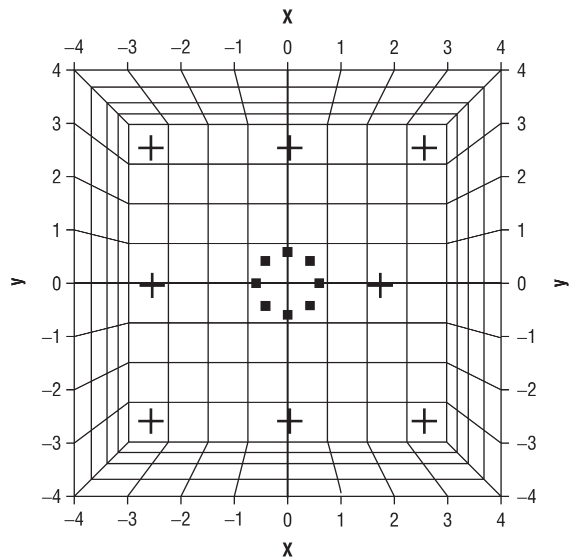
A two-dimensional plane separating points in three dimensional space.

Decision Surface



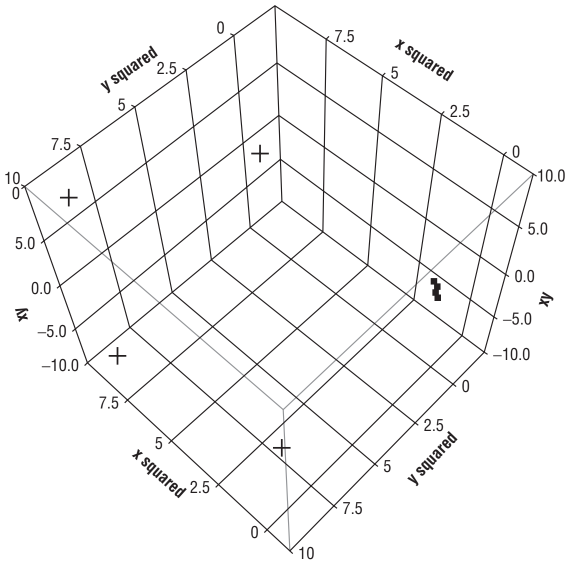
A one-dimensional line separating points on a two-dimensional plane.

Support Vectors



On the plane, boundary between the two classes is not a straight line.

Class Boundaries



After application of the kernel function, the two classes are easily separated.

Kernel Function

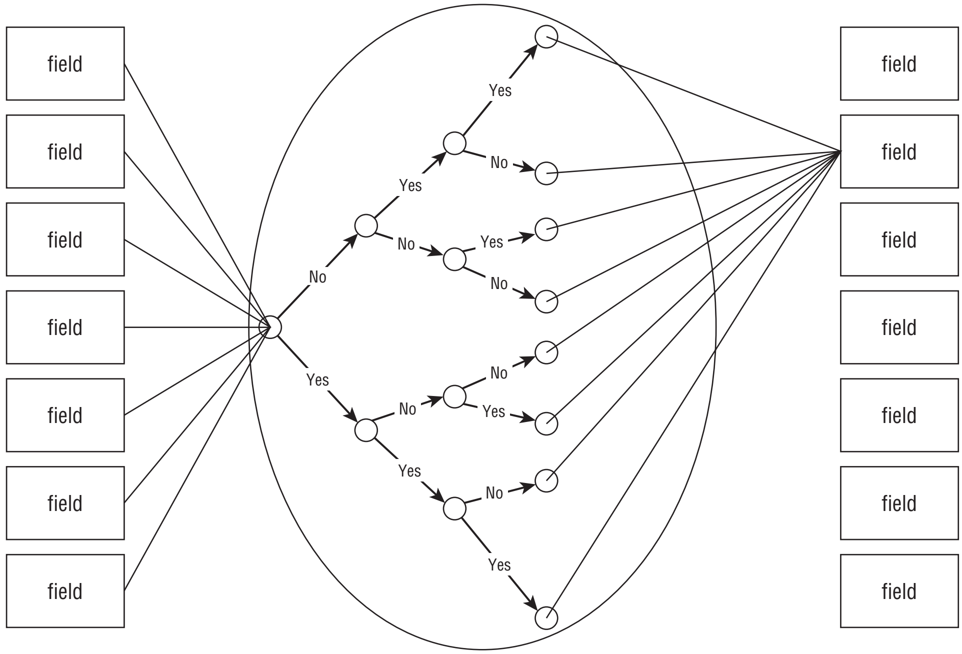
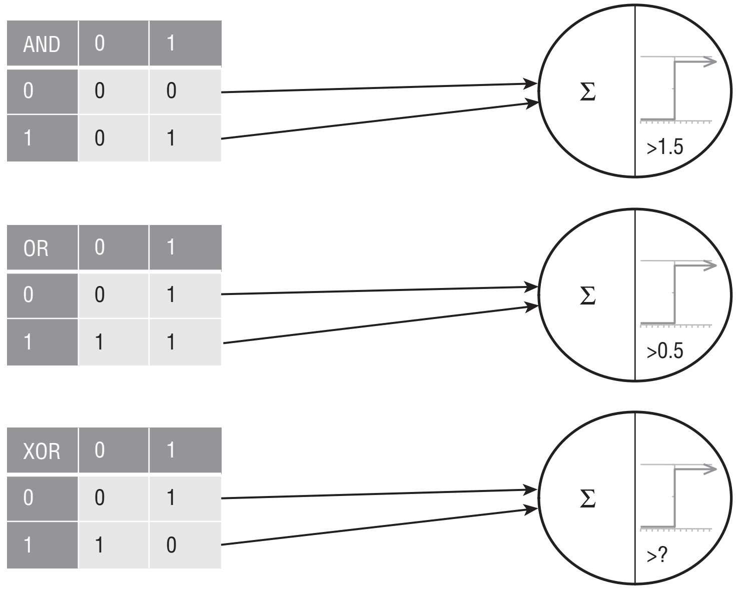


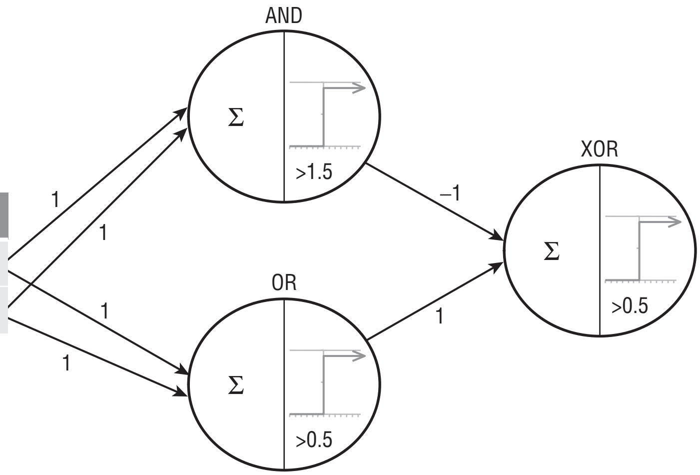
Figure 7-14: A decision tree uses values from one snapshot to create the next snapshot in time.



The XOR function cannot be implemented by a single-layer perceptron.

Function Tables

XOR	0	1
0	0	1
1	1	0



The XOR function is easily implemented by a two-layer perceptron.

Two-Layer Perceptron

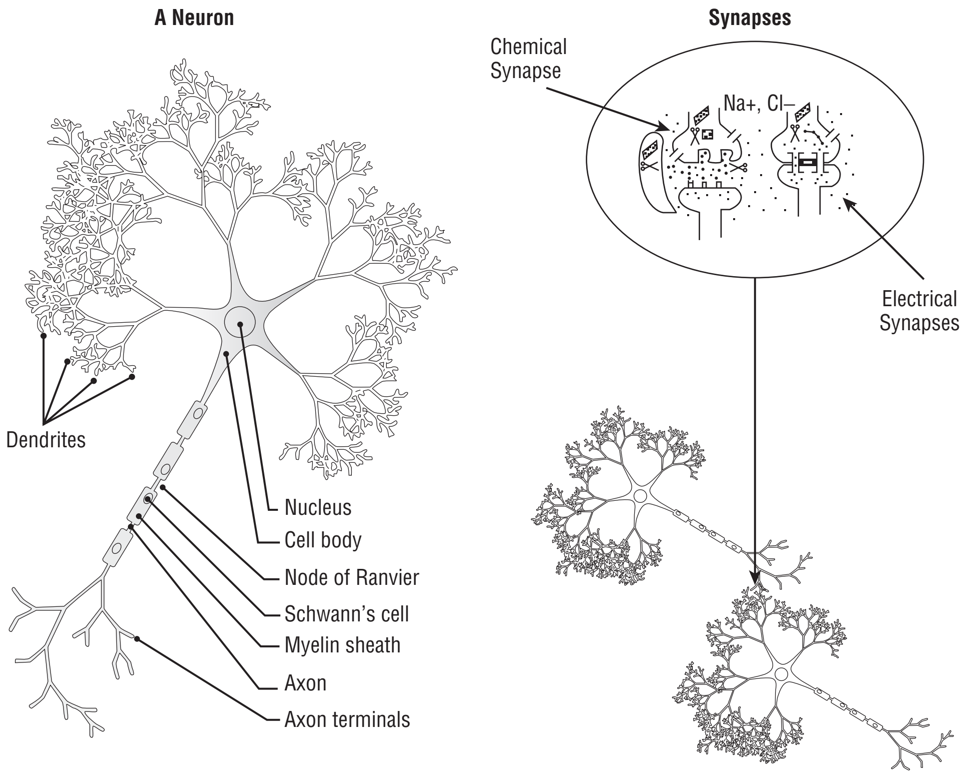


Figure 8-1: A neuron combines input signals from many other neurons to produce an output signal.

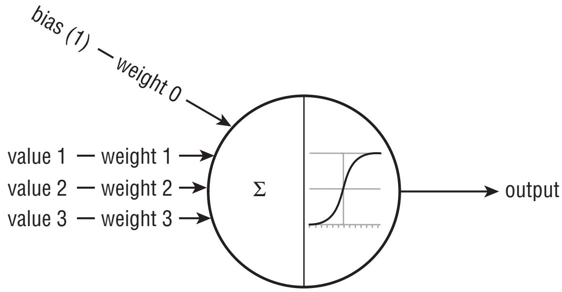


Figure 8-2: The output of the unit is typically a nonlinear combination of its inputs.

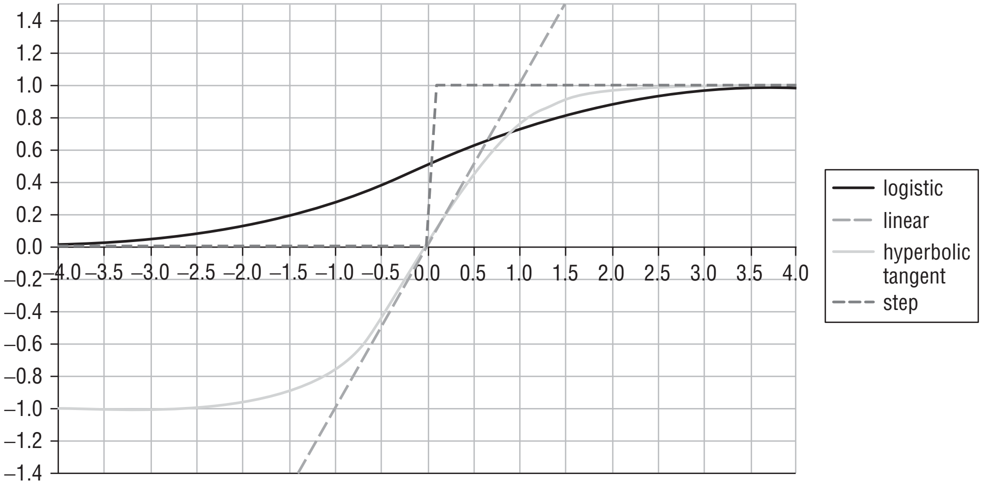


Figure 8-3: Four common transfer functions are the step, linear, logistic, and hyperbolic tangent functions.

$$\textbf{logistic}(x) = \frac{1}{(1 + e^{-x})}$$

Equation 22

$$\tanh(x) = \frac{e^{2x} - 1}{e^{2x} + 1}$$

Equation 23

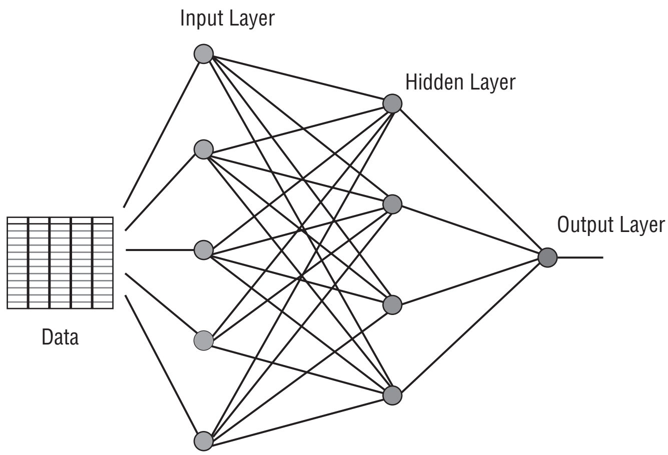


Figure 8-4: A multi-layer perceptron with a single hidden layer.

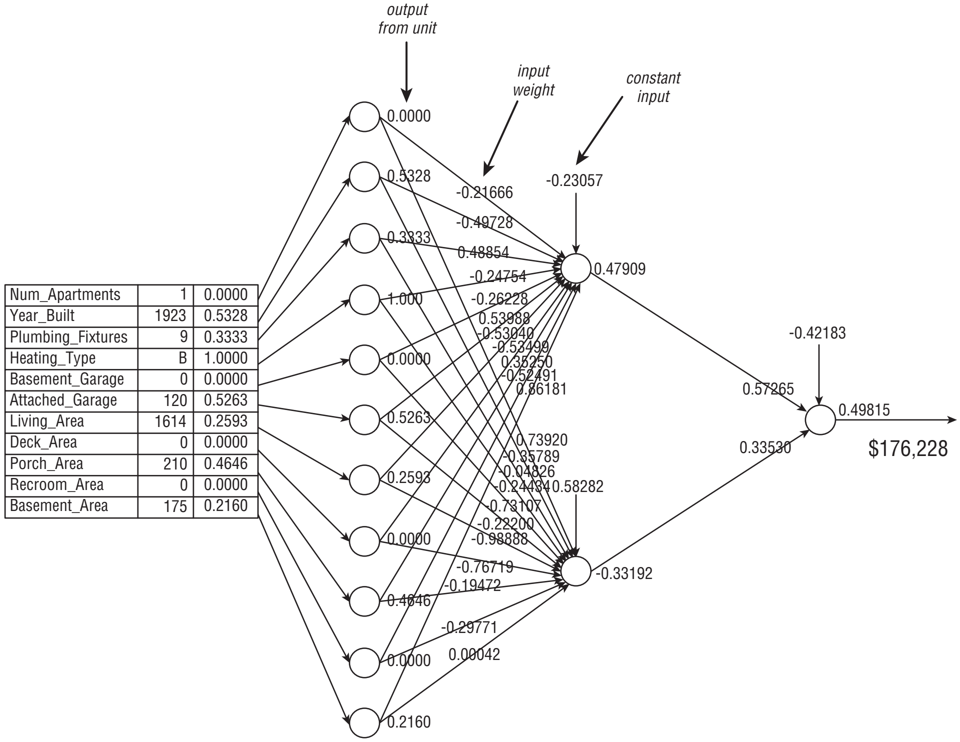


Figure 8-5: The real estate training example shown here provides the input into a neural network and illustrates that a network is filled with seemingly meaningless weights.

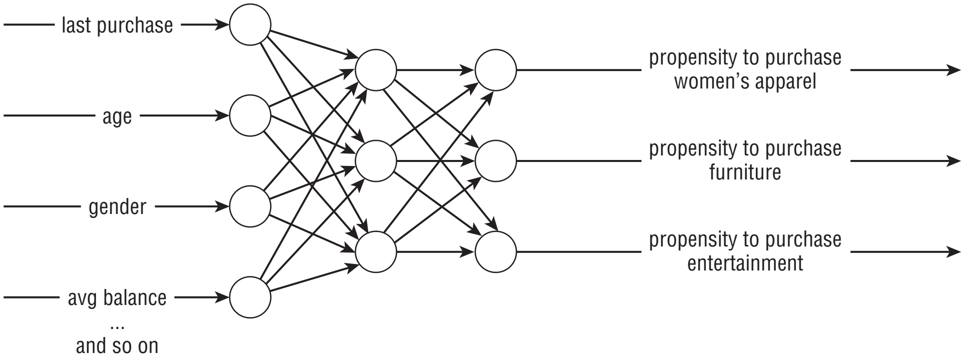


Figure 8-6: This network has more than one output and is used to estimate the probability that customers will make a purchase in each of three departments.

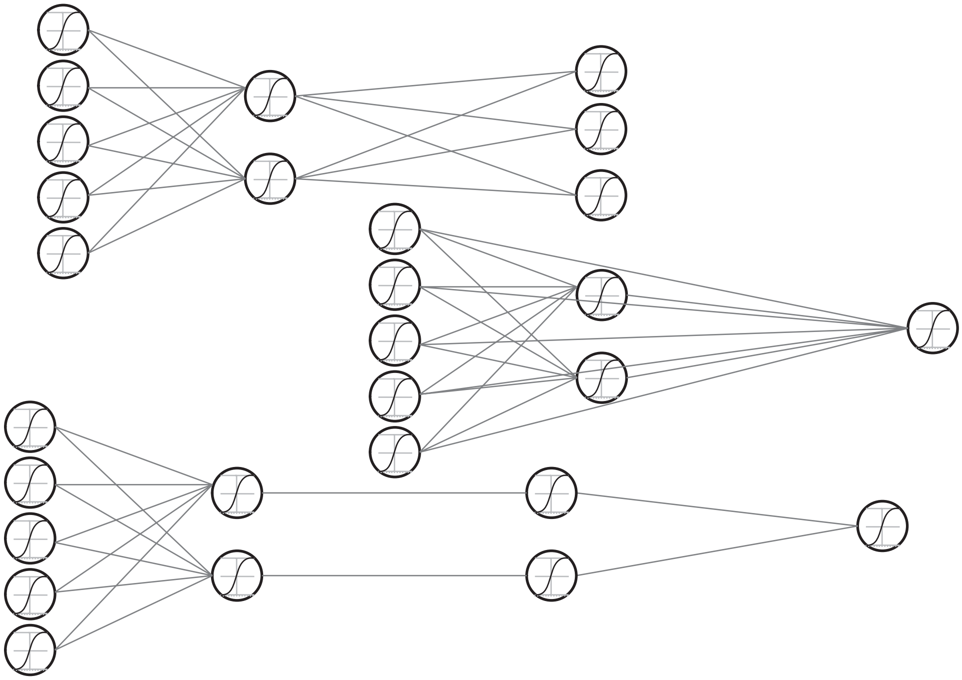


Figure 8-7: There are many variations on the basic neural network architecture.

Table 8-1: Common Features Describing a House

FEATURE	DESCRIPTION	RANGE OF VALUES
Num_Apartments	Number of dwelling units	Integer: 1–3
Year_Built	Year built	Integer: 1850–1986
Plumbing_Fixtures	Number of plumbing fixtures	Integer: 5–17
Heating_Type	Heating system type	Coded as A or B
Basement_Garage	Basement garage (number of cars)	Integer: 0–2
Attached_Garage	Attached frame garage area (in square feet)	Integer: 0–228
Living_Area	Total living area (square feet)	Integer: 714–4185
Deck_Area	Deck / open porch area (square feet)	Integer: 0–738
Porch_Area	Enclosed porch area (square feet)	Integer: 0–452
Recroom_Area	Recreation room area (square feet)	Integer: 0–672
Basement_Area	Finished basement area (square feet)	Integer: 0–810

Table 8-2: Sample Record from Training Set with Values Scaled to Range –1 to 1

FEATURE	RANGE OF VALUES	ORIGINAL VALUE	SCALED VALUE
Months_Ago	0–23	4	–0.6522
Num_Apartments	1-3	1	–1.0000
Year_Built	1850–1986	1923	+0.0730
Plumbing_Fixtures	5–17	9	–0.3077
Heating_Type	Coded as A or B	B	+1.0000
Basement_Garage	0–2	0	–1.0000
Attached_Garage	0–228	120	+0.0524
Living_Area	714–4185	1,614	–0.4813
Deck_Area	0–738	0	–1.0000
Porch_Area	0–452	210	–0.0706
Recroom_Area	0–672	0	–1.0000
Basement_Area	0–810	175	–0.5672

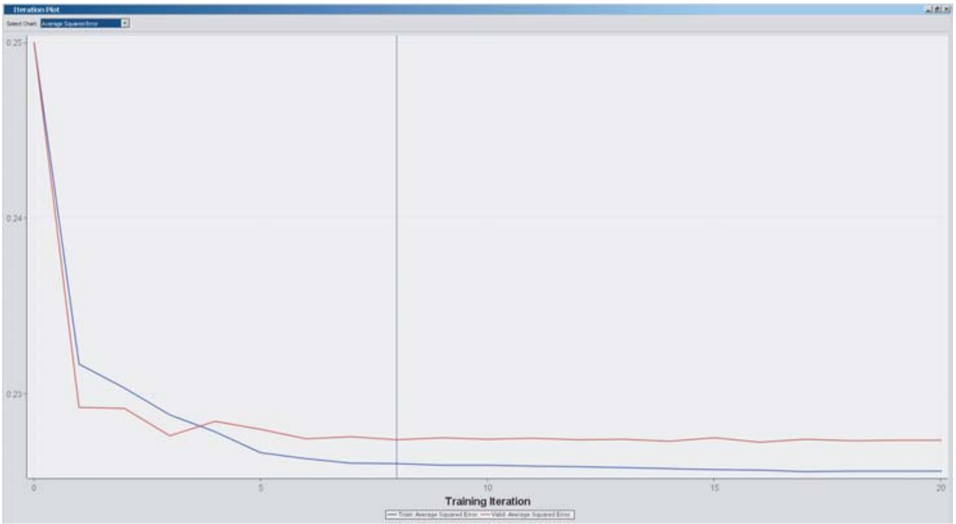


Figure 8-8: After twenty training iterations, error on the training data is nearly zero, but error on the validation data reached its lowest value after just seven iterations.

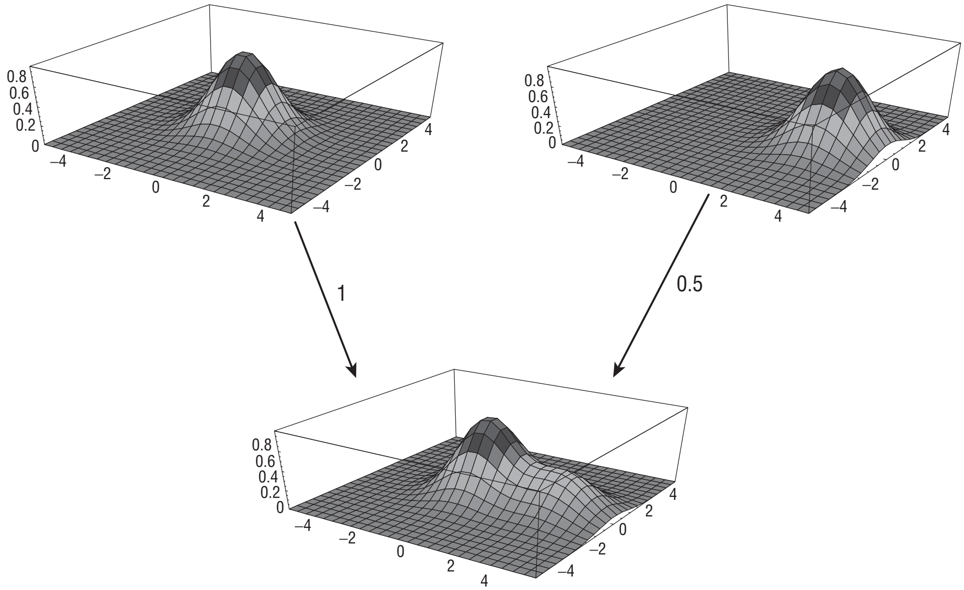


Figure 8-9: Two Gaussian surfaces are added to produce the output surface.

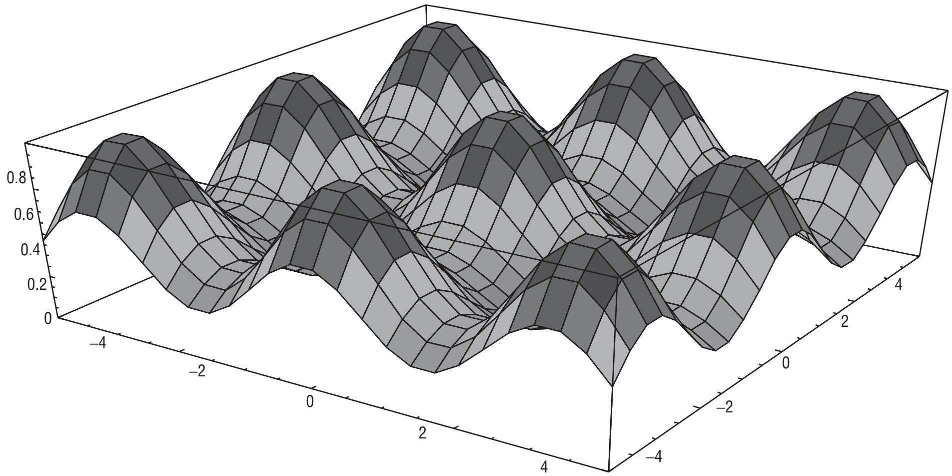


Figure 8-10: Radial basis functions can be placed in a grid to provide even coverage of the input space.

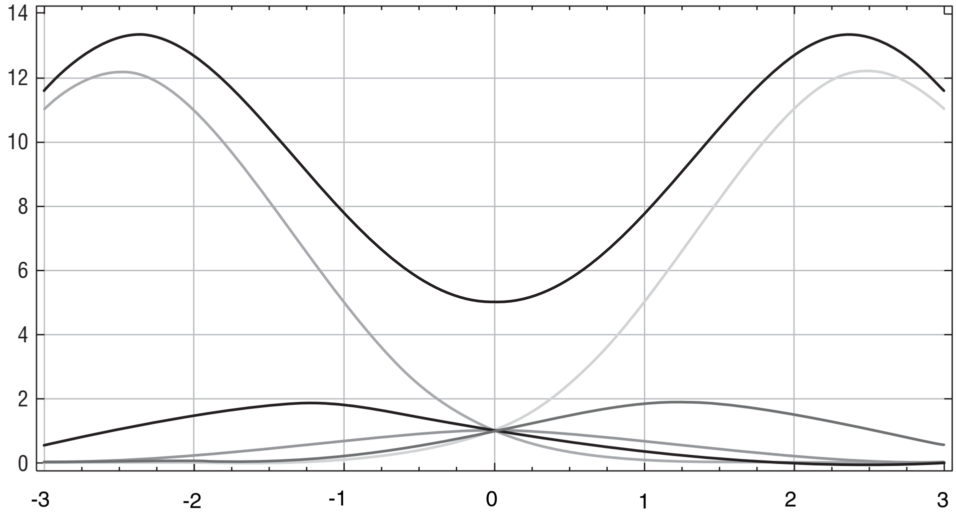


Figure 8-11: Several bell-shaped curves are added to produce a sinusoidal output curve.

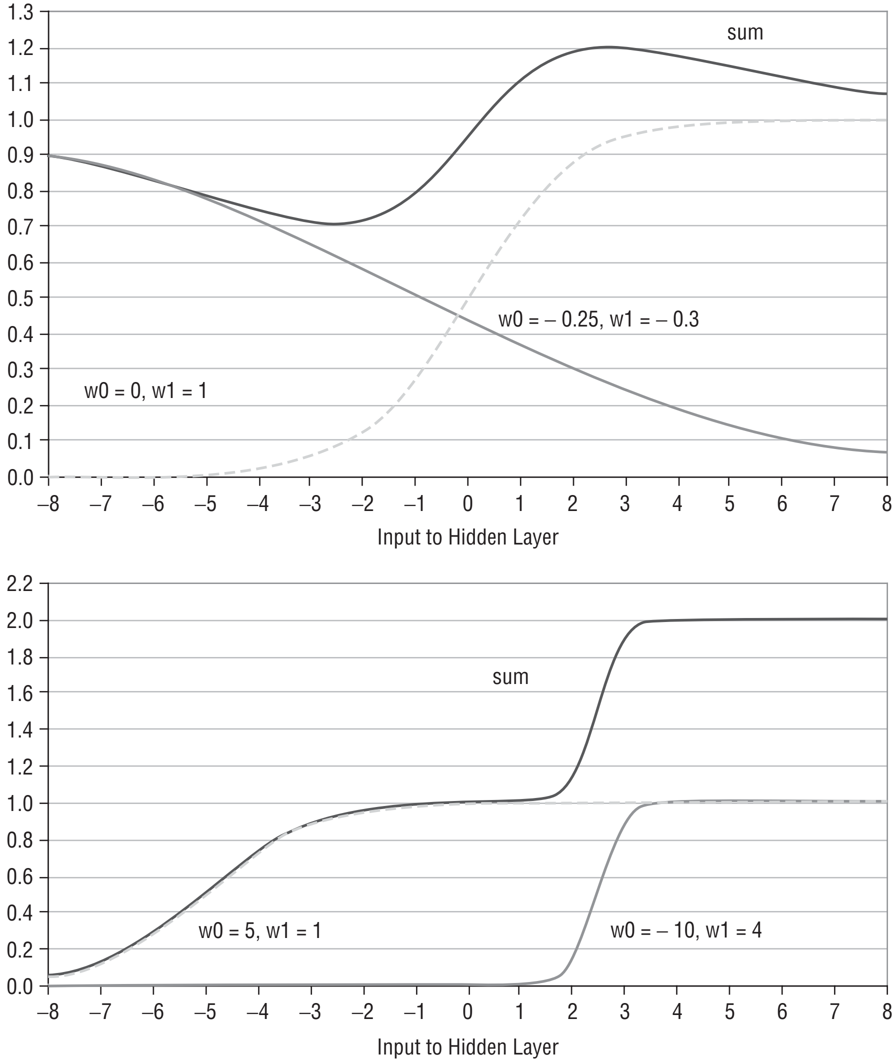
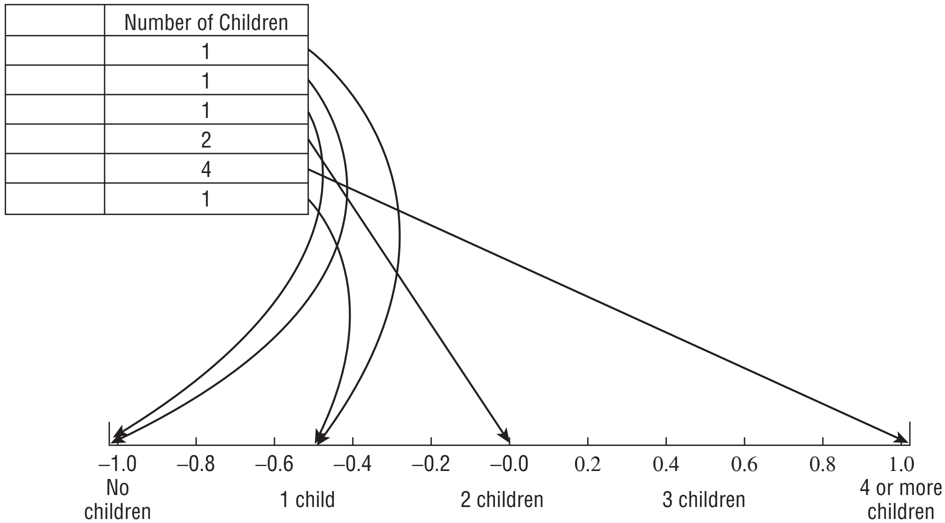


Figure 8-12: Varying the weights in an MLP with two hidden layer nodes leads to a variety of output curves.

0	$\rightarrow$	**0 0 0 0**	**= 0/16 = 0.0000**
1	$\rightarrow$	**1 0 0 0**	**= 8/16 = 0.5000**
2	$\rightarrow$	**1 1 0 0**	**= 12/16 = 0.7500**
3	$\rightarrow$	**1 1 1 0**	**= 14/16 = 0.8750**



When codes have an inherent order, they can be mapped onto the unit interval.

Thermometer Codes

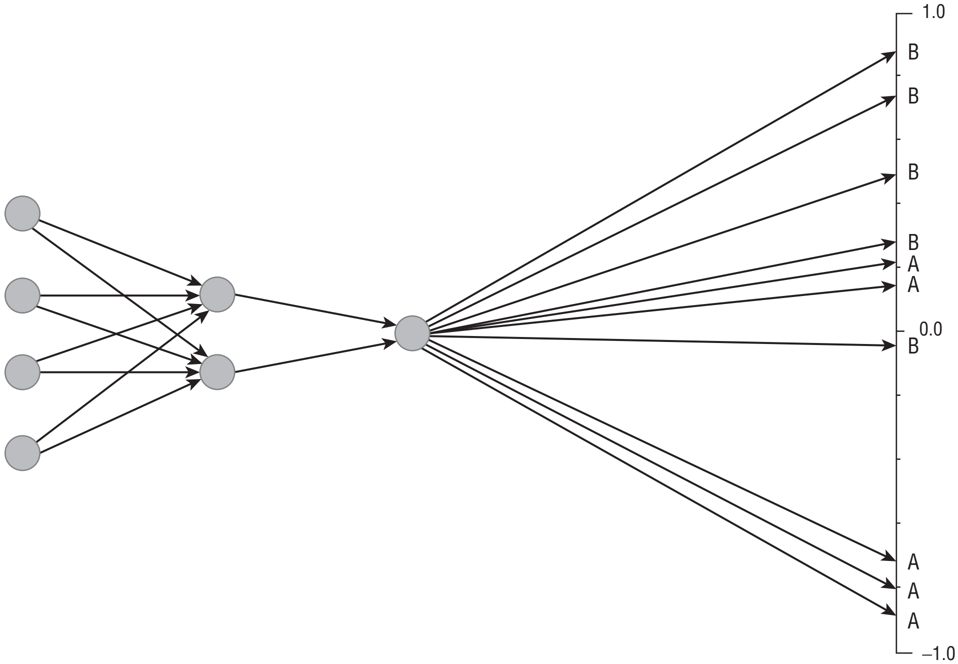


Figure 8-13: Running a neural network on examples from the validation set can help determine how to interpret results.

Table 8-3: Time Series

DATA ELEMENT	DAY-OF-WEEK	CLOSING PRICE
1	1	$40.25
2	2	$41.00
3	3	$39.25
4	4	$39.75
5	5	$40.50
6	1	$40.50
7	2	$40.75
8	3	$41.25
9	4	$42.00
10	5	$41.50

Table 8-4: Time Series with Time Lag

DATA ELEMENT	DAY-OF-WEEK	CLOSING PRICE	PREVIOUS CLOSING PRICE	PREVIOUS-1 CLOSING PRICE
1	1	$40.25		
2	2	$41.00	$40.25	
3	3	$39.25	$41.00	$40.25
4	4	$39.75	$39.25	$41.00
5	5	$40.50	$39.75	$39.25
6	1	$40.50	$40.50	$39.75
7	2	$40.75	$40.50	$40.50
8	3	$41.25	$40.75	$40.50
9	4	$42.00	$41.25	$40.75
10	5	$41.50	$42.00	$41.25

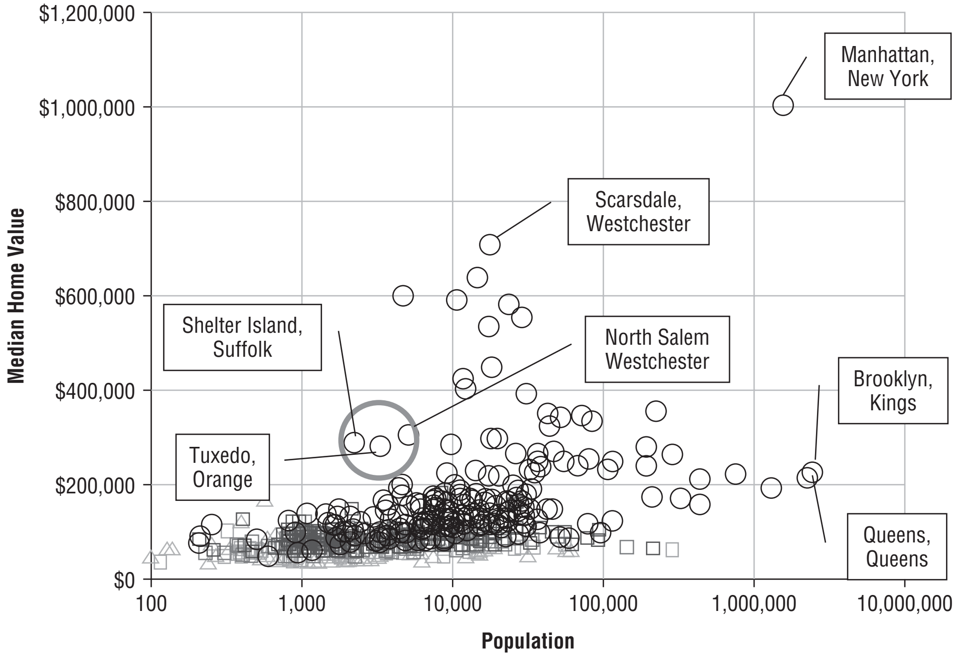


Figure 9-1: Based on 2000 census population and home value, the town of Tuxedo in Orange County has Shelter Island and North Salem as its two nearest neighbors.

Table 9-1: The Neighbors

TOWN	POP.	RENTING HOUSE-HOLDS	MEDIAN RENT	RENT <$500	RENT $750	RENT $1000	RENT $1,500	RENT >$1,500	NON-CASH
Shelter Island	2,228	160	$804	3.1%	34.6%	31.4%	10.7%	3.1%	17.0%
North Salem	5,173	244	$1,150	3.0%	10.2%	21.6%	30.9%	24.2%	10.2%
Tuxedo	*3,334*	*349*	*$907*	*4.6%*	*27.2%*	*29.6%*	*23.8%*	*3.8%*	*14.8%*

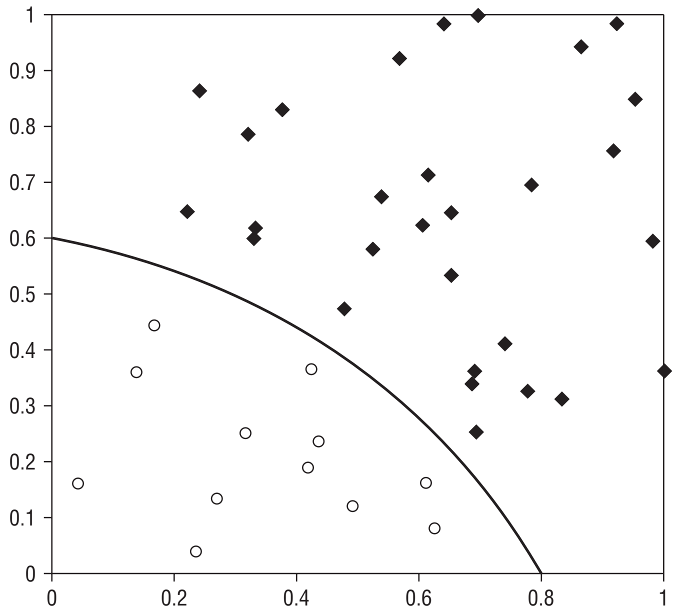


Figure 9-2: Perhaps the cleanest training set for MBR is one that divides neatly into two disjoint sets.

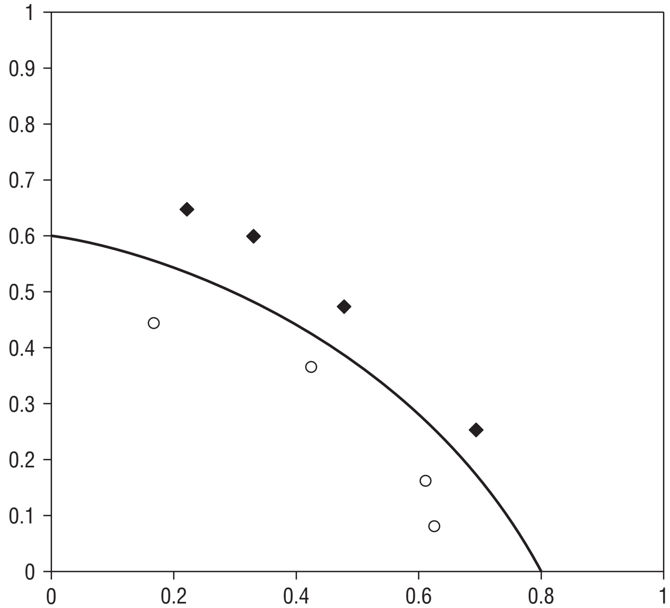


Figure 9-3: This smaller set of points returns the same results as in Figure 9-2 using MBR.

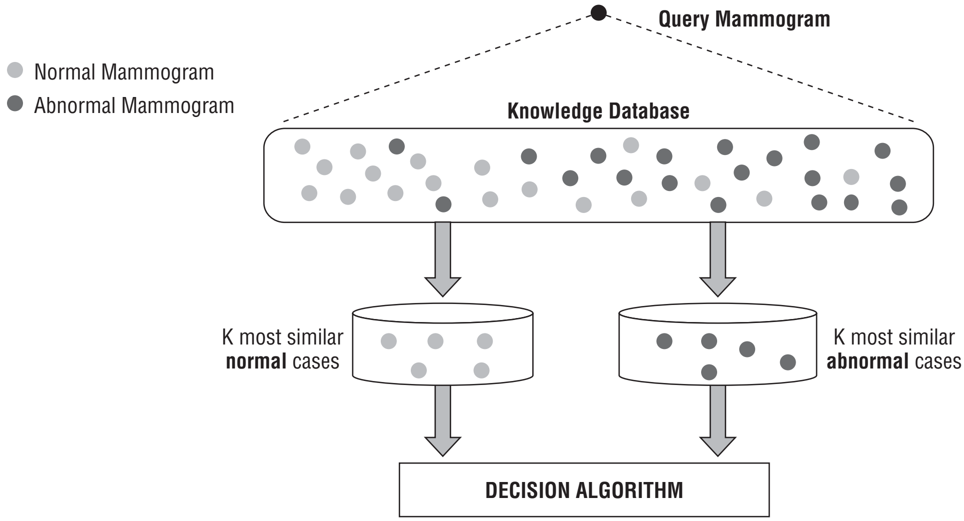


Figure 9-4: The basic idea for automated diagnosis of mammogram abnormalities using MBR finds similar normal and abnormal cases in the knowledge base, and then decides which to present to the physician. (Courtesy of Dr. Tourassi)

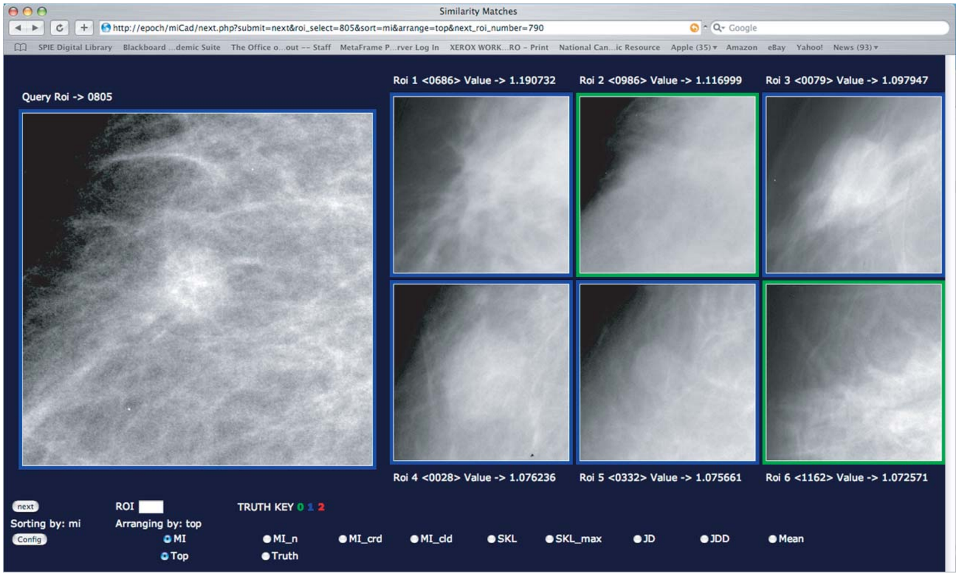


Figure 9-5: Similarity matches for a mammogram suggest whether or not the mammogram is normal or abnormal — and provide nearby examples for further investigation.

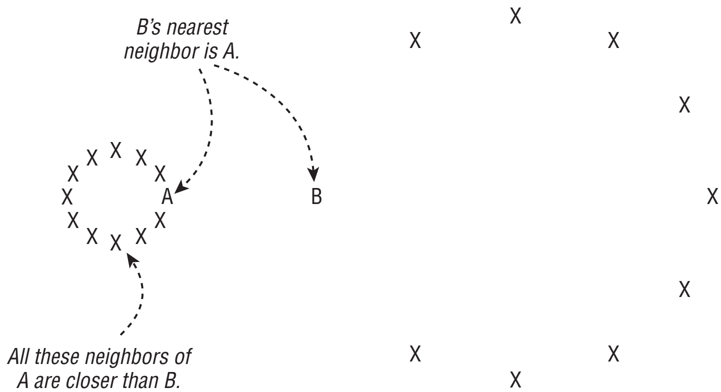


Figure 9-6: B's nearest neighbor is A, but A has many neighbors closer than B.

Table 9-2: Five Customers in a Marketing Database

RECNUM	GENDER	AGE	SALARY
1	Female	27	$ 19,000
2	Male	51	$ 64,000
3	Male	52	$105,000
4	Female	33	$ 55,000
5	Male	45	$ 45,000

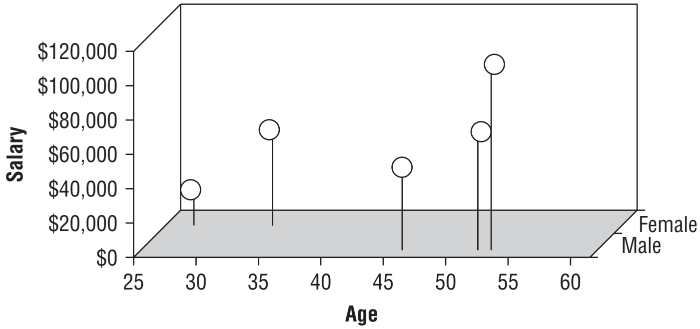


Figure 9-7: This scatter plot shows the five records from Table 9-2 in three dimensions — age, salary, and gender — and suggests that standard distance is a good metric for nearest neighbors.

Table 9-3: Distance Matrix Based on Ages of Customers

	27	51	52	33	45
27	0.00	0.96	1.00	0.24	0.72
51	0.96	0.00	0.04	0.72	0.24
52	1.00	0.04	0.00	0.76	0.28
33	0.24	0.72	0.76	0.00	0.48
45	0.72	0.24	0.28	0.48	0.00

Table 9-4: Set of Nearest Neighbors for Three Distance Functions, Ordered Nearest to Farthest

	D_{SUM}	D_{NORM}	D_{EUCLID}
1	1,4,5,2,3	1,4,5,2,3	1,4,5,2,3
2	2,5,3,4,1	2,5,3,4,1	2,5,3,4,1
3	3,2,5,4,1	3,2,5,4,1	3,2,5,4,1
4	4,1,5,2,3	4,1,5,2,3	4,1,5,2,3
5	5,2,3,4,1	5,2,3,4,1	5,2,3,4,1

Table 9-5: New Customer

RECNUM	GENDER	AGE	SALARY
New	Female	45	$100,000

Table 9-6: Set of Nearest Neighbors for New Customer

	1	2	3	4	5	NEIGHBORS
d_{sum}	1.662	1.659	1.338	1.003	1.640	4,3,5,2,1
d_{Euclid}	0.781	1.052	1.251	0.494	1.000	4,1,5,2,3

Table 9-7: Customers with Attrition History

RECNUM	GENDER	AGE	SALARY	INACTIVE
1	Female	27	$19,000	no
2	Male	51	$64,000	yes
3	Male	52	$105,000	yes
4	Female	33	$55,000	yes
5	Male	45	$45,000	no
New	Female	45	$100,000	?

Table 9-8: Using MBR to Determine Whether the New Customer Will Become Inactive

	NEIGHBORS	NEIGHBOR ATTRITION	K = 1	K = 2	K = 3	K = 4	K = 5
d_{sum}	4,3,5,2,1	Y,Y,N,Y,N	yes	yes	yes	yes	yes
d_{Euclid}	4,1,5,2,3	Y,N,N,Y,Y	yes	?	no	?	yes

Table 9-9: Attrition Prediction with Confi dence

	K = 1	K = 2	K = 3	K = 4	K = 5
d_{sum}	yes, 100%	yes, 100%	yes, 67%	yes, 75%	yes, 60%
d_{Euclid}	yes, 100%	yes, 50%	no, 67%	yes, 50%	yes, 60%

Table 9-10: Attrition Prediction with Weighted Voting

	K = 1	K = 2	K = 3	K = 4	K = 5
d_{sum}	**0.749** to 0	**1.441** to 0	**1.441** to 0.647	**2.085** to 0.647	**2.085** to 1.290
d_{Euclid}	**0.669** to 0	**0.669** to 0.562	0.669 to **1.062**	**1.157** to 1062	**1.601** to 1.062

Table 9-11: Confi dence with Weighted Voting

	1	2	3	4	5
d_{sum}	yes, 100%	yes, 100%	yes, 69%	yes, 76%	yes, 62%
d_{Euclid}	yes, 100%	yes, 54%	no, 61%	yes, 52%	yes, 60%

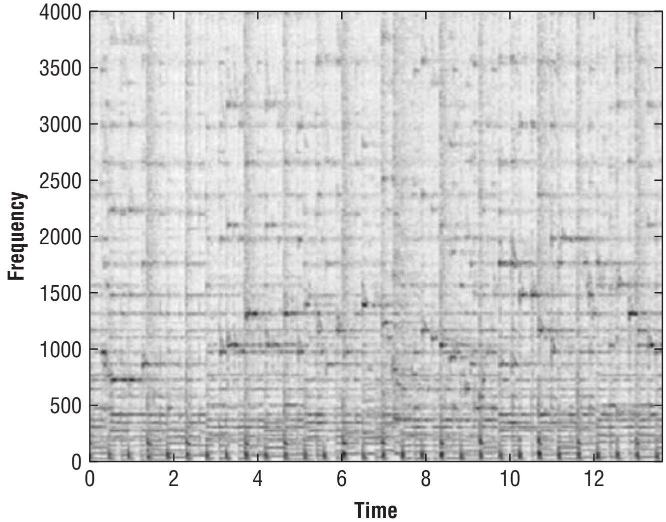


Figure 9-8: A spectrogram is a picture of a song in the frequency domain, with frequencies sampled every half second.

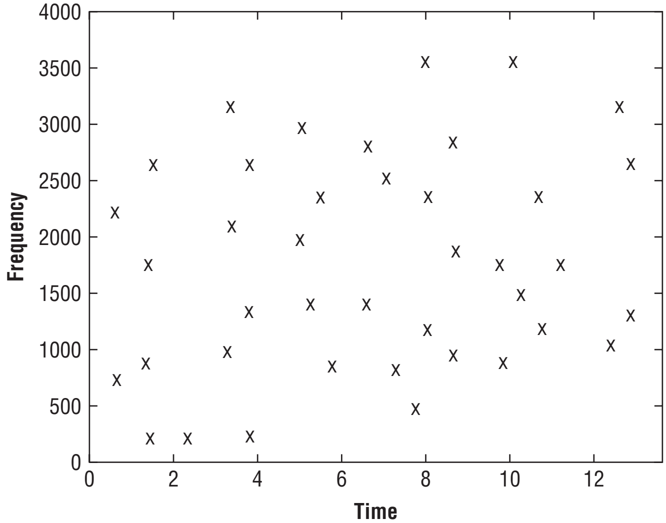


Figure 9-9: A constellation is a picture of the peaks of frequencies for a song in the frequency domain.

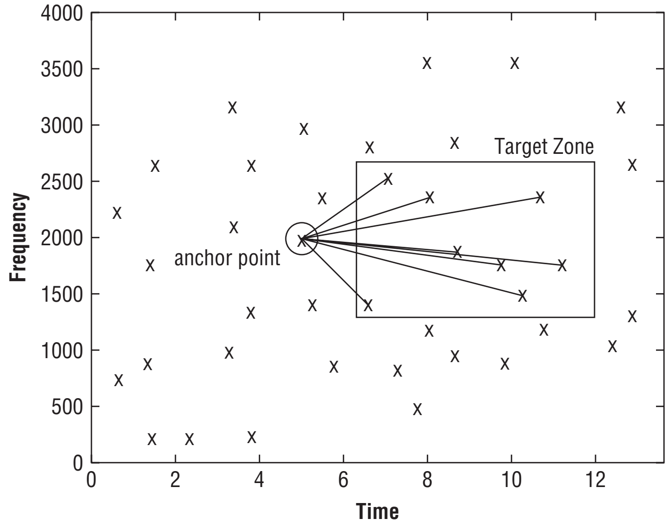


Figure 9-10: An anchor point is defined only by the set of peaks within a particular range of frequencies and times after the point in question.

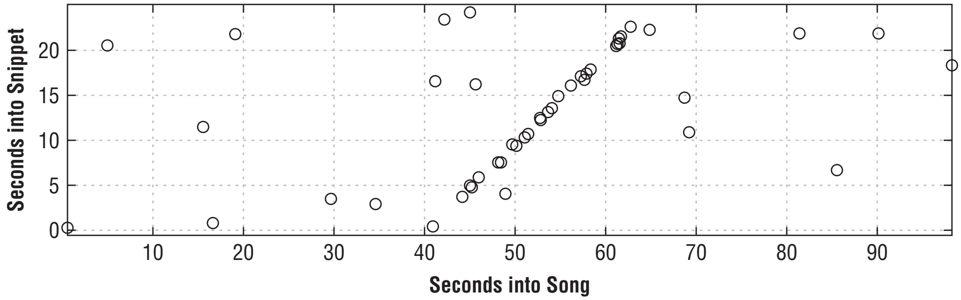


Figure 9-11: Anchor points that match are plotted in the absolute timeframe of both the song and the snippet. The vertical line starting at 41 seconds indicates that the snippet is matching that portion of the song.

$$\left(\tfrac{1}{2}\,(-1) + \tfrac{1}{4}\,(-4)\right) / \left(\tfrac{1}{2} + \tfrac{1}{4}\right) = -1.5/0.75 = -2.$$

Equation 24

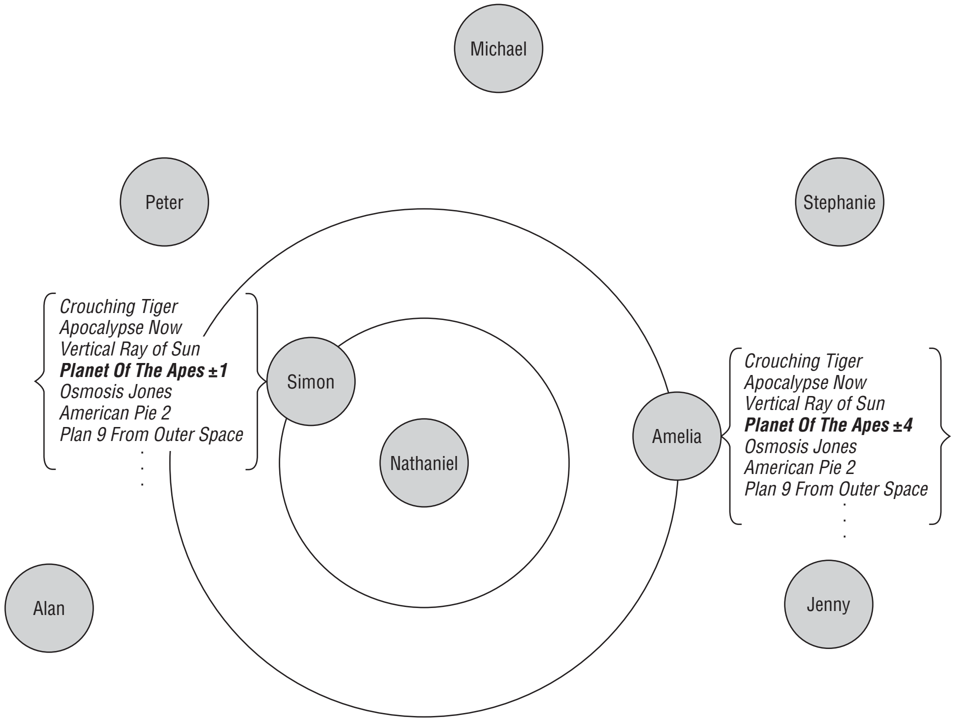


Figure 9-12: The predicted rating for *Planet of the Apes* is –2.66.

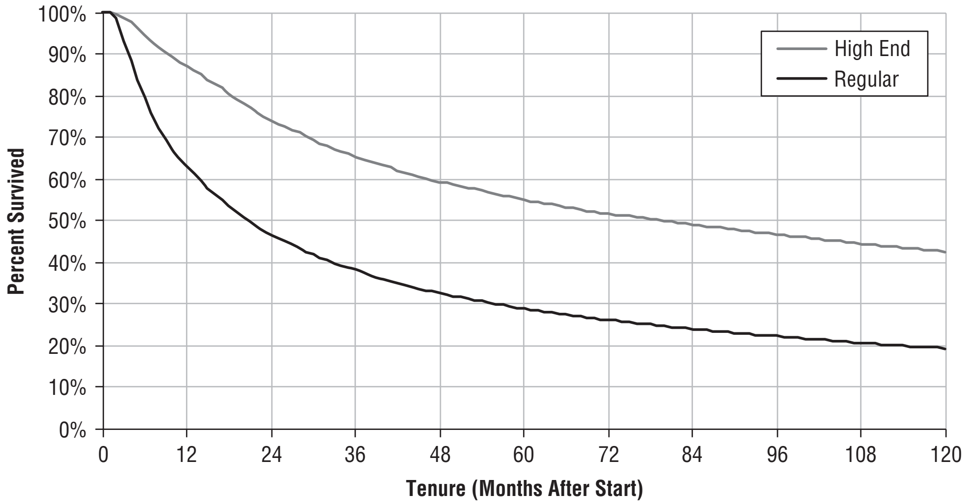


Figure 10-1: Survival curves show that high-end customers stay around longer.

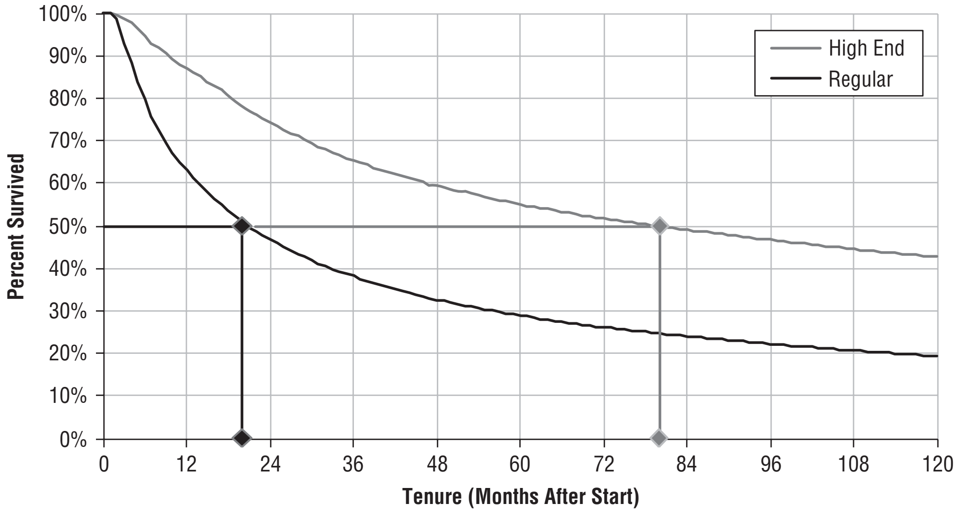


Figure 10-2: The median customer lifetime is where the retention curve crosses the 50 percent point.

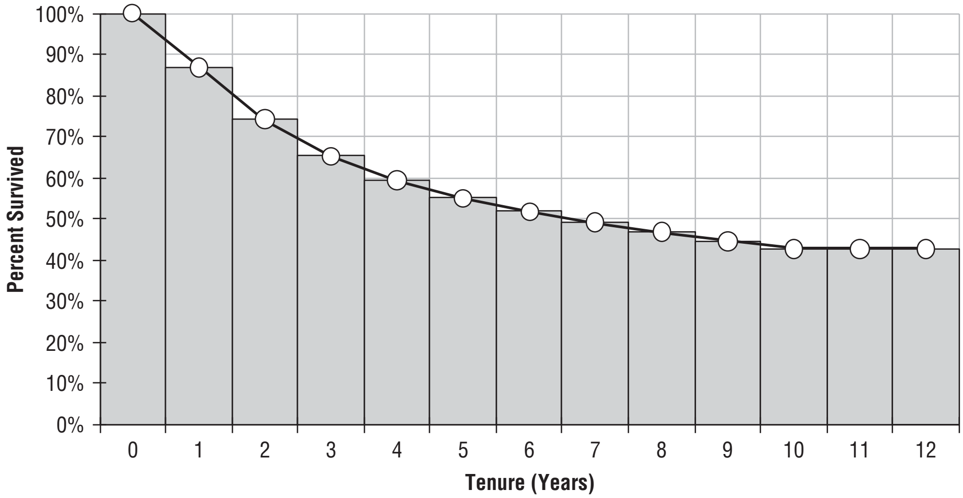


Figure 10-3: Circumscribing each point with a rectangle makes it clear how to approximate the area under the survival curve.

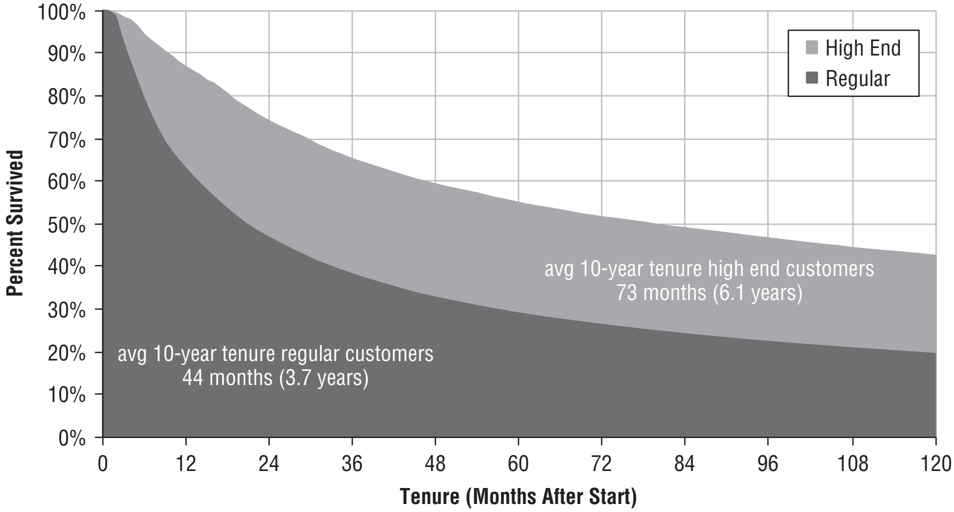
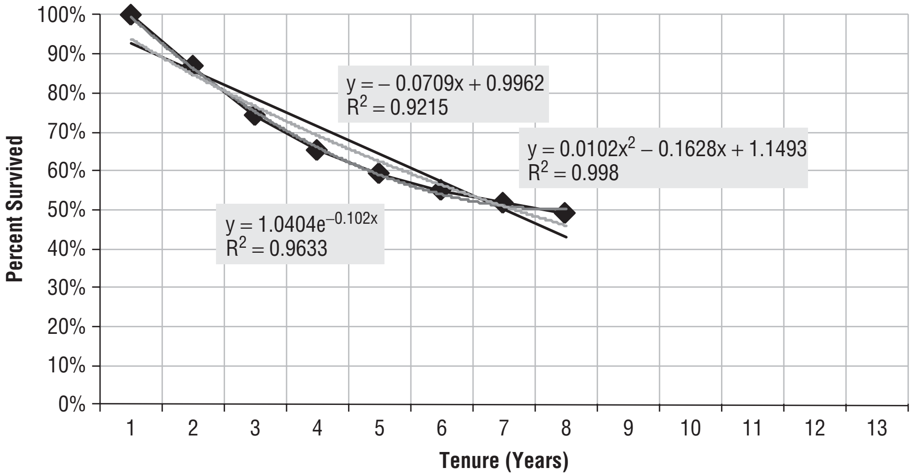
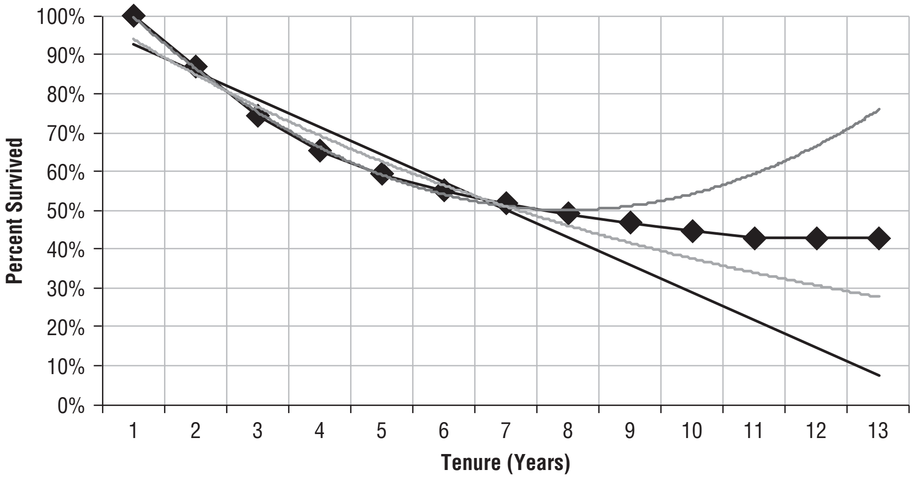


Figure 10-4: Average customer lifetime for different groups of customers can be compared using the areas under the survival curve.



Fitting parametric curves to a survival curve is easy.

Survival Curve



The parametric curves that fit a retention curve do not fit well beyond the range where they are defined.

Parametric Curve

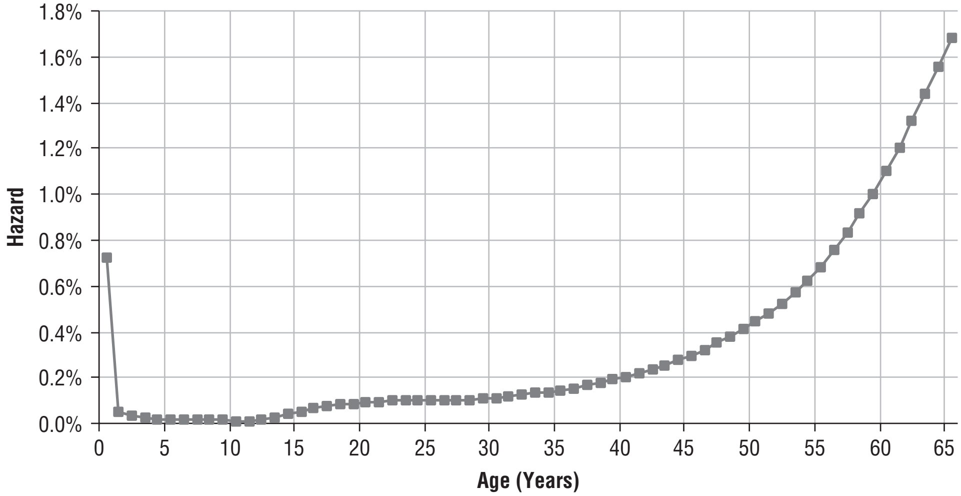


Figure 10-5: The shape of a bathtub-shaped hazard function starts high, plummets, and then gradually increases again.

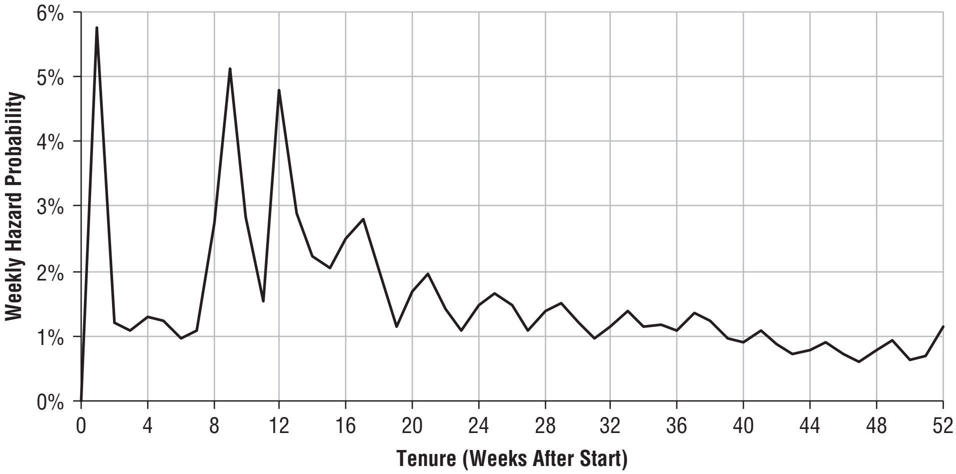


Figure 10-6: A subscription business has customer hazard probabilities that look like this.

Table 10-1: Tenure Data for Several Customers

CUSTOMER	CENSORED	TENURE
1	Y	12
2	N	6
3	N	6
4	N	3
5	Y	3
6	N	5
7	N	6
8	Y	9

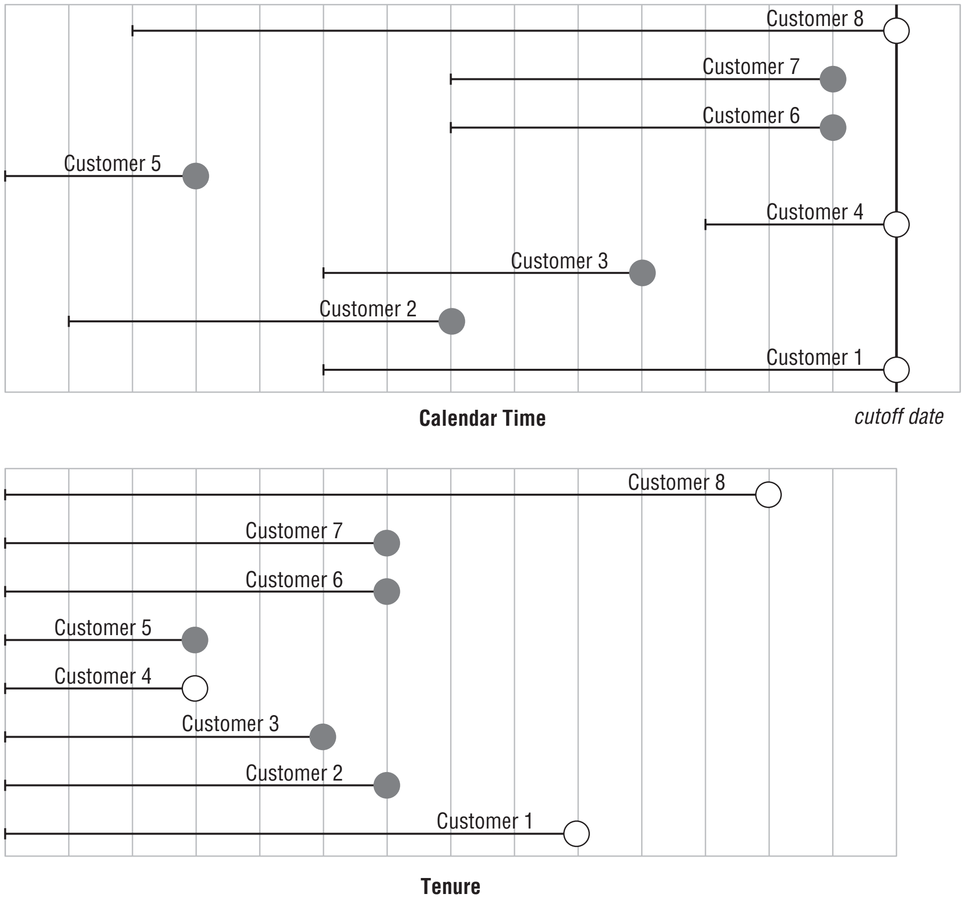


Figure 10-7: The top chart shows a group of customers who all start at different times; some customers are censored because they are still active. The bottom chart shows the same customers on the tenure time scale.

Table 10-2: Tracking Customers over Several Time Periods (A=Active; S=Stopped; blank=Censored)

	TENURE PERIOD												
CUSTOMER	0	1	2	3	4	5	6	7	8	9	10	11	12
1	A	A	A	A	A	A	A	A	A	A	A	A	A
2	A	A	A	A	A	A	S						
3	A	A	A	A	A	A	S						
4	A	A	A	S									
5	A	A	A	A									
6	A	A	A	A	A	S							
7	A	A	A	A	A	A	S						
8	A	A	A	A	A	A	A	A	A	A			

Table 10-3: From Times to Hazards

	TENURE PERIOD												
	0	1	2	3	4	5	6	7	8	9	10	11	12
ACTIVE	8	8	8	7	6	5	2	2	2	2	1	1	1
STOPPED	0	0	0	1	0	1	3	0	0	0	0	0	0
CENSORED	0	0	0	0	2	2	3	6	6	6	7	7	7
HAZARD	0.0%	0.0%	0.0%	12.5%	0.0%	16.7%	60.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%

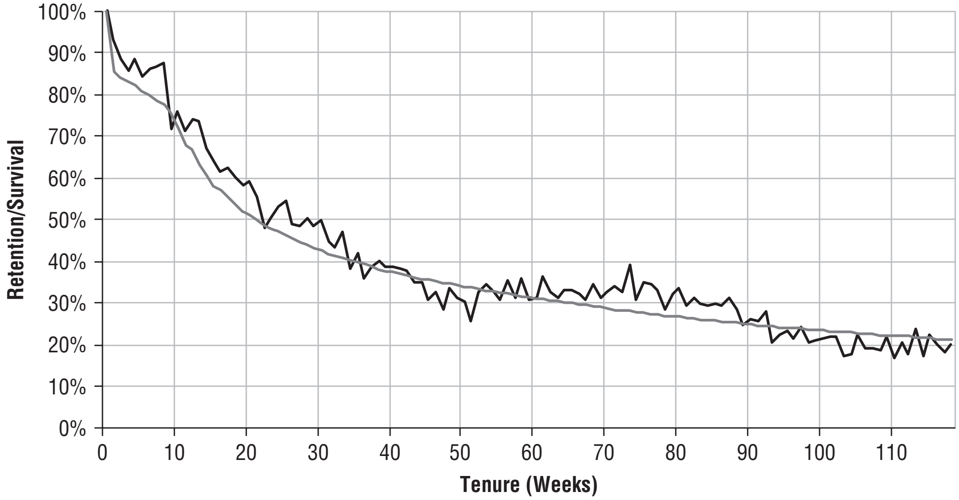


Figure 10-8: A retention curve might be quite jagged, especially in comparison to the survival curve for the same data.

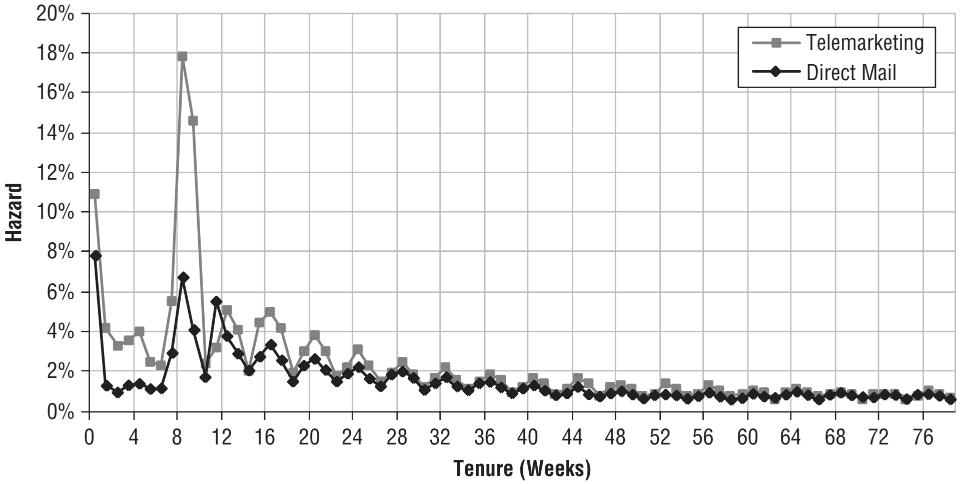


Figure 10-9: These two hazard functions suggest that the risk of attrition is about one and a half times as great for customers acquired through telemarketing versus direct mail, although the ratio does differ somewhat by tenure.

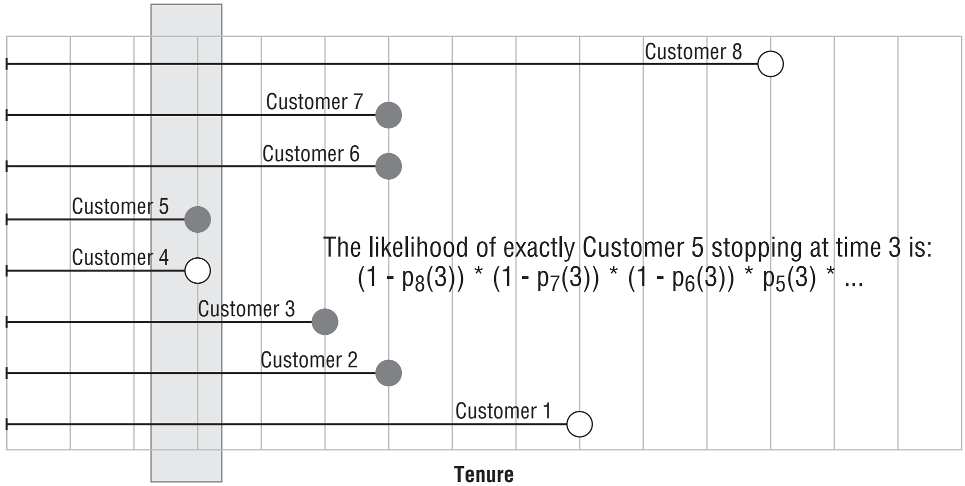


Figure 10-10: Cox's insightful observation that led to proportional hazards modeling is to look at all customers at a given tenure and ask, "What is the likelihood that exactly one set of customers stops when the rest remain active?"

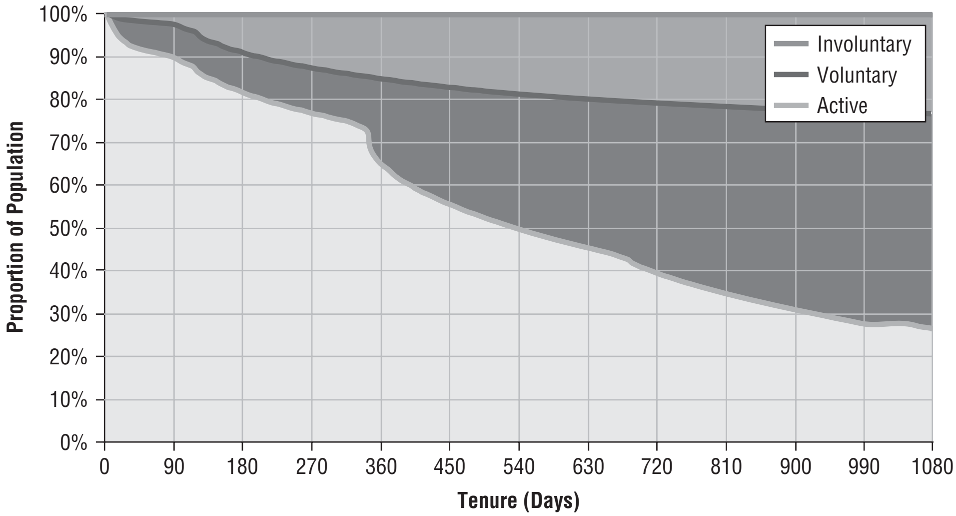


Figure 10-11: Using competing risks, creating a chart that shows the proportion of customers that succumb to each risk at any given tenure is possible.

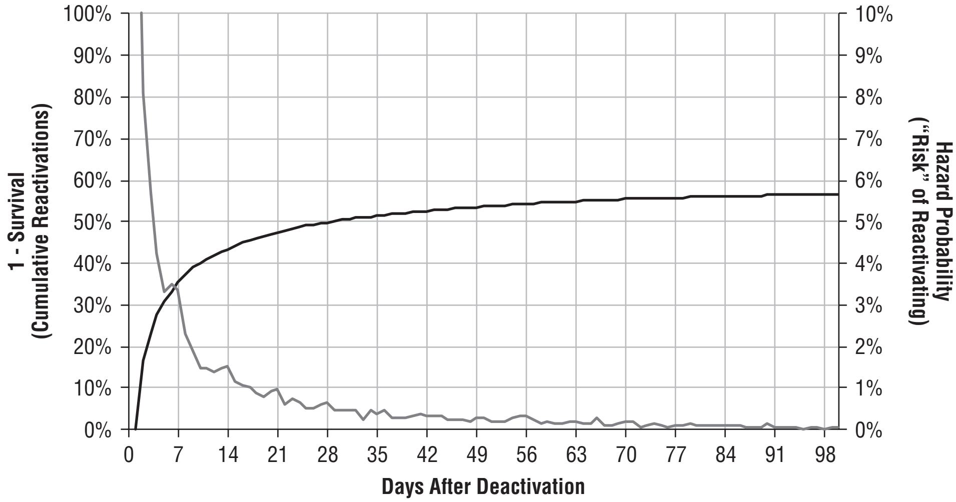


Figure 10-12: This chart shows 1-Survival, the cumulative number of reactivations as well as the "hazard probability" of reactivation.

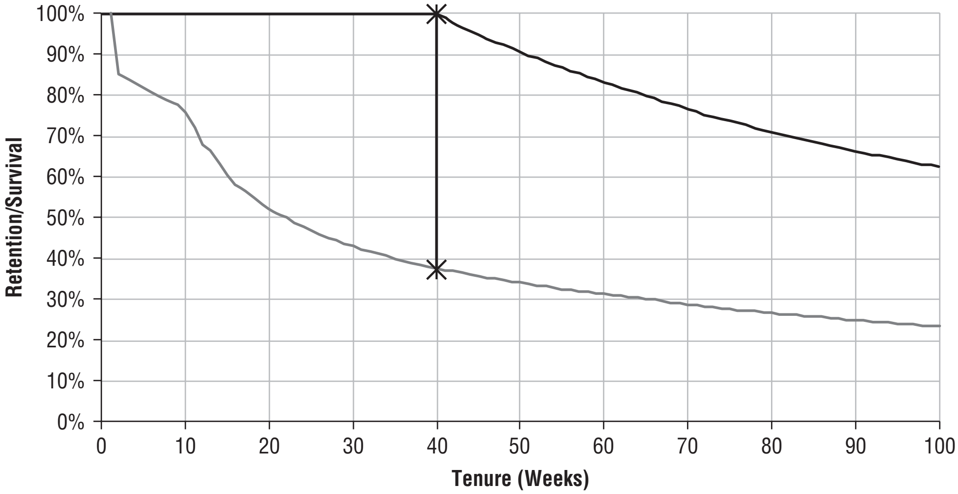


Figure 10-13: The conditional survival is the survival, assuming that a customer has survived to a particular tenure. It is calculated by dividing the survival value by the value at that tenure.

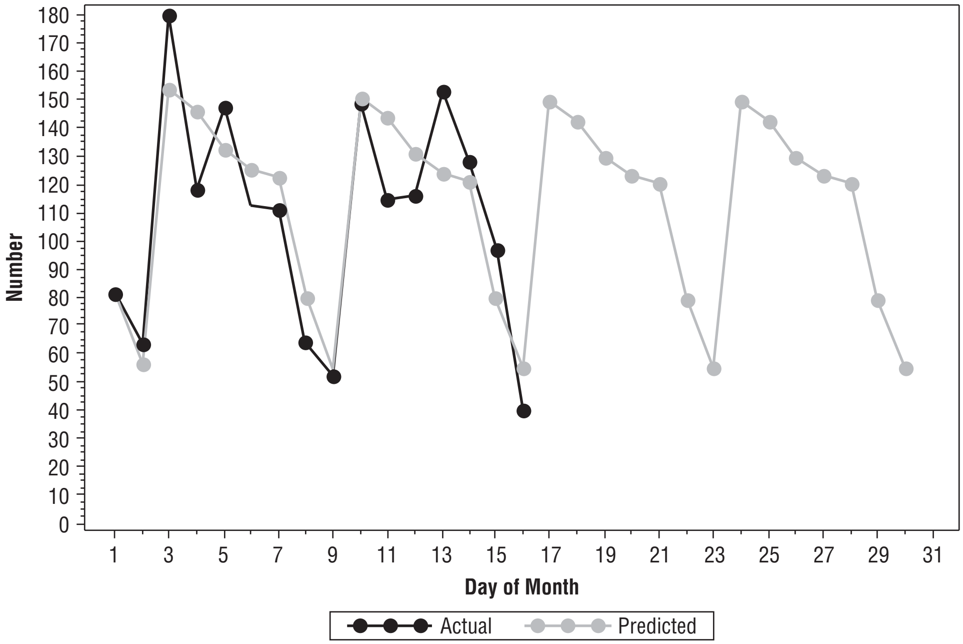


Figure 10-14: You can also use survival analysis for forecasting customer stops.

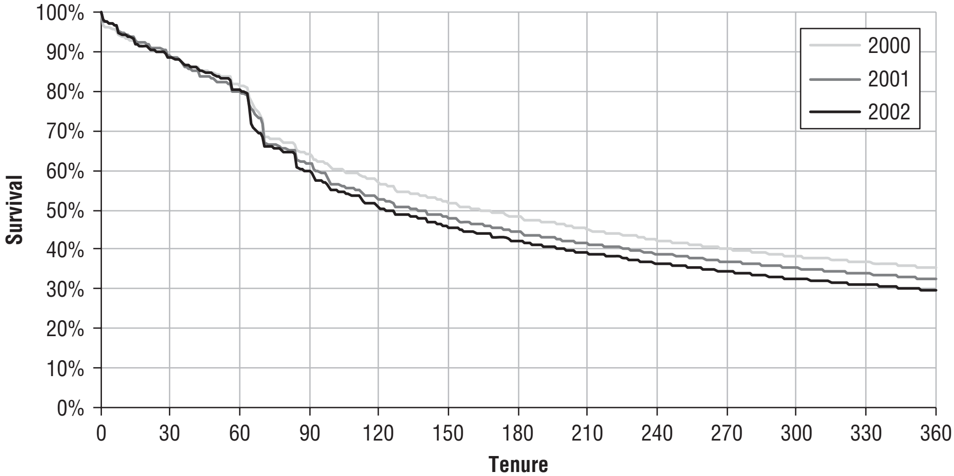


Figure 10-15: A time-window technique allows you to see changes in survival over time.

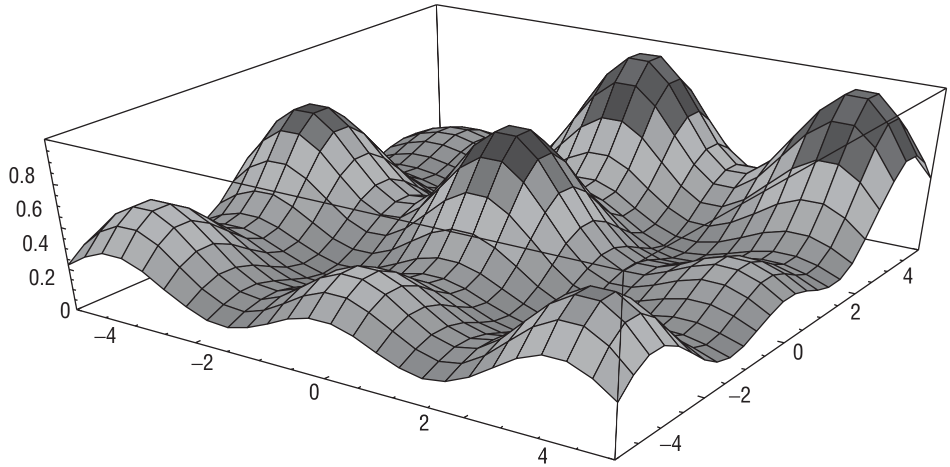
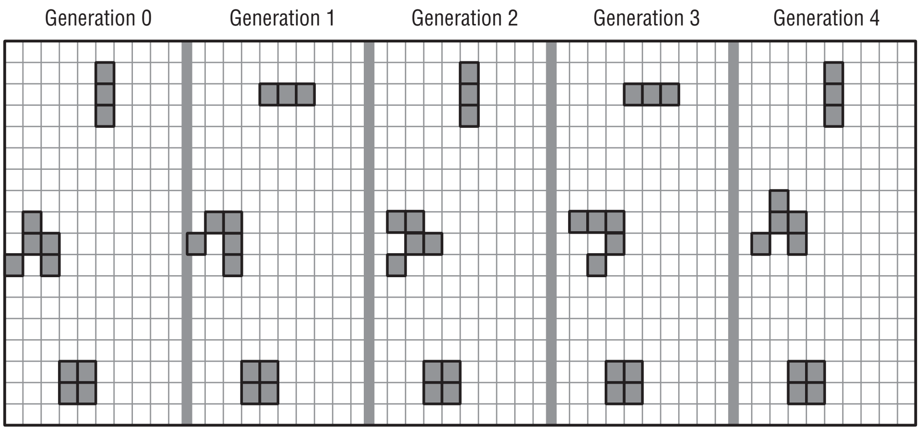


Figure 11-1: The optimization challenge is to find the highest hill.



Several generations of the Game of Life from two starting patterns.

Game of Life Patterns

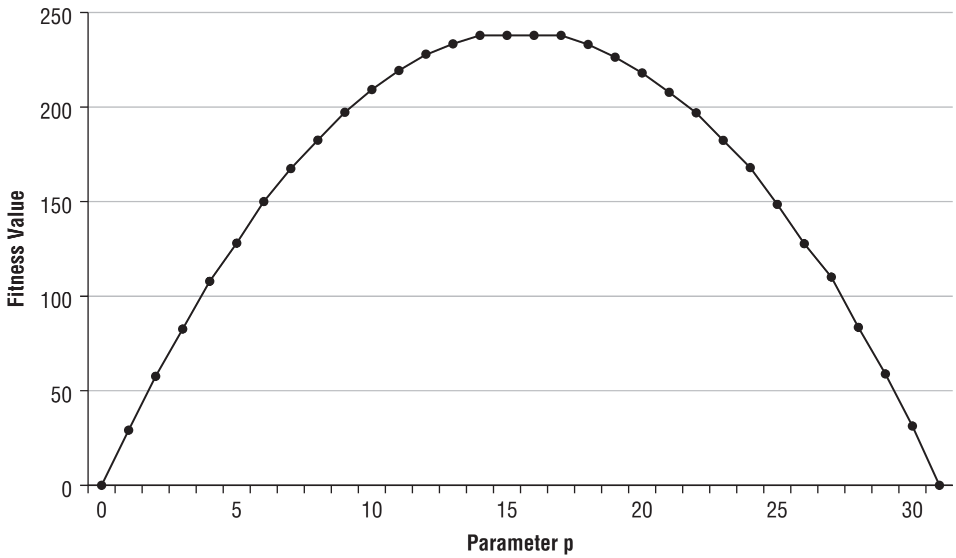


Figure 11-2: Finding the maximum of this simple function helps illustrate genetic algorithms.

Table 11-1: Ten Randomly Generated Genomes

COUNT	16	8	4	2	1	P	FITNESS
1	0	1	1	1	0	14	238
1	0	1	0	0	0	8	184
1	1	0	1	1	1	23	184
1	0	1	0	1	0	10	210
1	1	1	0	0	0	24	168
1	1	1	1	1	0	30	30
1	0	0	1	0	0	4	108
1	0	1	1	0	1	13	234
1	1	1	0	0	1	25	150
1	0	0	0	1	1	3	84

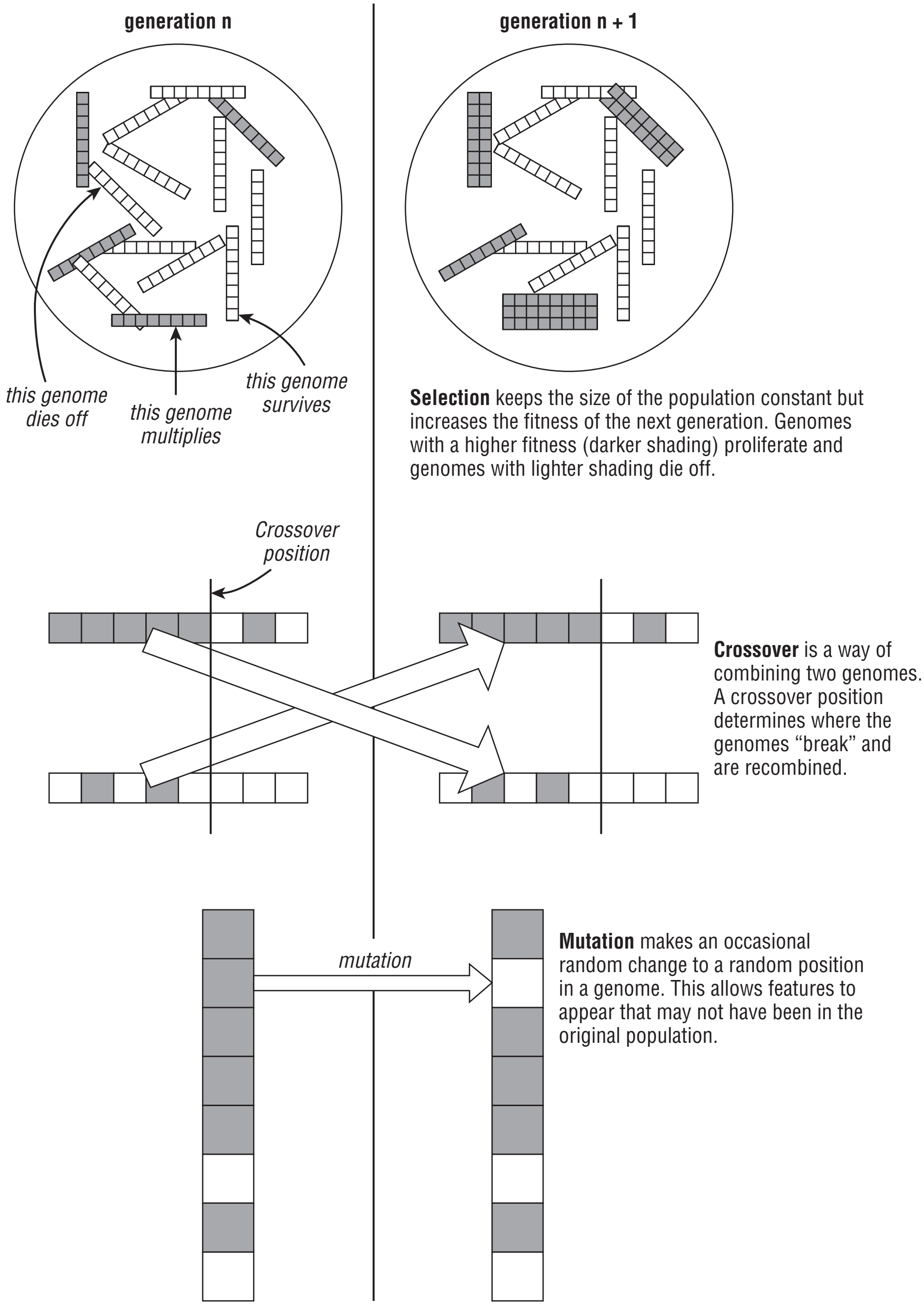
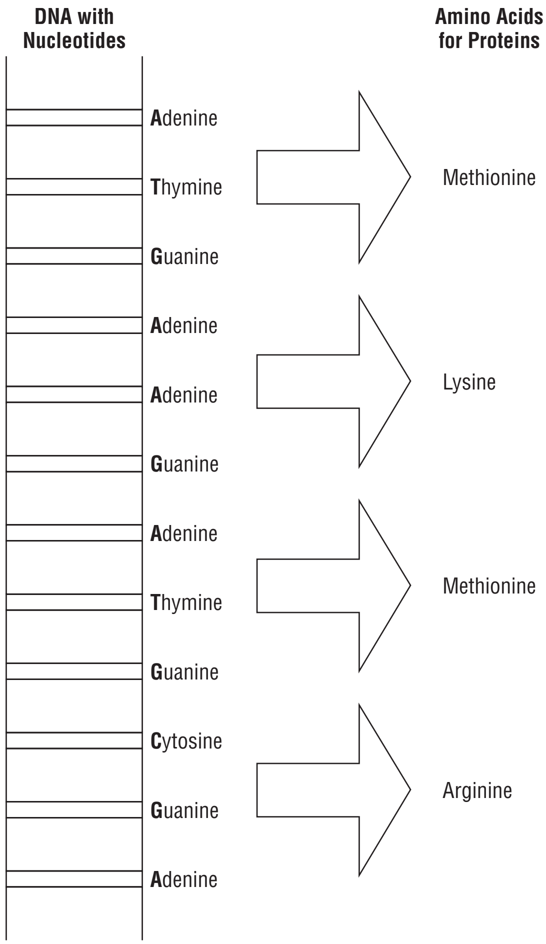


Figure 11-3: The basic operators in genetic algorithms are selection, crossover, and mutation.



Nucleotides in DNA code for amino acids that make up proteins.

Nucleotides in DNA

Table 11-2: The Population After Selection

COUNT	16	8	4	2	1	P	FITNESS
1	1	0	0	0	0	16	240
1	0	1	0	0	0	8	184
1	1	0	1	1	1	23	184
2	0	1	0	1	0	10	210
1	1	1	0	0	0	24	168
1	0	0	1	0	0	4	108
2	0	1	1	0	1	13	234
1	1	1	0	0	1	25	150

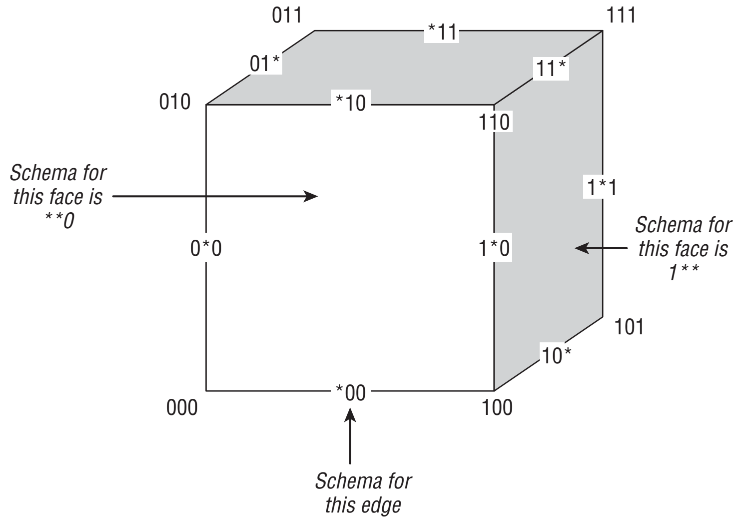


Figure 11-4: A cube is a useful representation of schemata on three bits. The corners represent the genomes, the edges represent the schemata of order 2, the faces, the schemata of order 1, and the entire cube, the schema of order 0.

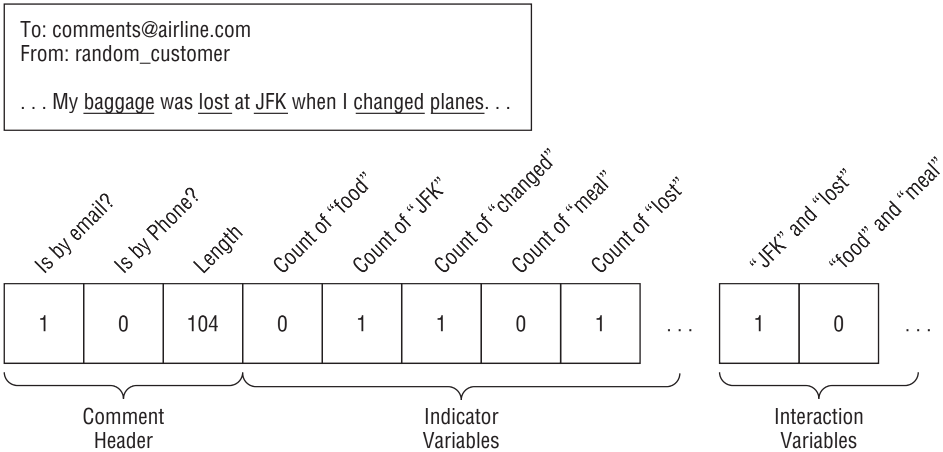


Figure 11-5: The comment signature describes the text in the comment.

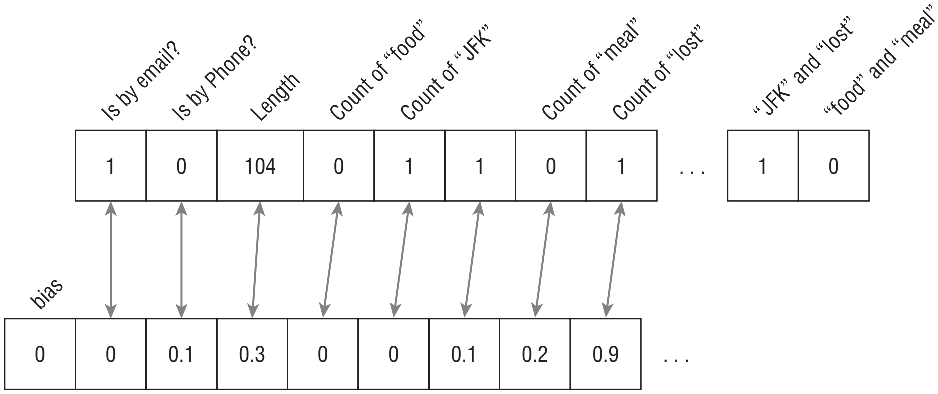


Figure 11-6: The genome has a weight for each field in the comment signature, plus an additional weight called a bias.

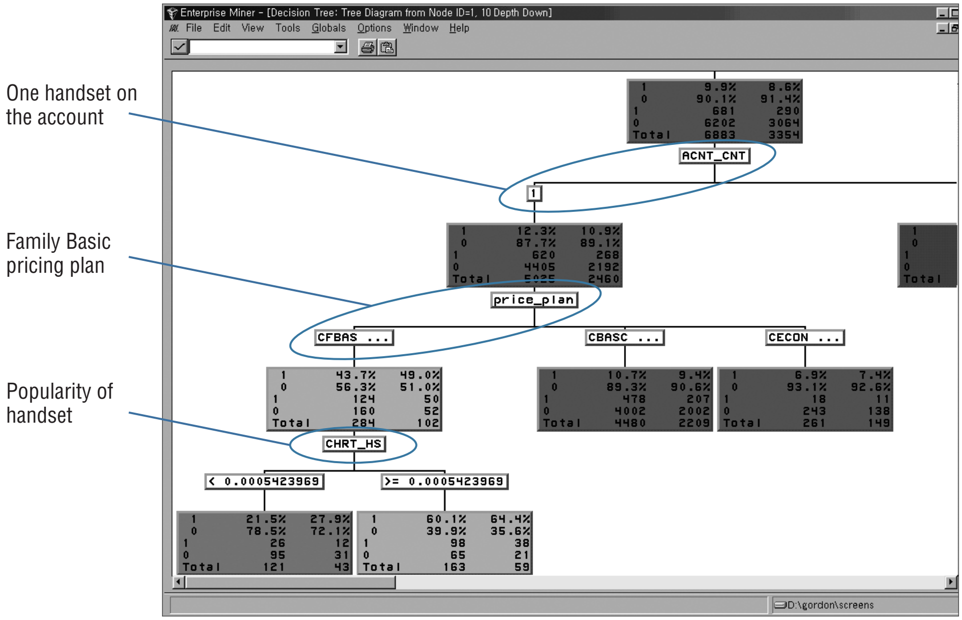
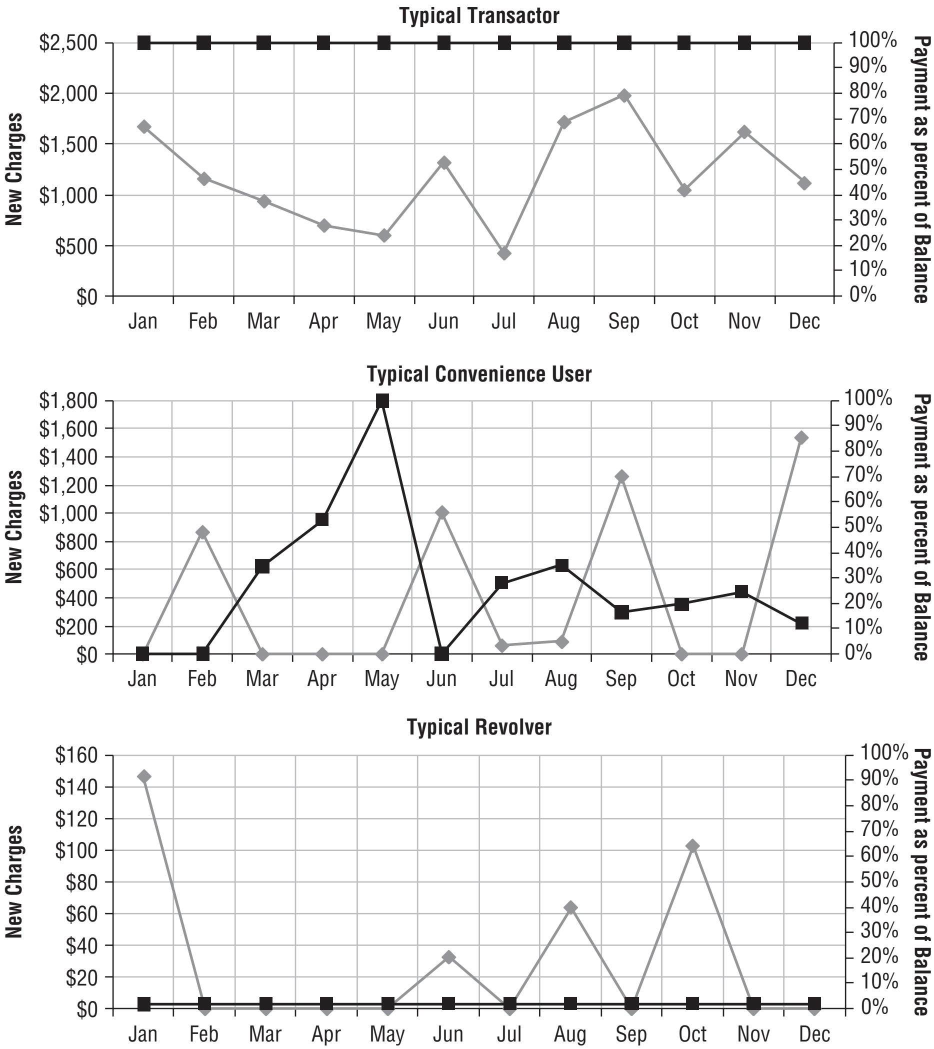


Figure 12-1: This decision tree reveals an interesting pattern, unrelated to the target variable, that is not obvious without knowledge of the business.



Three different types of credit card customers differ in their payment and usage patterns.

Credit Card Customer Graphs

	Transactor (Limit $2,000)			Convenience User (Limit $2,000)			Revolver (Limit $2,000)		
	Charge	% Of Limit	Measure	Charge	% Of Limit	Measure	Charge	% Of Limit	Measure
Jan	$1,250.44	62.5%	1.00	$1,172.51	58.6%	1.00	$0.00	0.0%	0.00
Feb	$1,546.52	77.3%	1.00	$0.00	0.0%	0.00	$135.95	6.8%	0.27
Mar	$1,661.93	83.1%	1.00	$0.00	0.0%	0.00	$90.28	4.5%	0.18
Apr	$522.87	26.1%	1.00	$47.28	2.4%	0.09	$0.00	0.0%	0.00
May	$1,937.79	96.9%	1.00	$0.00	0.0%	0.00	$25.86	1.3%	0.05
Jun	$863.30	43.2%	1.00	$738.99	36.9%	1.00	$0.00	0.0%	0.00
Jul	$841.93	42.1%	1.00	$0.00	0.0%	0.00	$113.94	5.7%	0.23
Aug	$1,237.68	61.9%	1.00	$53.56	2.7%	0.11	$0.00	0.0%	0.00
Sep	$1,741.01	87.1%	1.00	$60.57	3.0%	0.12	$0.00	0.0%	0.00
Oct	$959.30	48.0%	1.00	$1,086.34	54.3%	1.00	$151.61	7.6%	0.30
Nov	$1,954.05	97.7%	1.00	$0.00	0.0%	0.00	$88.15	4.4%	0.18
Dec	$1,051.92	52.6%	1.00	$0.00	0.0%	0.00	$0.00	0.0%	0.00
Overall			**1.00**			**0.28**			**0.10**

Charge Measure for Transactors

	Transactor (Limit $2,000)			Convenience User (Limit $2,000)			Revolver (Limit $2,000)		
	Balance	Payment	Measure	Charge	Payment	Measure	Charge	Payment	Measure
Jan	$1,250.44	$1,250.44	1.00	$1,172.51	$0.00	0.00	$1,500.00	$30.00	0.02
Feb	$1,546.52	$1,546.52	1.00	$1,172.51	$300.00	0.26	$1,620.95	$29.70	0.02
Mar	$1,661.93	$1,661.93	1.00	$872.51	$300.00	0.34	$1,696.37	$32.12	0.02
Apr	$522.87	$522.87	1.00	$619.79	$300.00	0.48	$1,680.31	$33.61	0.02
May	$1,937.79	$1,937.79	1.00	$319.79	$300.00	0.94	$1,689.37	$33.27	0.02
Jun	$863.30	$863.30	1.00	$758.77	$19.79	0.03	$1,672.73	$33.45	0.02
Jul	$841.93	$841.93	1.00	$738.99	$300.00	0.41	$1,769.95	$33.12	0.02
Aug	$1,237.68	$1,237.68	1.00	$492.55	$300.00	0.61	$1,753.39	$35.07	0.02
Sep	$1,741.01	$1,741.01	1.00	$253.12	$192.55	0.76	$1,735.85	$34.72	0.02
Oct	$959.30	$959.30	1.00	$1,146.91	$60.57	0.05	$1,870.10	$34.37	0.02
Nov	$1,954.05	$1,954.05	1.00	$1,086.34	$300.00	0.28	$1,941.07	$37.06	0.02
Dec	$1,051.92	$1,051.92	1.00	$786.34	$300.00	0.38	$1,922.54	$38.45	0.02
Overall			**1.00**			**0.38**			**0.02**

Payment Measure for Transactors

Cost Per Month	$500,000											
Number of Months	6											
Cost Per New Customer	$250.00											
Expected Discount Rate	1.0%											
Revenue Per Customer Month	$30											
Attrition Rate Per Month	5.0%											
	1	2	3	4	5	6	7	8	9	10	11	12
	Jan	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**
Cost	$500,000	$500,000	$500,000	$500,000	$500,000	$500,000	$0	$0	$0	$0	$0	$0
Starts	2,000	2,000	2,000	2,000	2,000	2,000	0	0	0	0	0	0
Attrition Rate	5%	5%	5%	5%	5%	5%	5%	5%	5%	5%	5%	5%
New Customers	1,000.0	1,000.0	1,000.0	1,000.0	1,000.0	1,000.0	0.0	0.0	0.0	0.0	0.0	0.0
Contribution to Base Customers	0.0	1,900.0	3,705.0	5,419.8	7,048.8	8,596.3	10,066.5	9,563.2	9,085.0	8,630.8	8,199.2	7,789.3
Revenue/Customer	$30	$30	$30	$30	$30	$30	$30	$30	$30	$30	$30	$30
Revenue	$30,000	$87,000	$141,150	$192,593	$241,463	$287,890	$301,995	$286,895	$272,551	$258,923	$245,977	$233,678
Cumulative Cost	$500,000	$1,000,000	$1,500,000	$2,000,000	$2,500,000	$3,000,000	$3,000,000	$3,000,000	$3,000,000	$3,000,000	$3,000,000	$3,000,000
Cumulative Revenue	$30,000	$117,000	$258,150	$450,743	$692,205	$980,095	$1,282,090	$1,568,986	$1,841,537	$2,100,460	$2,346,437	$2,580,115
Monthly Discount Rate	1.0%	1.0%	1.0%	1.0%	1.0%	1.0%	1.0%	1.0%	1.0%	1.0%	1.0%	1.0%
Net Discount Rate	1.0%	2.0%	3.0%	3.9%	4.9%	5.9%	6.8%	7.7%	8.6%	9.6%	10.5%	11.4%
Discounted Revenue	$29,700	$85,269	$136,958	$185,004	$229,629	$271,042	$281,479	$264,731	$248,980	$234,165	$220,233	$207,129
Cum Discounted Revenue	$29,700	$114,969	$251,926	$436,930	$666,559	$937,601	$1,219,081	$1,483,812	$1,732,792	$1,966,957	$2,187,190	$2,394,319
Cum Costs	$500,000	$1,000,000	$1,500,000	$2,000,000	$2,500,000	$3,000,000	$3,000,000	$3,000,000	$3,000,000	$3,000,000	$3,000,000	$3,000,000
Net Revenue	–$470,300	–$885,031	–$1,248,074	–$1,563,070	–$1,833,441	–$2,062,399	–$1,780,919	–$1,516,188	– $1,267,208	–$1,033,043	–$812,810	–$605,681

Figure 12-2: This financial spreadsheet model calculates the impact of a marketing campaign for acquiring new customers.

Cost Per Month	500000	
Number of Months	6	
Cost Per New Customer	250	
Expected Discount Rate	0.01	
Revenue Per Customer Month	30	
Attrition Rate Per Month	0.05	
	1	2
	Jan	Feb
Cost	=IF(B$8<=$B$2, $B$1, 0)	= F(C$8<=$B$2, $B$1, 0)
Starts	=B10/$B$3	=C10/$B$3
Attrition Rate	=$B$6	=$B$6
New Customers	=B11/2	=C11/2
Contribution to Base Customers	=IF(ISNUMBER(A14), (A14+A11)*(1–B12), 0)	=IF(ISNUMBER(B14), (B14+B11)*(1–C12), 0)
Revenue/Customer	=$B$5	=$B$5
Revenue	=B15*(B14+B13)	=C15*(C14+C13)
Cumulative Cost	=B10+IF(ISNUMBER(A17), A17, 0)	=C10+IF(ISNUMBER(B17), B17, 0)
Cumulative Revenue	=B16+IF(ISNUMBER(A18), A18, 0)	=C16+IF(ISNUMBER(B18), B18, 0)
Monthly Discount Rate	=$B$4	=$B$4
Net Discount Rate	=1–IF(ISNUMBER(A20), 1–A20, 1)*(1–B19)	=1–IF(ISNUMBER(B20), 1–B20, 1)*(1–C19)
Discounted Revenue	=B16*(1–B20)	=C16*(1–C20)
Cum Discounted Revenue	=IF(ISNUMBER(A22), A22, 0)+B21	=IF(ISNUMBER(B22), B22, 0)+C21
Cum Costs	=IF(ISNUMBER(A23), A23, 0)+B10	=IF(ISNUMBER(B23), B23, 0)+C10
Net Revenue	=B22–B23	=C22–C23

Figure 12-3: The spreadsheet performs the calculations needed for a financial spreadsheet model.

Table 12-1: Various Financial Measures for the Campaign

YEAR	COST	REVENUE	NET REVENUE	NUMBER OF CUSTOMERS
1	$3,000,000	$2,394,319	–$605,681	7,789.3
2	$3,000,000	$4,100,195	$1,100,195	4,209.0
3	$3,000,000	$4,917,253	$1,917,253	2,274.4
4	$3,000,000	$5,308,597	$2,308,597	1,229.0
5	$3,000,000	$5,496,038	$2,496,038	664.1

Cost Per Month	=IF(C1="uniform",RAND()*(E1-D1)+D1,IF(C1="normal",NORMINV(RAND(),D1,E1)))/B2	uniform	2700000	3300000
Number of Months	6	fixed		
Cost Per New Customer	=IF(C3="uniform",RAND()*(E3-D3)+D3,IF(C3="normal",NORMINV(RAND(),D3,E3)))	uniform	220	280
Expected Discount Rate	=IF(C4="uniform",RAND()*(E4-D4)+D4,IF(C4="normal",NORMINV(RAND(),D4,E4)))	normal	0.01	0.0015
Revenue Per Customer Month	=IF(C5="uniform",RAND()*(E5-D5)+D5,IF(C5="normal",NORMINV(RAND(),D5,E5)))	uniform	25	35
Attrition Rate Per Month	=IF(C6="uniform",RAND()*(E6-D6)+D6,IF(C6="normal",NORMINV(RAND(),D6,E6)))	normal	0.05	0.002

Figure 12-4: This spreadsheet introduces "uncertainty" into the financial model by having the inputs come from various distributions.

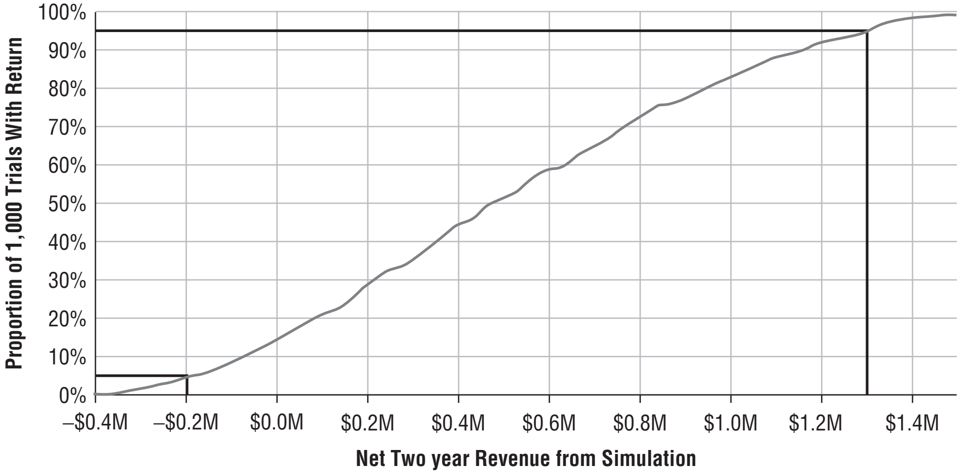


Figure 12-5: This chart shows the distribution of net revenue after two years, along with lines showing the 5 percent and 95 percent confidence range.

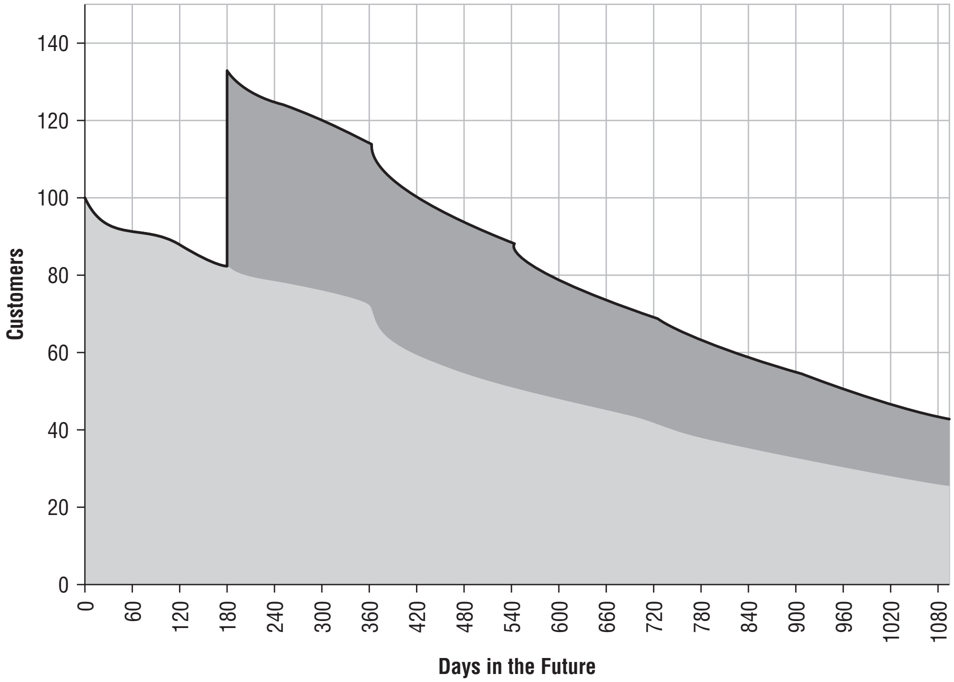


Figure 12-6: Assuming that 100 customers start on the first day of the forecast and 50 more start half a year later, survival curves determine how many customers are expected to still be around on any day in the future.

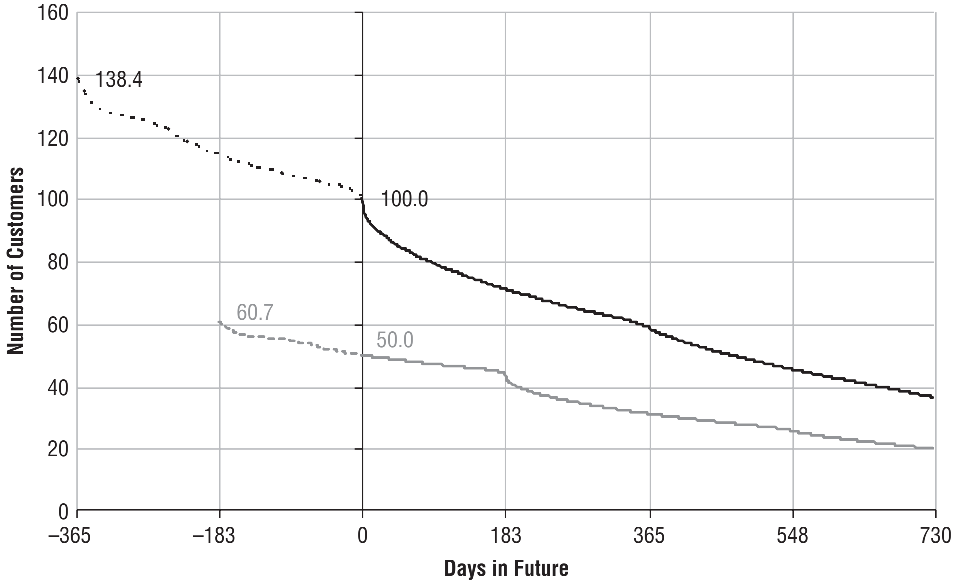
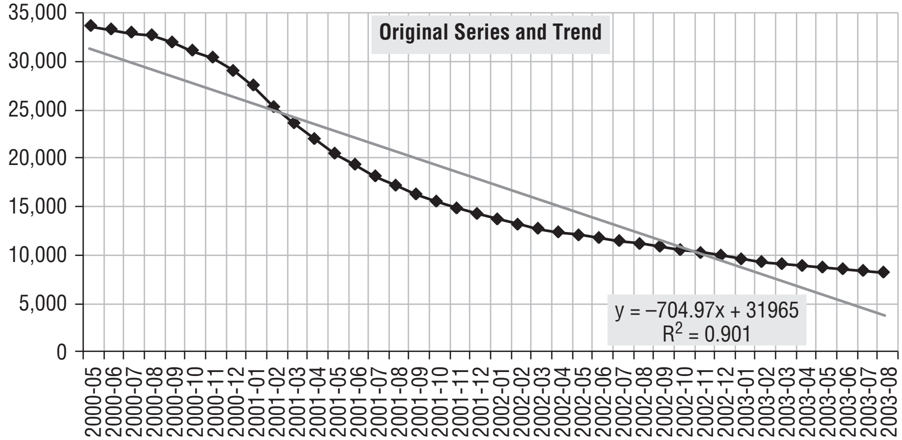
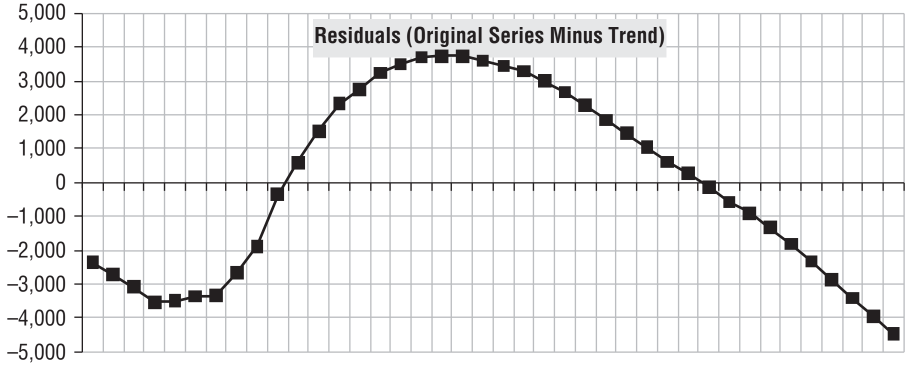


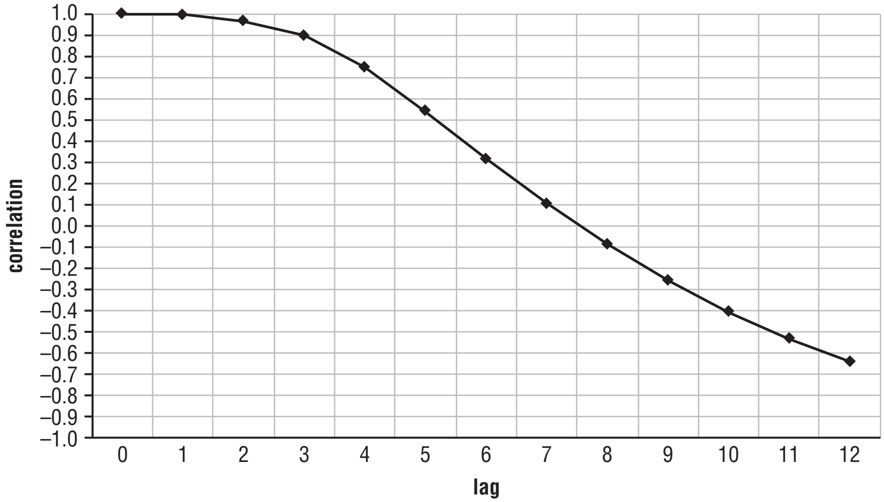
Figure 12-7: For customers who are active today, the survival curve can be retrofitted to the past and then extended into the future.



This chart shows a time series with its trend line.



The difference between the data and the trend line is called the *residuals*.



This chart shows a correlogram, which is the correlation coefficient of a time series using different lags.

Correlogram

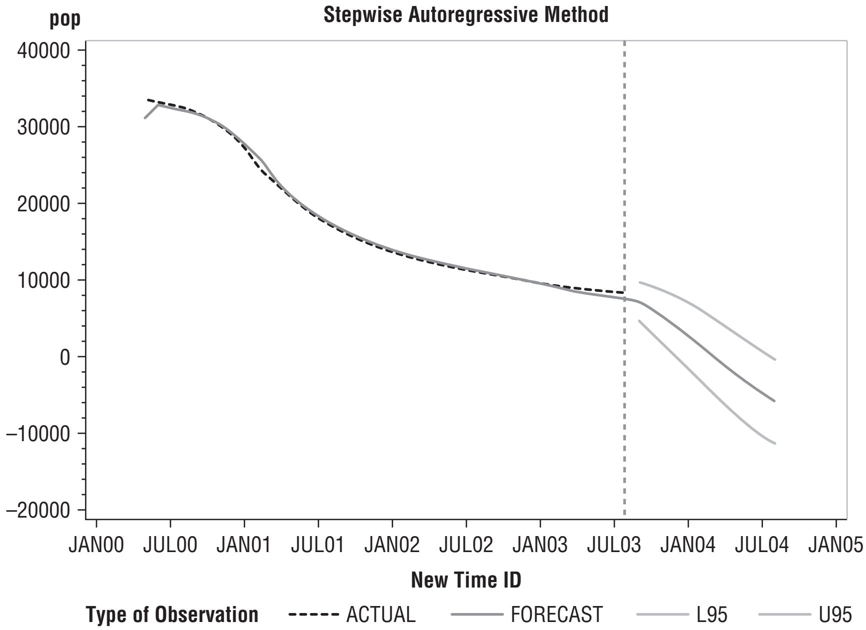
$$t = \frac{r}{\sqrt{(1-r^2)/(N-2)}}$$

Equation 25

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-12}.$$

$$y_t = -169.56 + 1.02\, y_{t-1} - 0.09 y_{t-12}$$

Equation 26



ARIMA forecasts often do well for the period of the forecast, but do less well when extrapolating into the future.

29	American Dreams
16	Bohemian Mix
07	Money & Brains
31	Urban Achievers
04	Young Digerati

Neighbor Segment 1

29	American Dreams
16	Bohemian Mix
07	Money & Brains
26	The Cosmopolitans
04	Young Digerati

Neighbor Segment 2

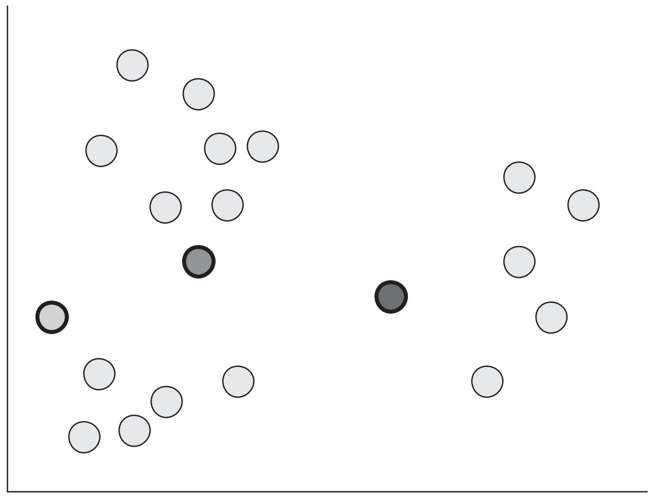


Figure 13-1: Three data points have been chosen as cluster seeds.

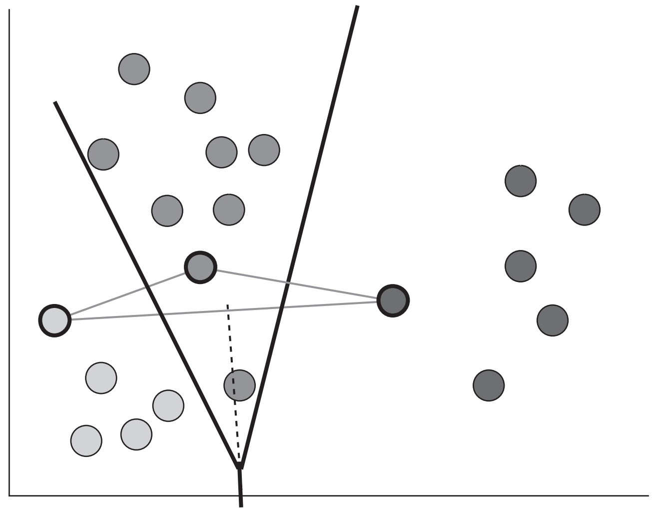


Figure 13-2: The initial clusters are formed by assigning each data point to the closest seed.

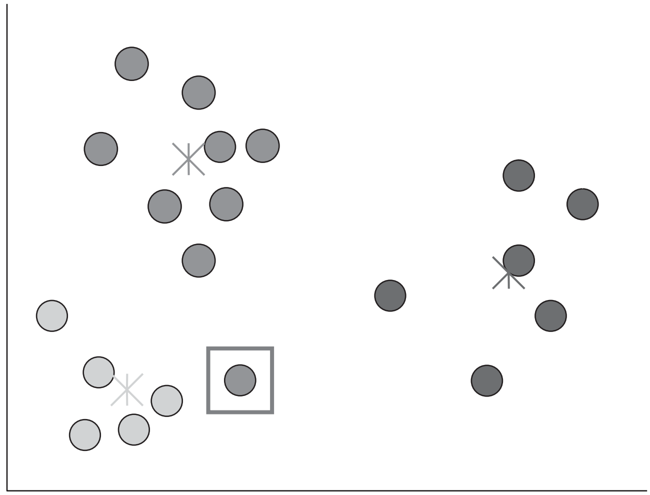


Figure 13-3: In the update step, the cluster centroid is calculated as the average value of the cluster members.

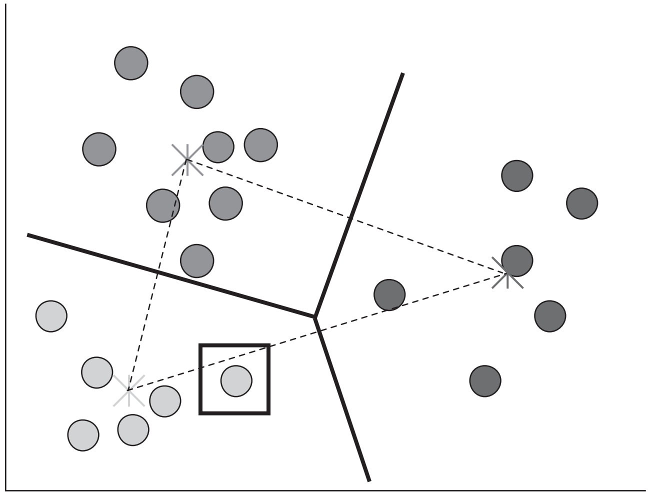
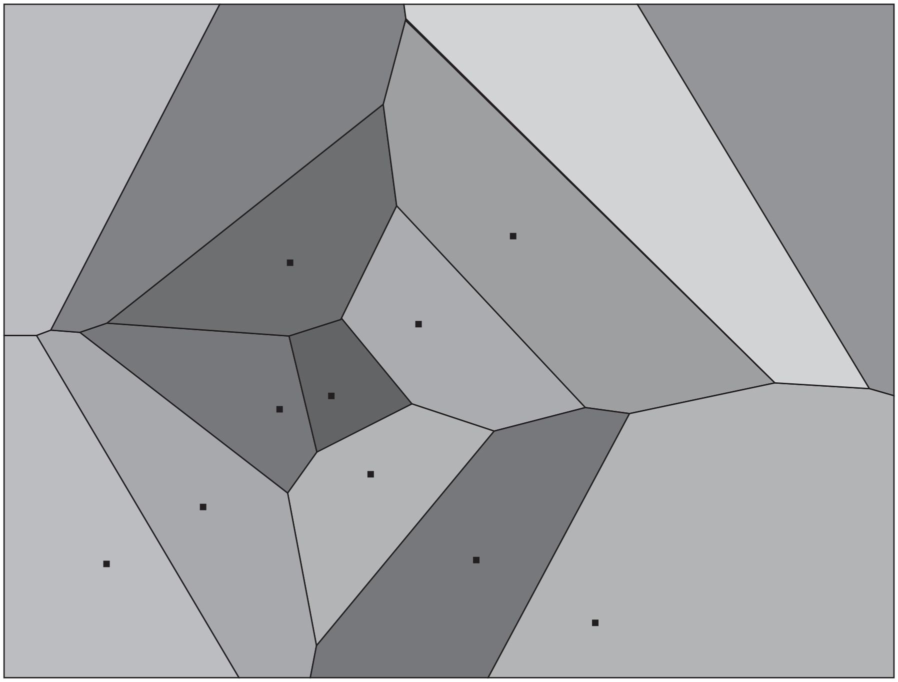


Figure 13-4: The k-means algorithm terminates when no records are reassigned following the latest relocation of the centroids.



The points in this diagram could represent stations on two metro lines that cross near the center of the map.

Voronoi Diagram

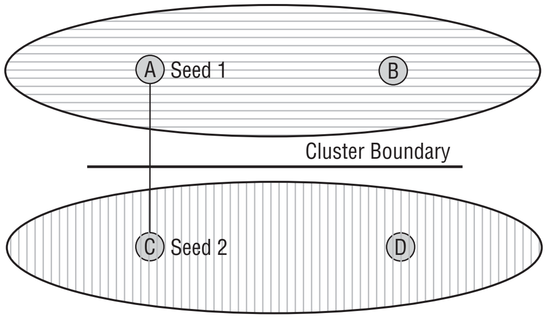


Figure 13-5: With K=2, choosing A and C as the cluster seeds leads to one cluster containing A and B and another containing C and D, which is clearly not the best pair of clusters.

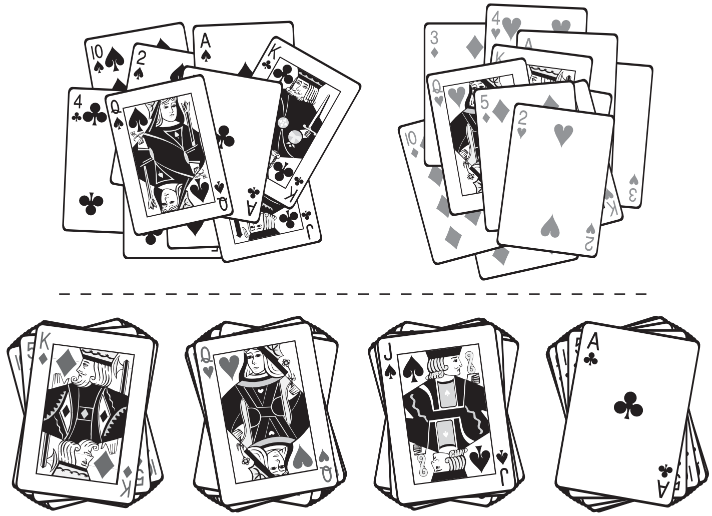


Figure 13-6: These examples of clusters of size 2 and 4 in a deck of playing cards illustrate that there is no one correct clustering.

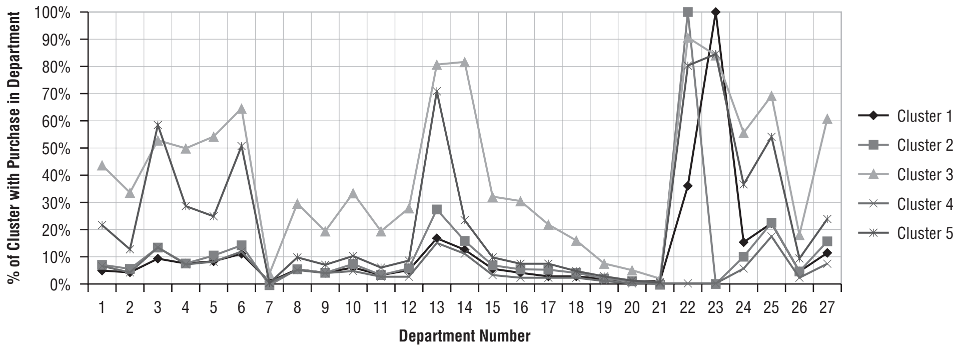


Figure 13-7: This parallel coordinates chart shows five clusters with the percentage of shoppers who have made a purchase in each department.

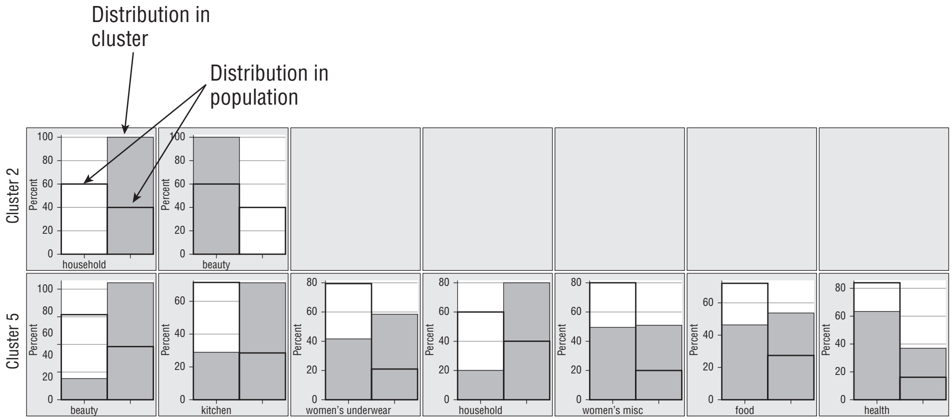


Figure 13-8: This chart compares the distribution of purchasers and non-purchasers in two clusters with the distribution in the overall population.

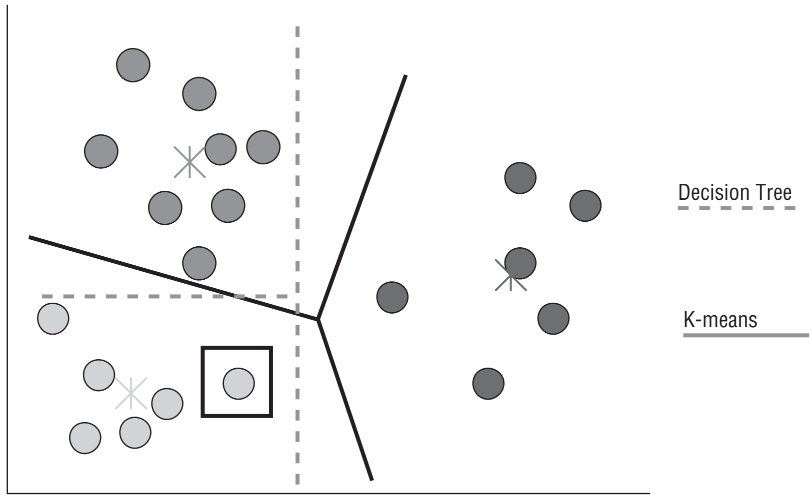


Figure 13-9: The directed clusters found by decision trees have boundaries that are parallel to the axes.

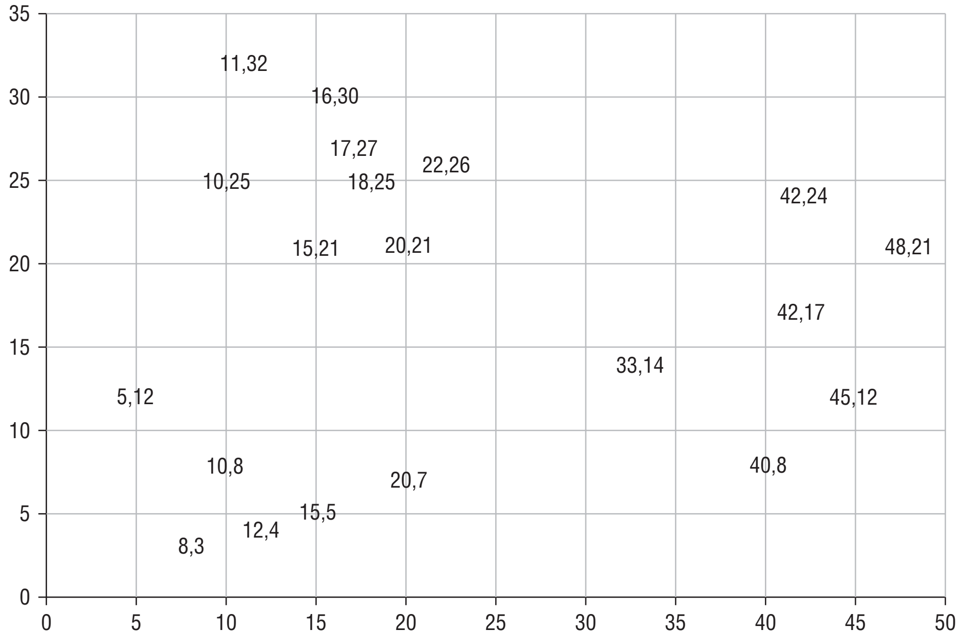


Figure 13-10: The distances used to illustrate the silhouette measure are based on the (x,y) coordinates shown here.

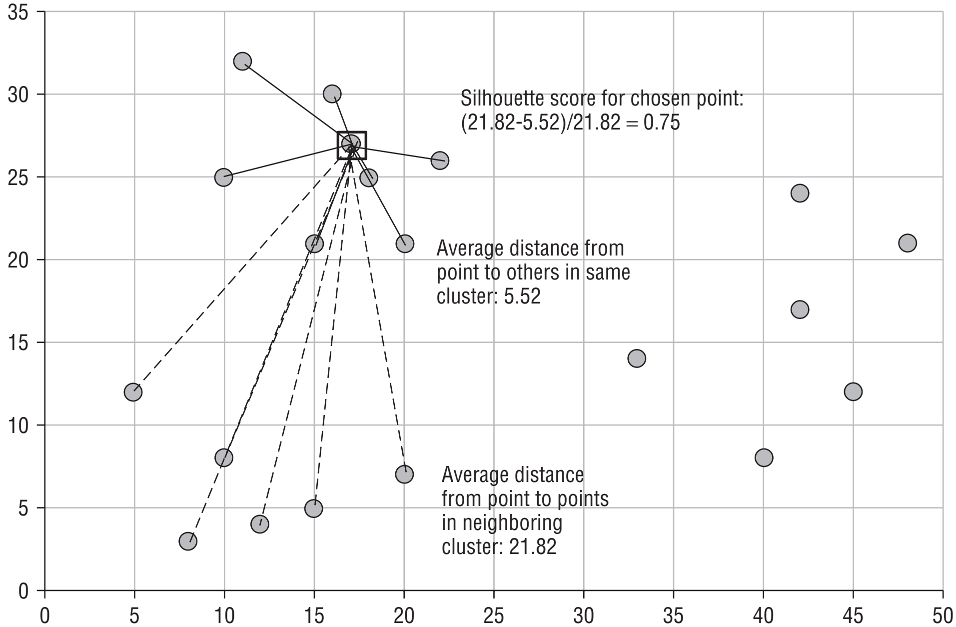


Figure 13-11: The dissimilarity score for a point depends on its distance from members of its own cluster and its distance from members of its neighboring cluster.

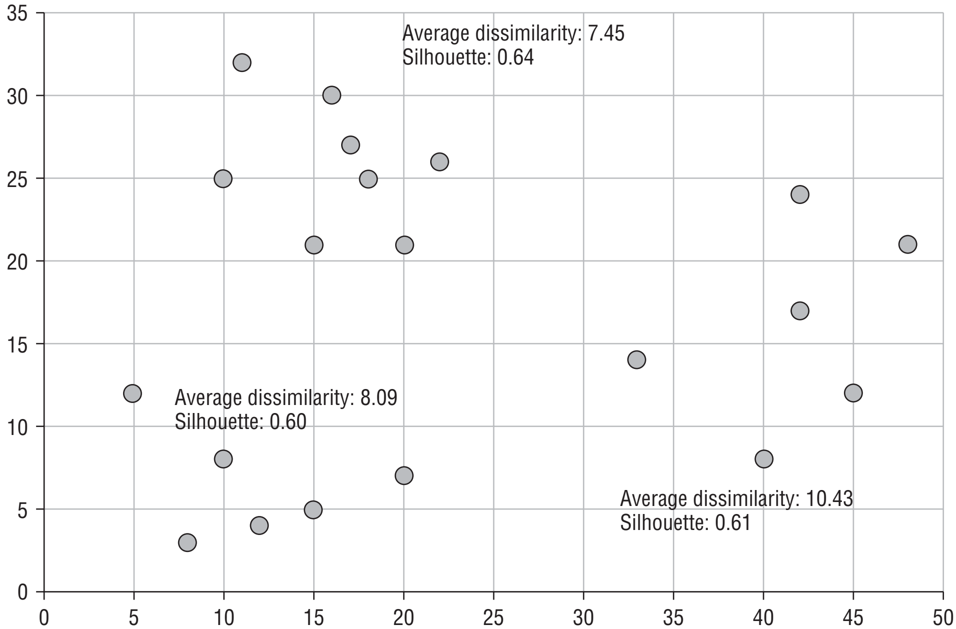


Figure 13-12: The silhouette scores of the cluster members are averaged to obtain the cluster silhouette.

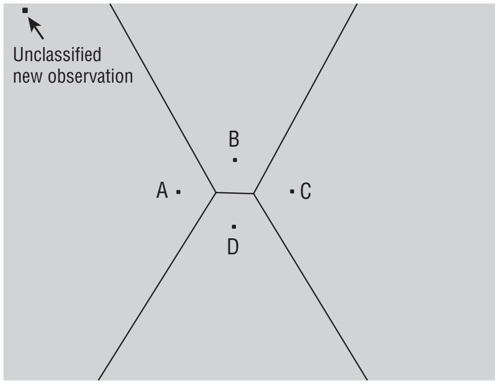


Figure 13-13: Should the new record really be assigned to Cluster A?

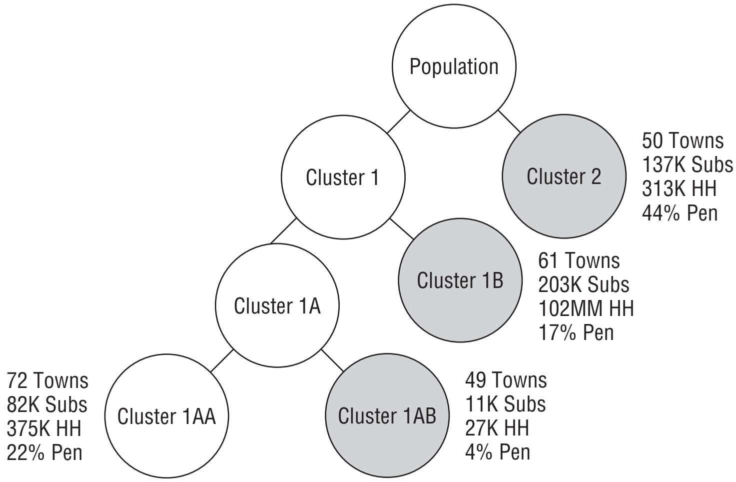


Figure 13-14: A cluster tree divides towns served by the *Boston Globe* into four distinct groups.

Table 13-1: Towns in the *City* and *West 1* Editorial Zones

TOWN	EDITORIAL ZONE	CLUSTER ASSIGNMENT
Boston	City	1B
Brookline	City	2
Cambridge	City	1B
Somerville	City	1B
Needham	West 1	2
Newton	West 1	2
Waltham	West 1	1B
Watertown	West 1	1B
Wellesley	West 1	2
Weston	West 1	2

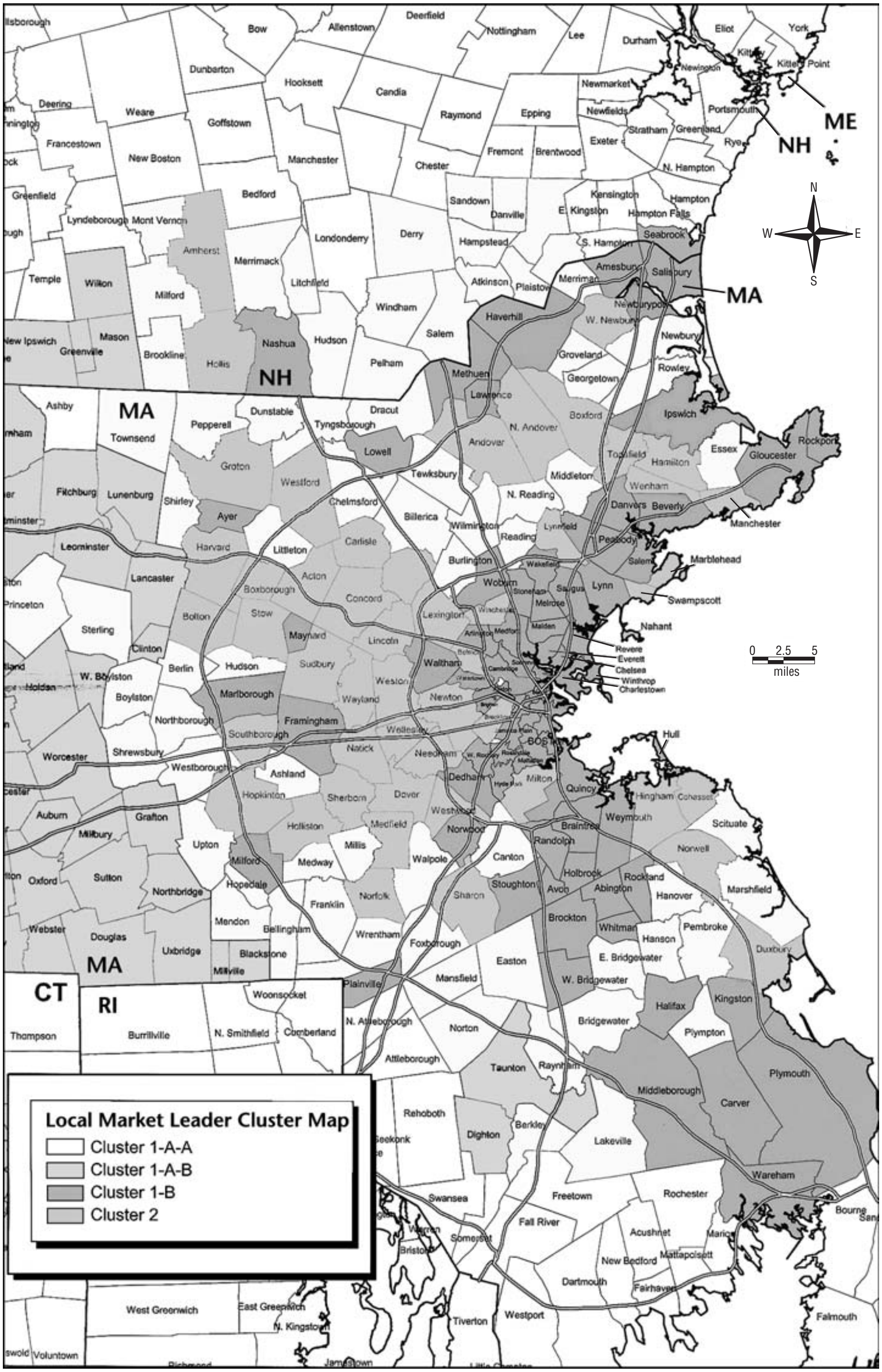
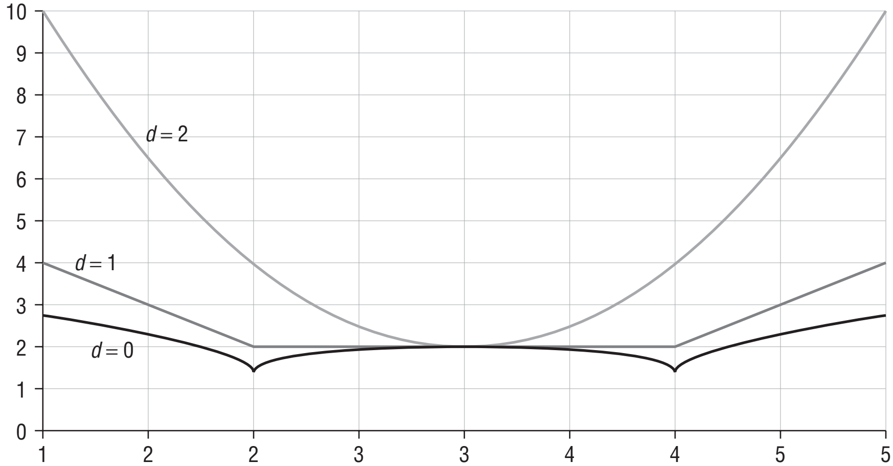


Figure 13-15: The map shows how the demographic clusters are distributed on a map of the *Globe's* coverage area.

$$(|\Delta X_1|^d + |\Delta X_2|^d + \cdots + |\Delta X_n|^d)^{1/d}.$$

Equation 27



The total distance to the points 2 and 4 is minimized at different points for different values of d using $|A - B|^d$ as the distance function.

Relationships Between Distance and Means, Medians, and Modes

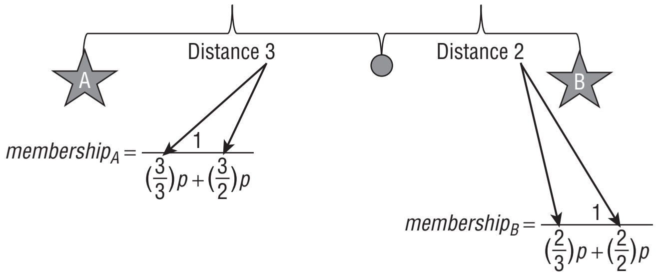


Figure 13-16: The data point shown here is to be assigned fuzzy membership in clusters A and B, which are represented by their centroids.

Table 13-2: Fuzzy Membership in A and B for Different Values of *P*

P	MEMBERSHIP IN A	MEMBERSHIP IN B
0	0.50	0.50
1	0.40	0.60
2	0.31	0.69
3	0.23	0.77

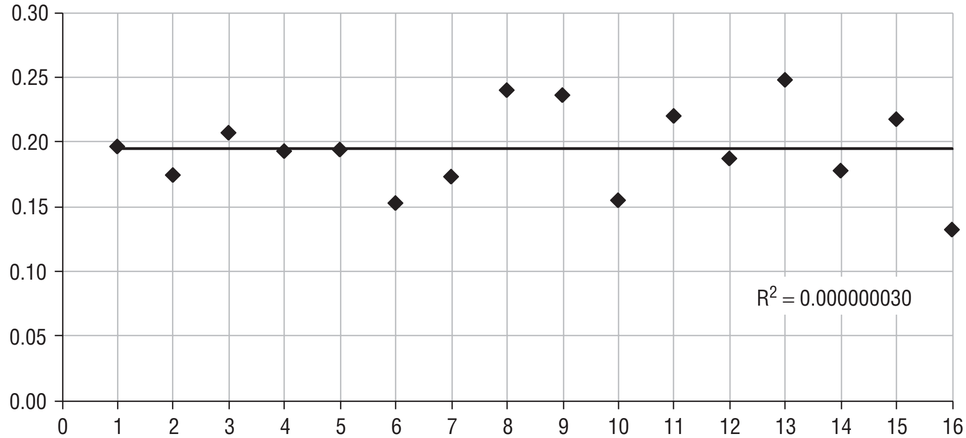


Figure 14-1: The line looks like a pretty good fit, but the R^2 value does not seem to agree. Sometimes measures of goodness do not do such a good job.

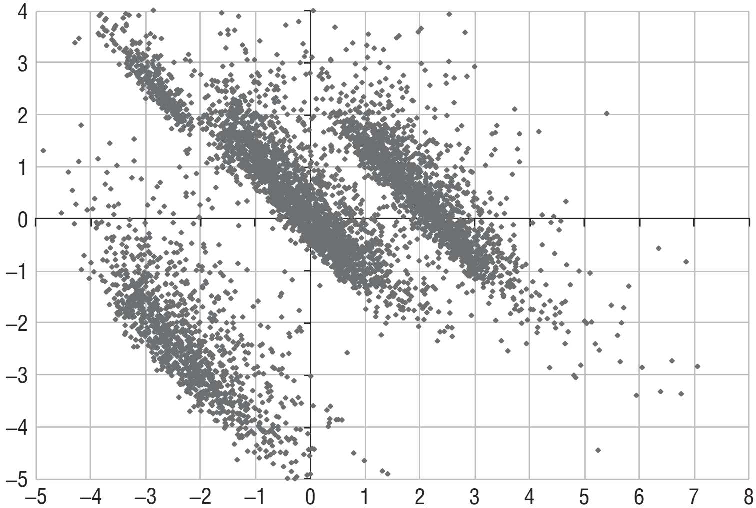


Figure 14-2: How many clusters can you see? There is no right answer.

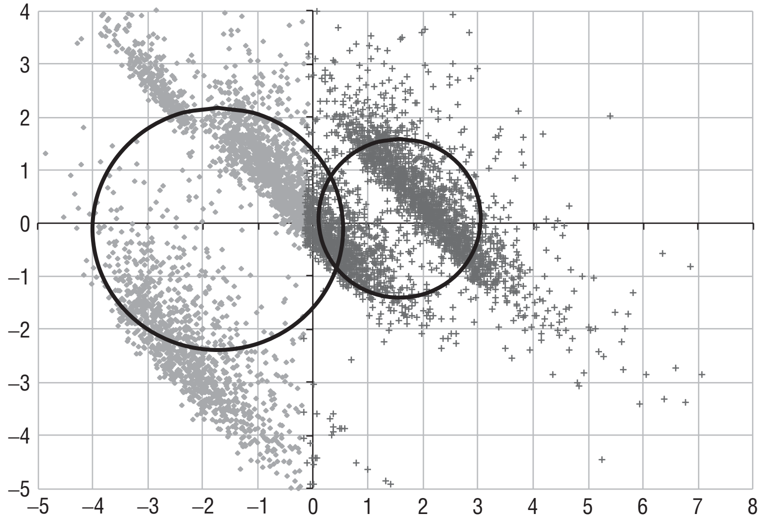


Figure 14-3: Is this really the best way to split the data into two clusters?

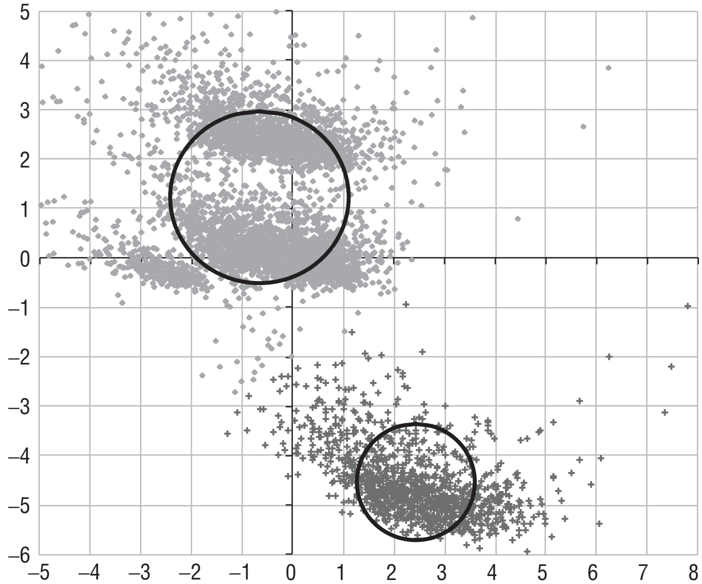


Figure 14-4: Much more intuitive clusters are generated after applying a simple linear transformation.

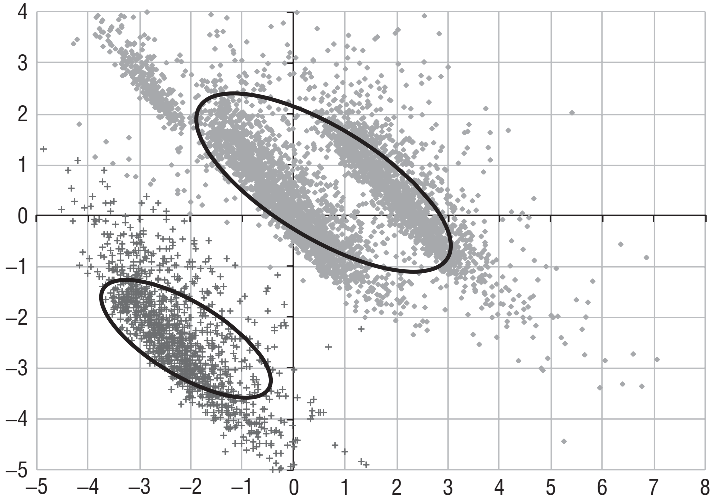


Figure 14-5: Back on the original data points, the clusters are characterized by ellipses rather than circles, and the ellipses are much more intuitive.

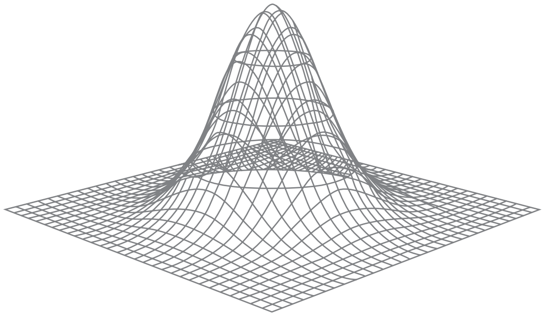


Figure 14-6: The normal distribution can be generalized to two or more dimensions.

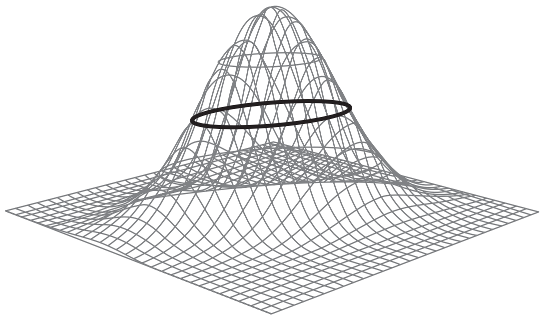


Figure 14-7: The cross-section for the normalized distribution in two dimensions is an ellipse.

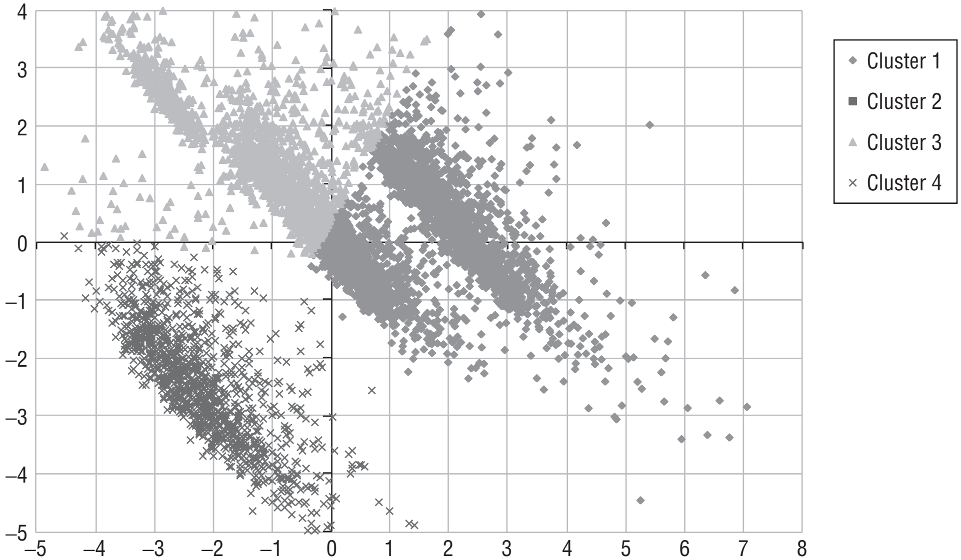


Figure 14-8: Four k-means clusters identify one of the clusters (on the lower left), but do not do a good job on the rest of the data.

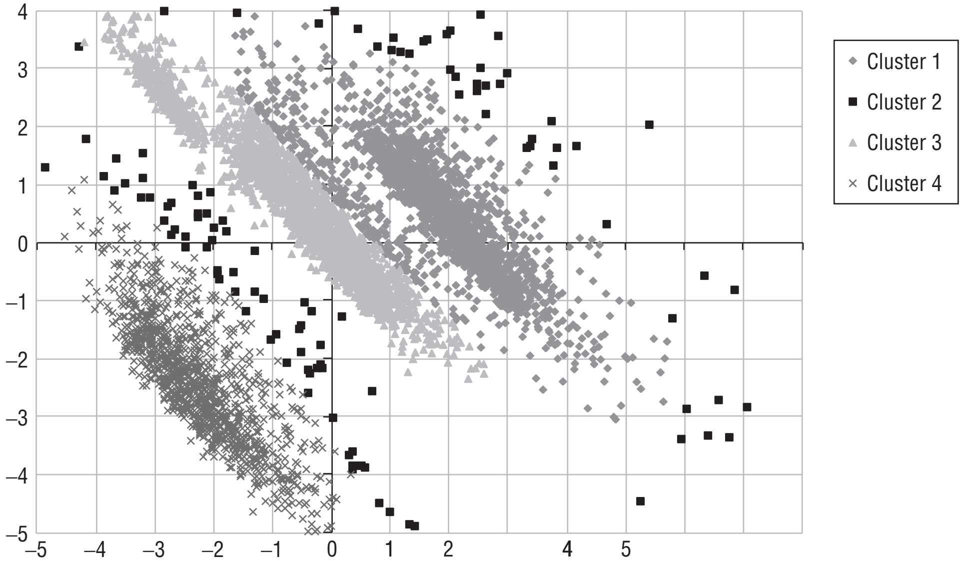


Figure 14-9: Four GMM clusters do a pretty good job of finding the obvious clusters in the data.

Table 14-1: A Contingency Table for the Chi-Square Calculation for Divisive Clustering on Categorical Variables

VARIABLE A	LEFT CHILD	RIGHT CHILD
Val 1	<count>	<count>
Val 2	<count>	<count>
…		
Val n	<count>	<count>

Table 14-2: First Level of Hierarchical Clustering, Combining Ages that are One Year Apart

AGE	DISTANCE 1
1	[1]
3	[3]
5	[5]
8	[8-9]
9	[8-9]
11	[11-13]
12	[11-13]
13	[11-13]
37	[37]
43	[43]
45	[45]
49	[49]
51	[51]
65	[65]

Table 14-3: Clustering of 15 Ages into 3 Clusters

AGE	DISTANCE 1	DISTANCE 2	DISTANCE 3	DISTANCE 4	DISTANCE 7
1	[1]	[1-5]	[1-13]	[1-13]	[1-13]
3	[3]	[1-5]	[1-13]	[1-13]	[1-13]
5	[5]	[1-5]	[1-13]	[1-13]	[1-13]
8	[8-9]	[8-13]	[1-13]	[1-13]	[1-13]
9	[8-9]	[8-13]	[1-13]	[1-13]	[1-13]
11	[11-13]	[8-13]	[1-13]	[1-13]	[1-13]
12	[11-13]	[8-13]	[1-13]	[1-13]	[1-13]
13	[11-13]	[8-13]	[1-13]	[1-13]	[1-13]
37	[37]	[37]	[37]	[37]	[37-51]
43	[43]	[43-45]	[43-45]	[43-51]	[37-51]
45	[45]	[43-45]	[43-45]	[43-51]	[37-51]
49	[49]	[49-51]	[49-51]	[43-51]	[37-51]
51	[51]	[49-51]	[49-51]	[43-51]	[37-51]
65	[65]	[65]	[65]	[65]	[65]

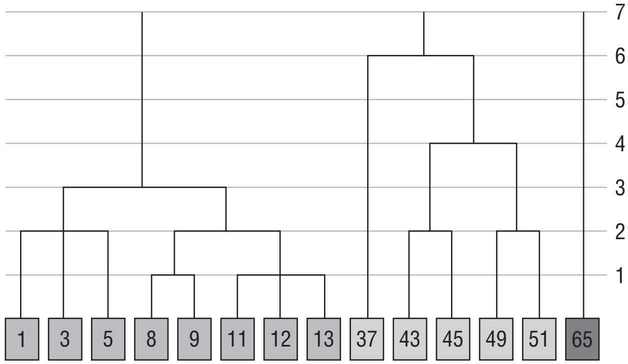


Figure 14-10: This visualization, called a dendogram, shows the clusters created by hierarchical clustering of ages.

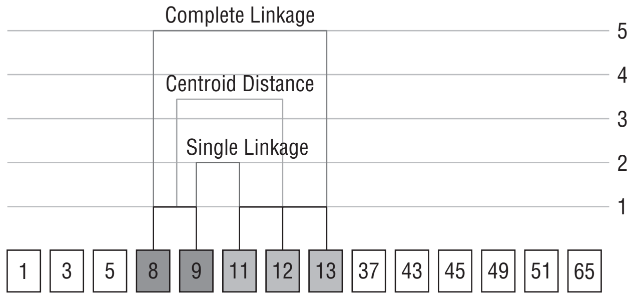


Figure 14-11: Single linkage, complete linkage, and centroid distance are three ways of combining clusters when they contain more than one data record.

Table 14-4: Distance Matrix for Ages

AGES	1	3	5	8	9	11	12	13	37	43	45	49	51	65
1		2	4	7	8	10	11	12	36	42	44	48	50	64
3	2		2	5	6	8	9	10	34	40	42	46	48	62
5	4	2		3	4	6	7	8	32	38	40	44	46	60
8	7	5	3		1	3	4	5	29	35	37	41	43	57
9	8	6	4	1		2	3	4	28	34	36	40	42	56
11	10	8	6	3	2		1	2	26	32	34	38	40	54
12	11	9	7	4	3	1		1	25	31	33	37	39	53
13	12	10	8	5	4	2	1		24	30	32	36	38	52
37	36	34	32	29	28	26	25	24		6	8	12	14	28
43	42	40	38	35	34	32	31	30	6		2	6	8	22
45	44	42	40	37	36	34	33	32	8	2		4	6	20
49	48	46	44	41	40	38	37	36	12	6	4		2	16
51	50	48	46	43	42	40	39	38	14	8	6	2		14
65	64	62	60	57	56	54	53	52	28	22	20	16	14	

Table 14-5: The Distance Matrix for the Ages After Combining 8- and 9-Year-Olds

AGES	1	3	5	8&9	11	12	13	37	43	45	49	51	65
1		2	4	7	10	11	12	36	42	44	9	50	64
3	2		2	5	8	9	10	34	40	42	46	48	62
5	4	2		3	6	7	8	32	38	40	44	46	60
8&9	7	5	3		1	2	3	4	28	34	36	40	42
11	10	8	6	2		1	2	26	32	34	38	40	54
12	11	9	7	3	1		1	25	31	33	37	39	53
13	12	10	8	4	2	1		24	30	32	36	38	52
37	36	34	32	28	26	25	24		6	8	12	14	28
43	42	40	38	34	32	31	30	6		2	6	8	22
45	44	42	40	36	34	33	32	8	2		4	6	20
49	48	46	44	40	38	37	36	12	6	4		2	16
51	50	48	46	42	40	39	38	14	8	6	2		14
65	64	62	60	56	54	53	52	28	22	20	16	14	

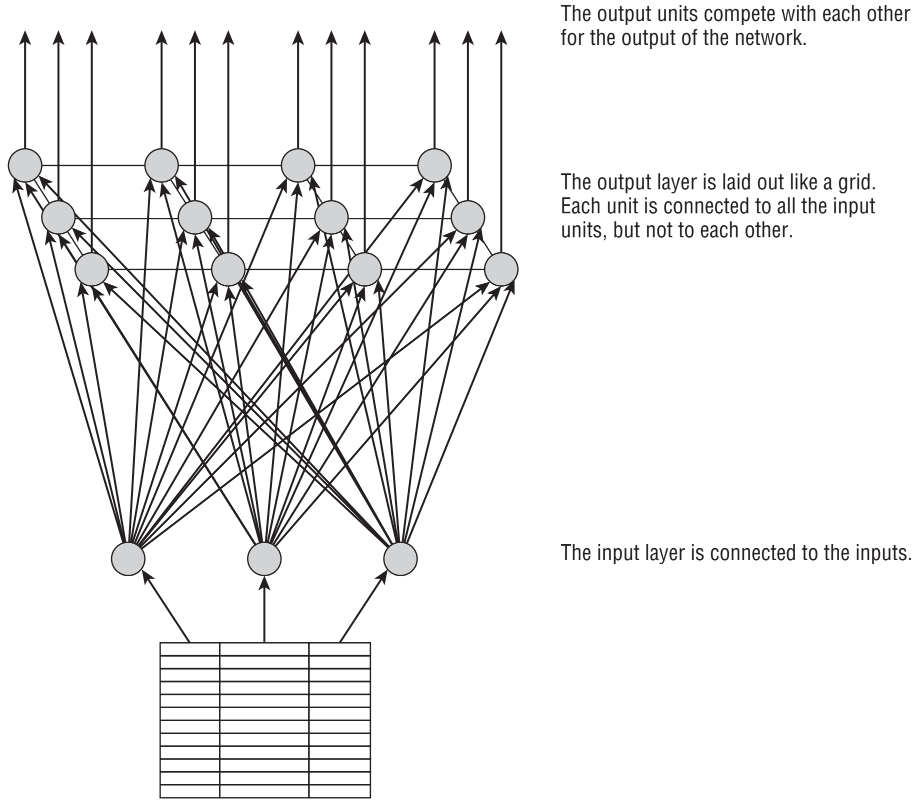


Figure 14-12: The self-organizing map is a special kind of neural network that can be used to detect clusters.

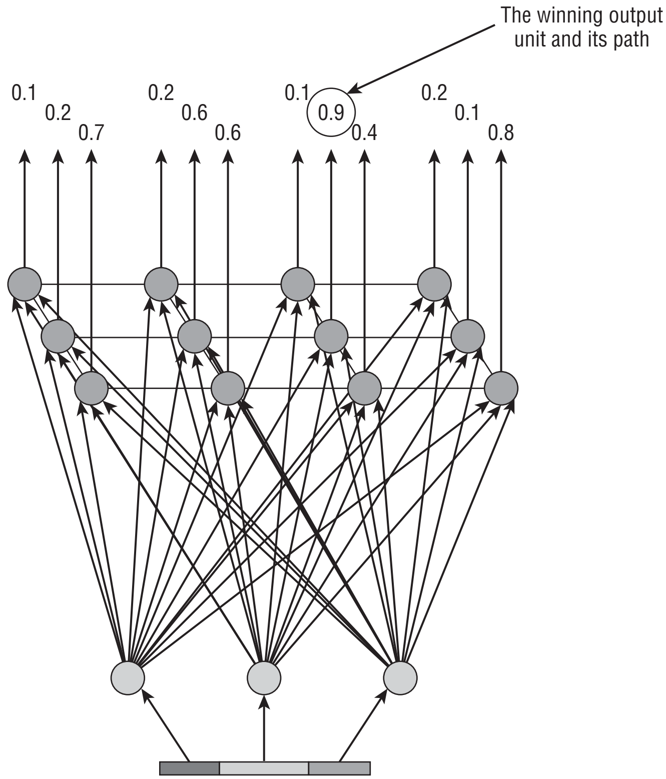


Figure 14-13: An SOM finds the output unit that does the best job of recognizing a particular input.

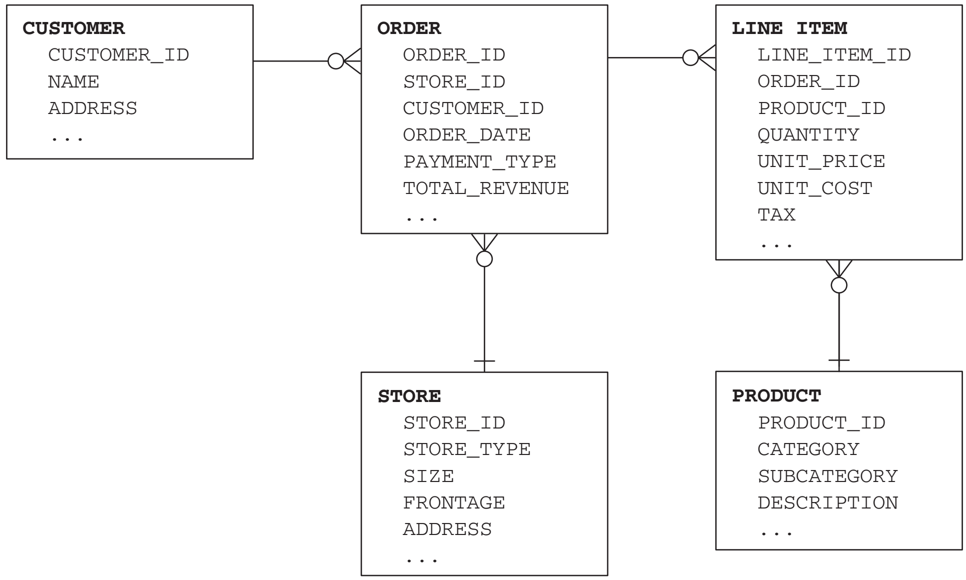


Figure 15-1: A logical data model for transaction-level market basket data has tables for the important entities related to market basket data.

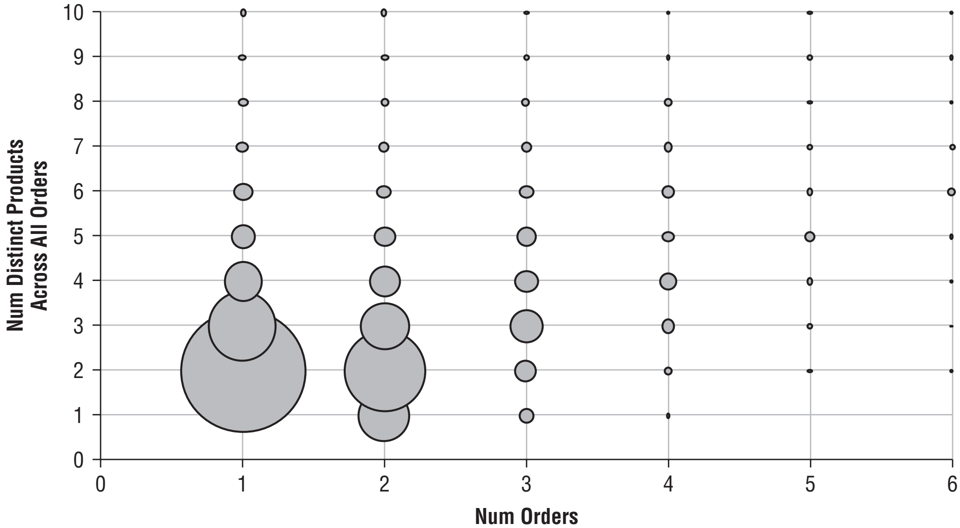


Figure 15-2: This bubble plot shows the breadth of customer relationships by the depth of the relationship.

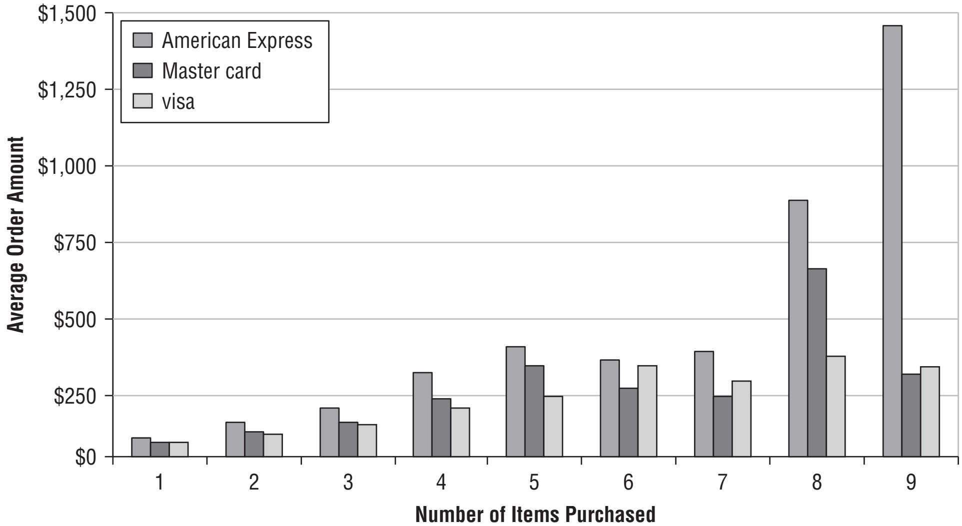


Figure 15-3: This chart shows the average amount spent by credit card type based on the number of items in the order for one particular retailer.

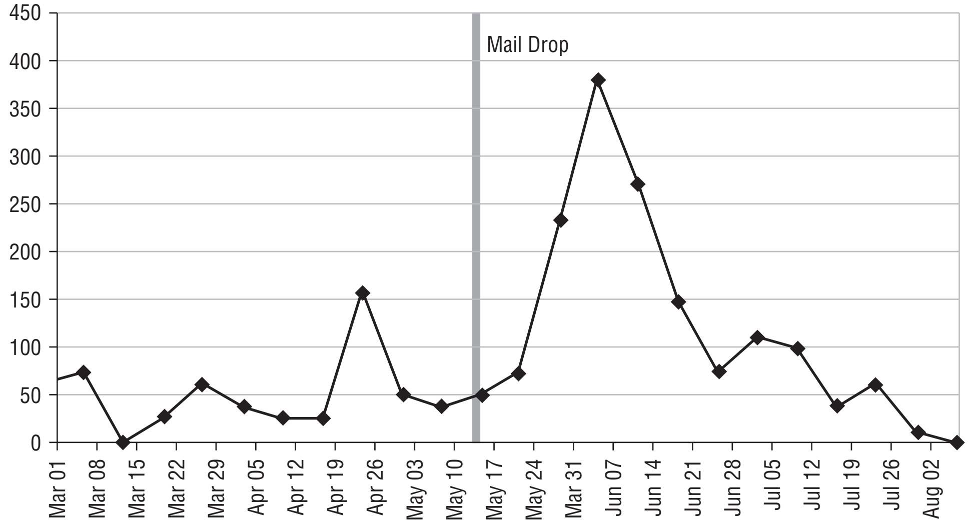


Figure 15-4: Showing marketing interventions and product sales on the same chart makes seeing effects of marketing efforts possible.

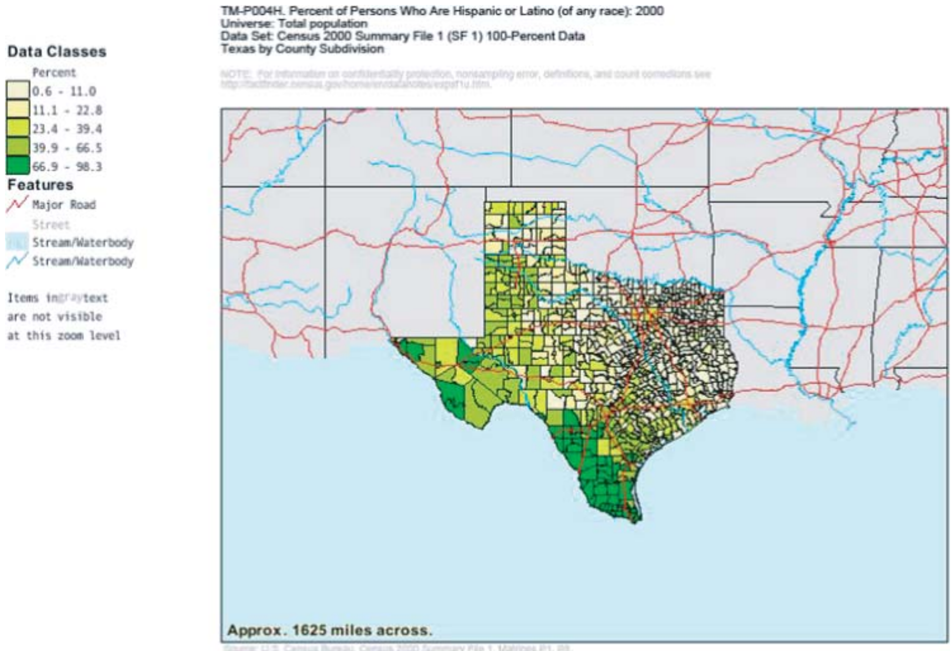


Figure 15-5: The proportion of Hispanics by county in Texas is quite high near the Mexican border, and then declines throughout the rest of the state.

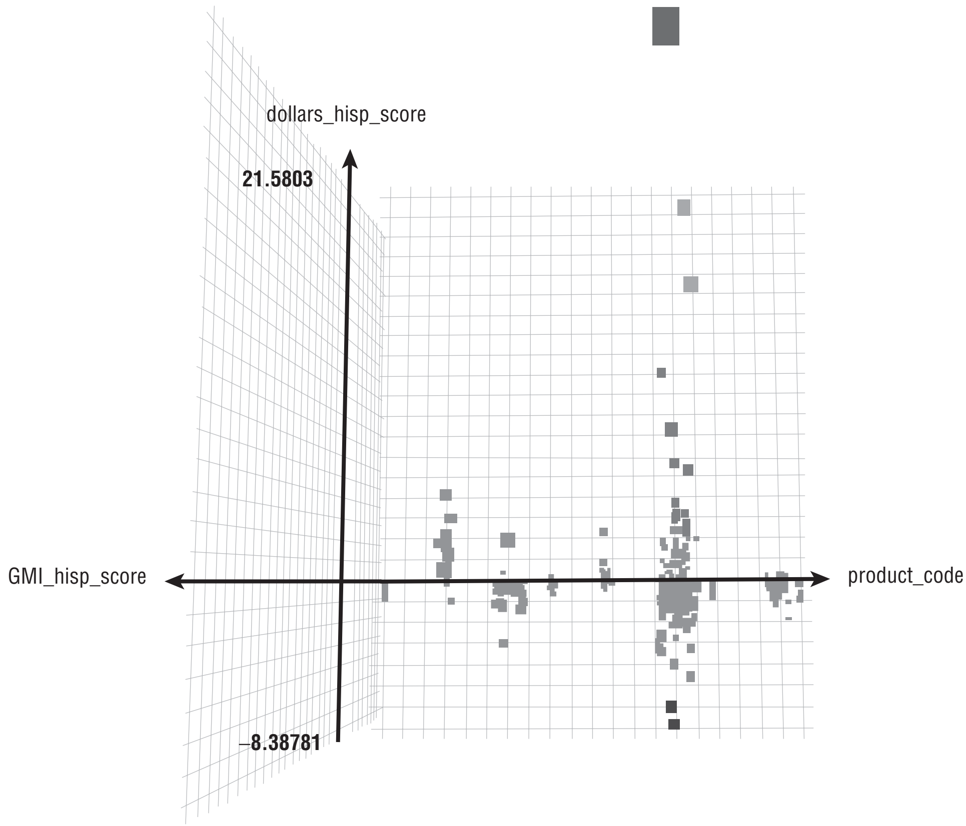


Figure 15-6: This chart shows that one product is both popular (because the cube is big) and has a high preference in Hispanic stores.

Table 15-1: Grocery Point-of-Sale Transactions

CUSTOMER	ITEMS
1	Orange juice, soda
2	Milk, orange juice, window cleaner
3	Orange juice, detergent
4	Orange juice, detergent, soda
5	Window cleaner, soda

Table 15-2: Co-Occurrence of Products

	OJ	WINDOW CLEANER	MILK	SODA	DETERGENT
OJ	4	1	1	2	1
Window Cleaner	1	2	1	1	0
Milk	1	1	1	0	0
Soda	2	1	0	3	1
Detergent	1	0	0	1	2

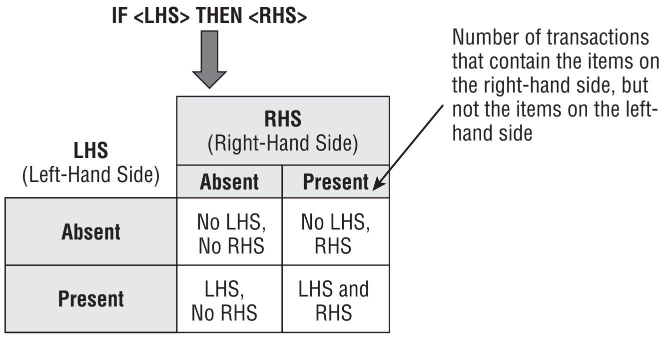


Figure 15-7: An association rule has a corresponding contingency table, where the two dimensions are based on the two sides of the rule. The cells in the table contain counts of the number of transactions that appear or do not appear on either side.

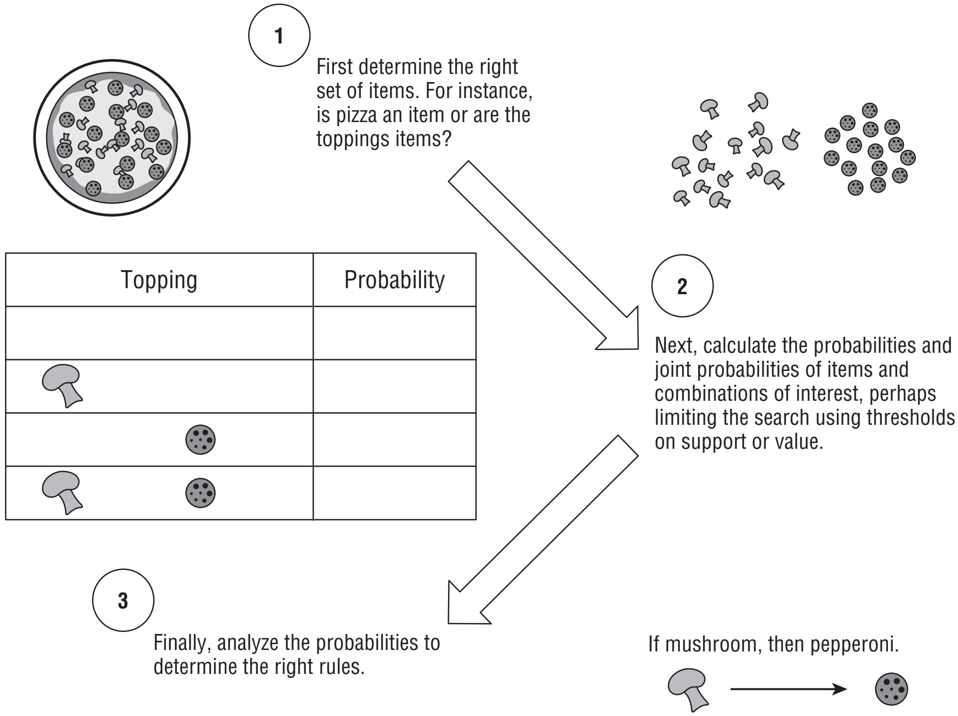


Figure 15-8: Finding association rules has these basic steps.

Table 15-3: Transactions with More Summarized Items

CUSTOMER	PIZZA	MILK	SUGAR	APPLES	COFFEE
1	X				
2		X	X		
3	X			X	X
4		X			X
5	X		X	X	X

Table 15-4: Transactions with More Detailed Items

CUSTOMER	EXTRA CHEESE	ONIONS	PEPPERS	MUSHROOMS	OLIVES
1	X	X			X
2			X		
3	X	X		X	
4		X			X
5	X		X	X	X

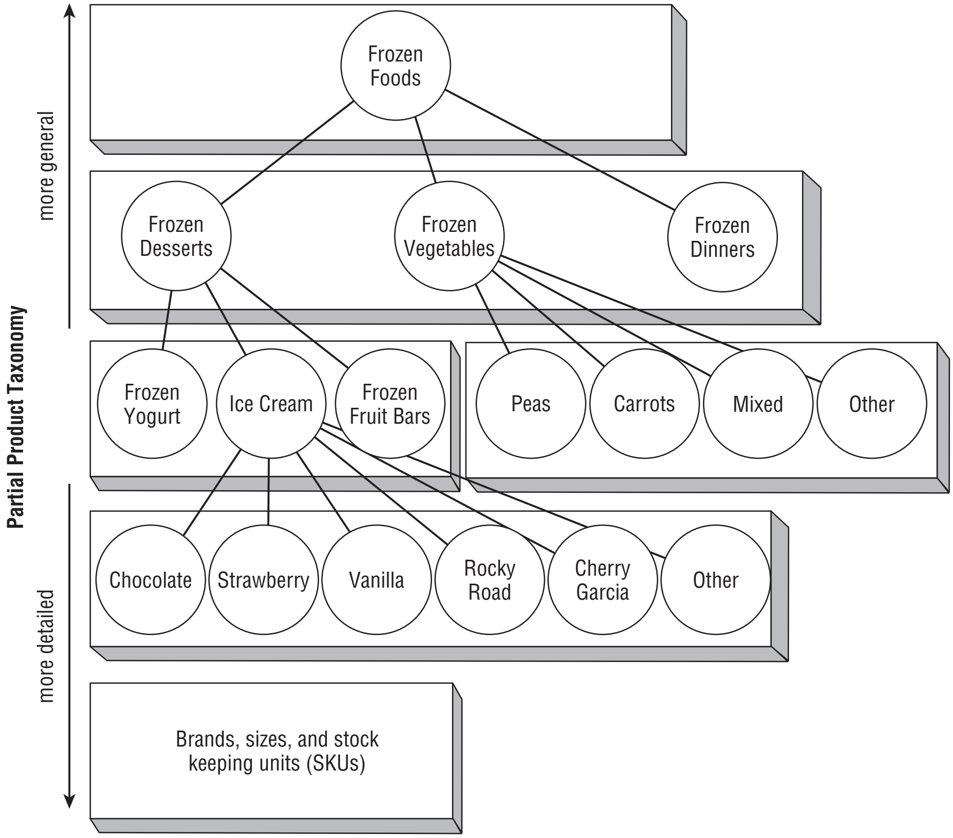


Figure 15-9: Product hierarchies start with the most general and move to increasing detail.

Table 15-5: Probabilities of Three Items and Their Combinations

COMBINATION	PROBABILITY
A	45.0%
B	42.5%
C	40.0%
A and B	25.0%
A and C	20.0%
B and C	15.0%
A and B and C	5.0%

Table 15-6: Confidence in Rules

RULE	P(CONDITION)	P(CONDITION AND RESULT)	CONFIDENCE
If A and B, then C	25%	5%	20%
If A and C, then B	20%	5%	25%
If B and C, then A	15%	5%	33%

$$lift = \frac{\frac{p(condition\ and\ result)}{p(condition)}}{p(result)} = \frac{p(condition\ and\ result)}{p(condition)\ p(result)}$$

Equation 28

Table 15-7: Lift Measurements for Four Rules

RULE	SUPPORT	CONFIDENCE	P(RESULT)	LIFT
If A and B, then C	5%	20%	40.0%	0.50
If A and C, then B	5%	25%	42.5%	0.59
If B and C, then A	5%	33%	45.0%	0.74
If A, then B	25%	59%	42.5%	1.31

Table 15-8: Transaction Counts for Data in Table 15-5

GROUPING	COUNT	PROPORTION
A only	100	5%
B only	150	8%
C only	200	10%
AB only	400	20%
AC only	300	15%
BC only	200	10%
ABC only	100	5%
None	550	28%

Table 15-9: Chi-Square Calculation for the Rule, "If A and B, then C"

	COUNTS		EXPECTED VALUES		CHI-SQUARE	
	NOT C	C	NOT C	C	NOT C	C
NOT AB	800	700	900	600	11.1	16.7
AB	400	100	300	200	33.3	50.0

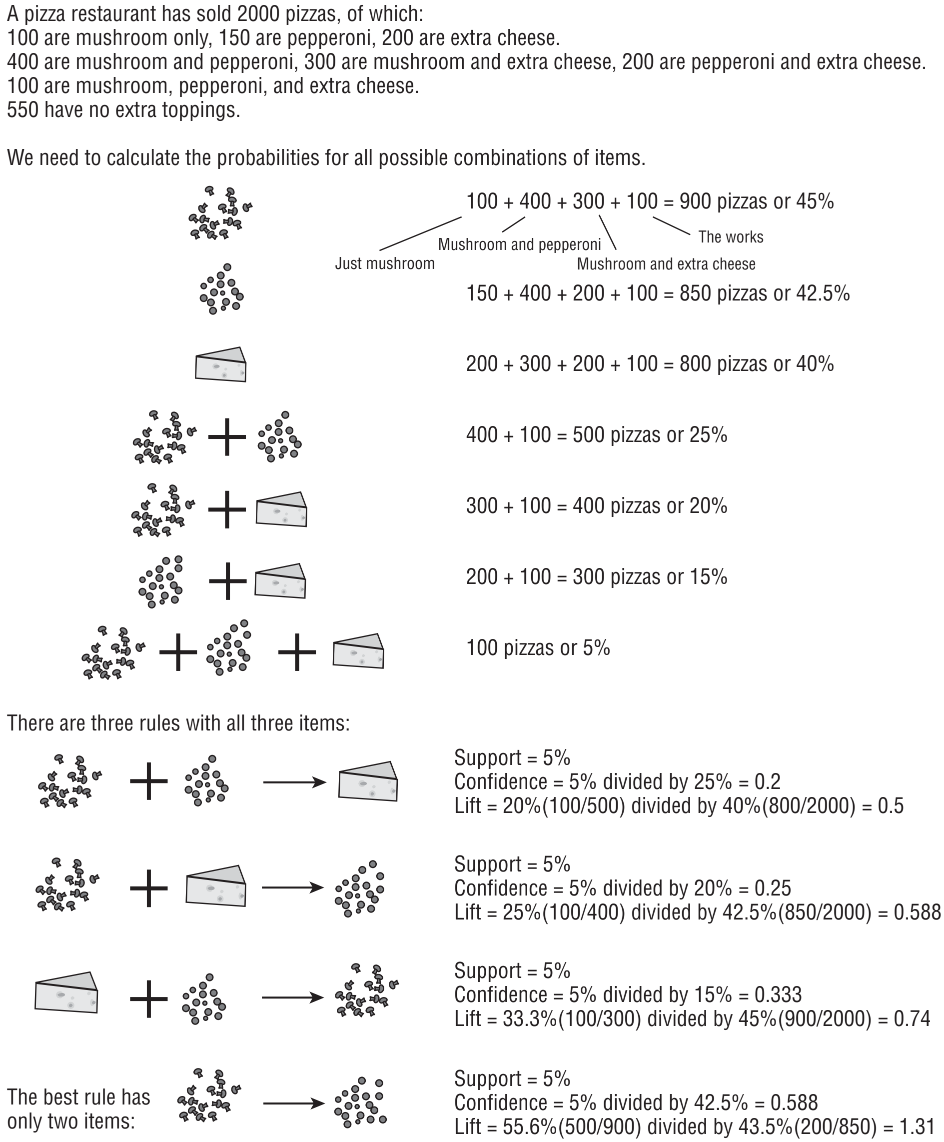


Figure 15-10: This example shows how to count up the frequencies on pizza sales for market basket analysis.

Clicks Imply Complaint Rules	Chi-Square
Telecom + Travel ==> Loans	299.0
Telecom + Government Grants ==> Credit Report	299.0
Government Grants + Gifts ==> Credit Report	299.0
Education + College/Scholarship ==> [Uncategorized]	149.0
Debt + Telecom ==> Credit Report	149.0
Debt + Government Grants ==> Credit Report	149.0
Debt + Gifts ==> Credit Report	149.0
Credit Card + Travel ==> Loans	99.0
Credit Card + Government Grants ==> Credit Report	99.0
Entrepreneurial + Credit Report ==> Home Improvement	74.0

Figure 15-11: Some combinations of clicks on e-mail offer types are more likely to lead to complaints on subsequent offers.

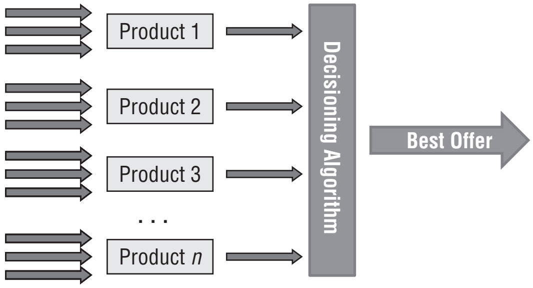


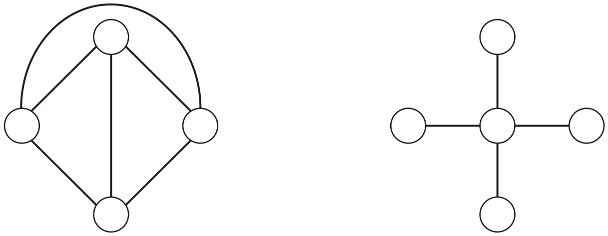
Figure 15-12: A typical cross-sell model builds propensities for each product and then has a decisioning algorithm to choose the best product for each customer.

Table 15-10: Prescription Sequences for One Calendar Year

PURCHASE PATTERN	PRESCRIPTIONS	PATIENTS	PERCENT
LLLL	4	12,099	12.2%
LLLLLLLLLLLL	12	11,910	12.0%
L	1	11,522	11.6%
LLL	3	9,261	9.3%
LLLLLLLLLLL	11	9,042	9.1%
LL	2	8,653	8.7%
LLLLL	5	6,328	6.4%
LLLLLLLLLL	10	6,325	6.4%
LLLLLL	6	6,013	6.1%
LLLLLLLLL	9	5,316	5.4%
LLLLLLL	7	5,147	5.2%
LLLLLLLL	8	4,992	5.0%
OTHER		2,701	2.7%

Table 15-11: Patients with 11 Zocor Prescriptions

SEQUENCE	LENGTH	PATIENTS	PERCENT
ZZZZZZZZZZZZ	12	8674	44.8%
ZZZZZZZZZZZ	11	7699	39.8%
ZZZZZZZZZZZZZ	13	2063	10.7%
ZZZZZZZZZZZZZZ	14	390	2.0%
ZZZZZZZZZZZV	12	180	0.9%
ZZZZZZZZZZZZZZZ	15	152	0.8%
ZZZZZZZZZZZZZZZZZZ	18	112	0.6%
ZZZZZZZZZZZZZZZZ	16	32	0.2%
ZZZZZZZZZZZCZZ	14	13	0.1%
ZZZZZZZZZZZVV	13	11	0.1%
ZZZZZZZZZZZZC	13	11	0.1%
ZZZZZZZZZZZZLL	14	11	0.1%
ZZZZZZZZZZZZZZZZZZZ	19	11	0.1%
ZZZZZZZZZZZZZZZZZZZZZZ	22	10	0.1%
ZZZZZZZZZZZZZZZZZZZZZZZZZ	25	10	0.1%
ZZZZZZZZZZZZZZZZZ	17	9	0.0%
ZZZZZZZZZZZZZZZZZZZZZZZ	23	9	0.0%
ZZZZZZZZZZZZZZZZZZZZL	21	8	0.0%
ZZZZZZZZZZZZV	13	7	0.0%
ZZZZZZZZZZZZZZVVZZ	18	7	0.0%
ZZZZZZZZZZZMM	13	6	0.0%
ZZZZZZZZZZZZZZZZLZ	18	4	0.0%
ZZZZZZZZZZZZZZZZZZO	19	4	0.0%



A fully connected graph with four nodes and six edges. In a fully connected graph, there is an edge between every pair of nodes.

A graph with five nodes and four edges.

Figure 16-1: The graph on the left is fully connected. The graph on the right has a hub and spokes.

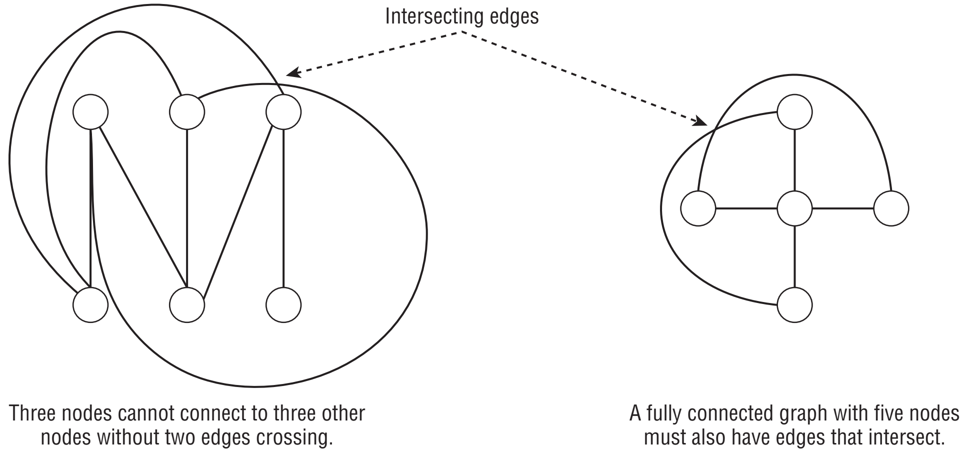


Figure 16-2: Some graphs cannot be drawn without crossing edges.

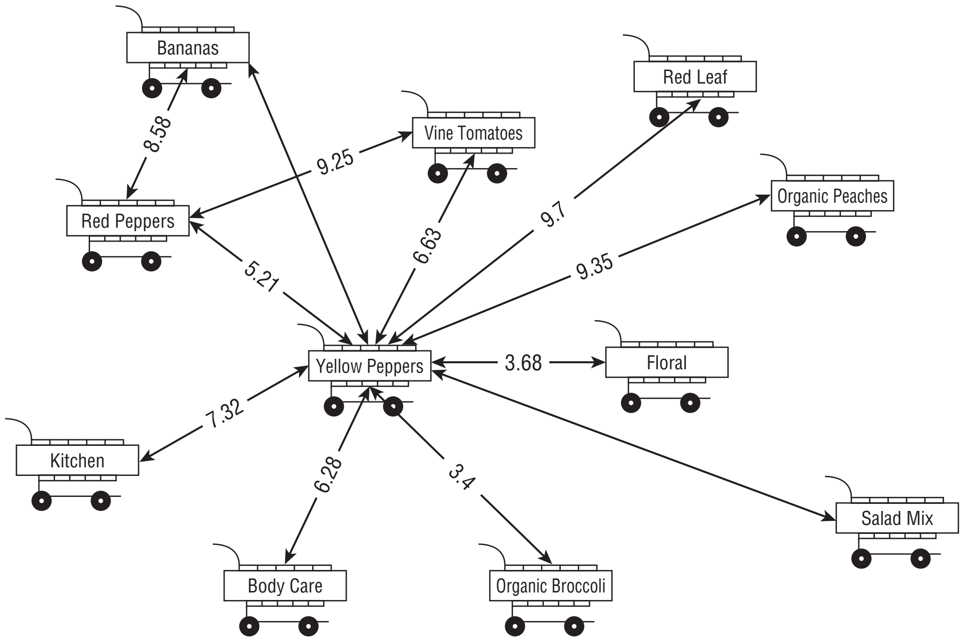


Figure 16-3: This is an example of a weighted graph where the edge weights are the number of transactions containing the items represented by the nodes at either end.

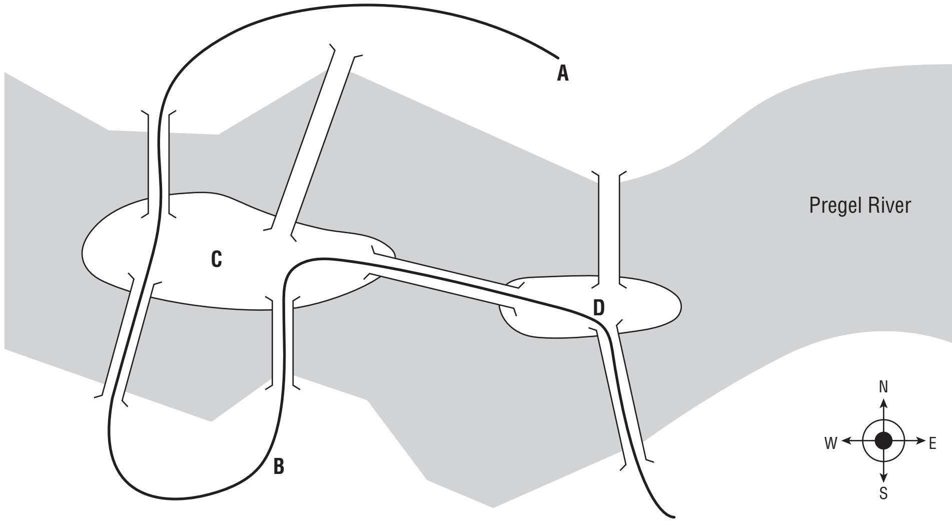


Figure 16-4: The Pregel River in Königsberg has two islands connected by a total of seven bridges, which played an important role in the development of graph theory.

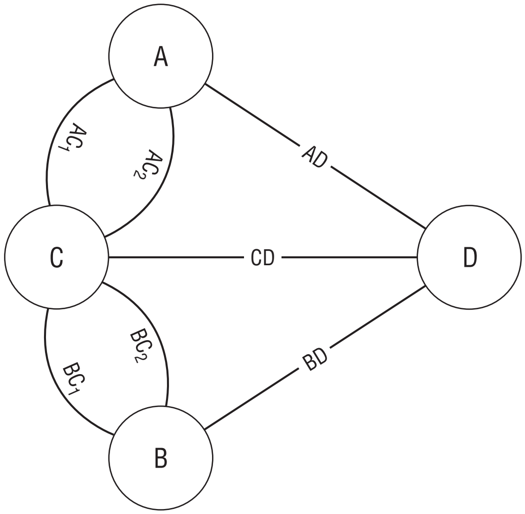


Figure 16-5: This graph represents the layout of Königsberg. The edges are bridges and the nodes are the riverbanks and islands.

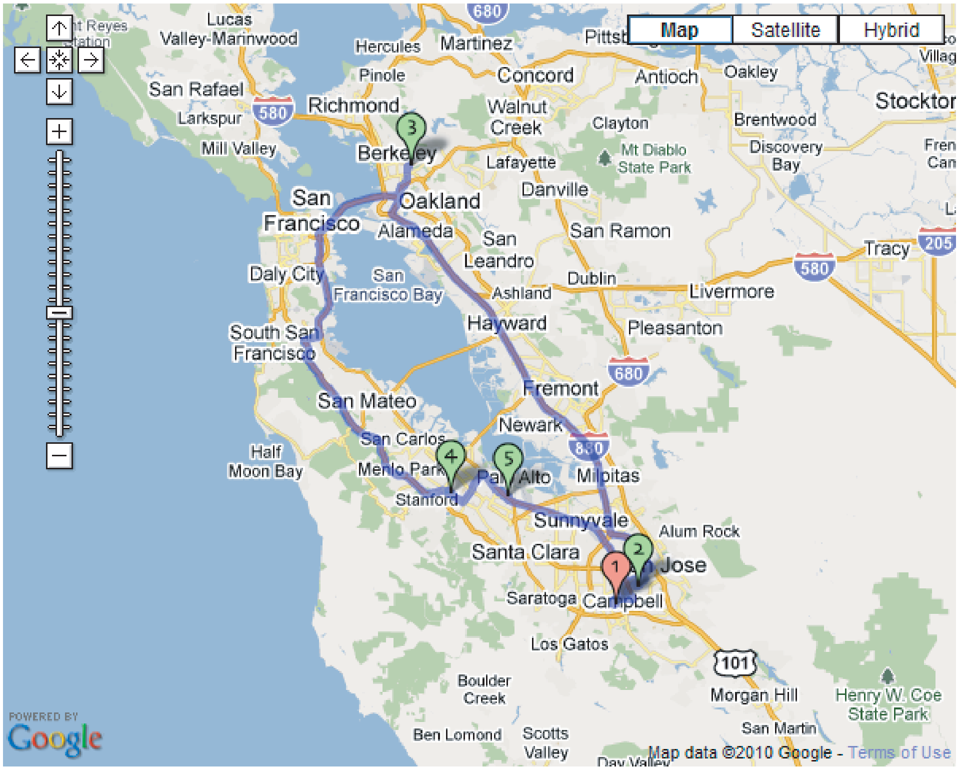


Figure 16-6: This route map, produced by Optimap using data provided by the Google Maps API, shows the best route (measured by driving time) for visiting several cities in the San Francisco Bay Area.

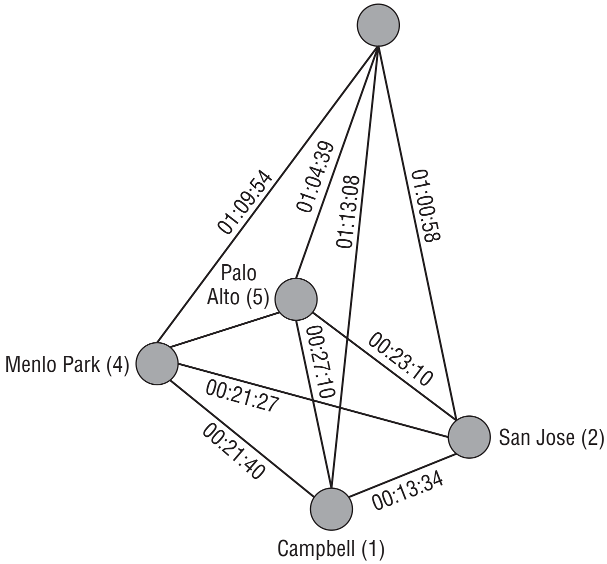
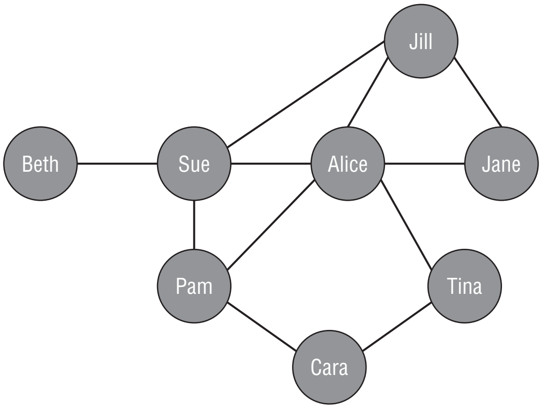


Figure 16-7: This weighted graph shows the expected driving time in hh:mm:ss between selected city pairs.

Table 16-1: Driving Times Between Addresses in Selected City Pairs

FROM/TO	CAMPBELL (1)	SAN JOSE (2)	BERKELEY (3)	MENLO PARK (4)	PALO ALTO (5)
CAMPBELL (1)	0	814 00:13:34	4,388 01:13:08	1,300 00:21:40	1,630 00:27:10
SAN JOSE (2)	814 00:13:34	0	3,658 01:00:58	1,287 00:21:27	1,390 00:23:10
BERKELEY (3)	4,388 01:13:08	3,658 01:00:58	0	4,194 01:09:54	3,879 01:04:39
MENLO PARK (4)	1,300 00:21:40	1,287 00:21:27	4,194 01:09:54	0	1,037 00:17:17
PALO ALTO (5)	1,630 00:27:10	1,390 00:23:10	3,879 01:04:39	1,037 00:17:17	0



Only Alice has five friends, but because of her, five people have a friend with five friends.

Conversation Paradox

Table 16-2: Five Telephone Calls

ID	ORIGINATING NUMBER	TERMINATING NUMBER	DURATION
1	353-3658	350-5166	00:00:41
2	353-3068	350-5166	00:00:23
3	353-4271	353-3068	00:00:01
4	353-3108	555-1212	00:00:42
5	353-3108	350-6595	00:01:22

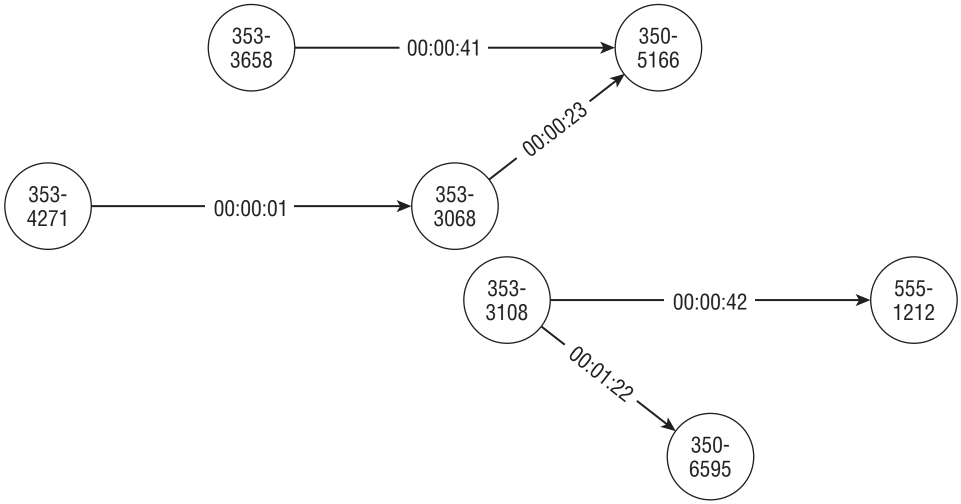


Figure 16-8: Five calls link seven telephone numbers.

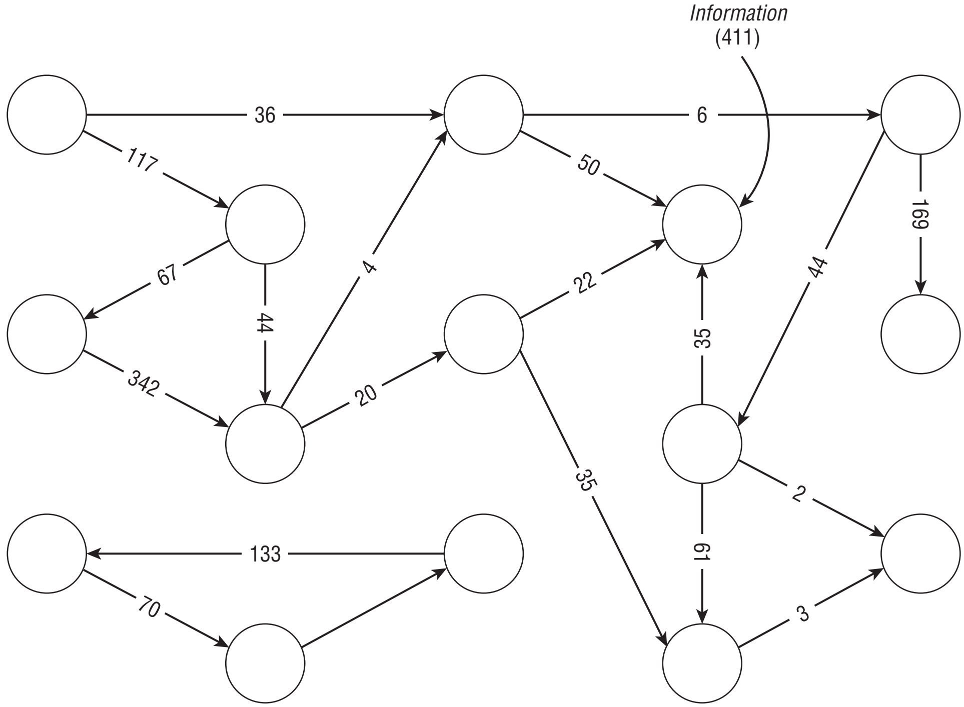
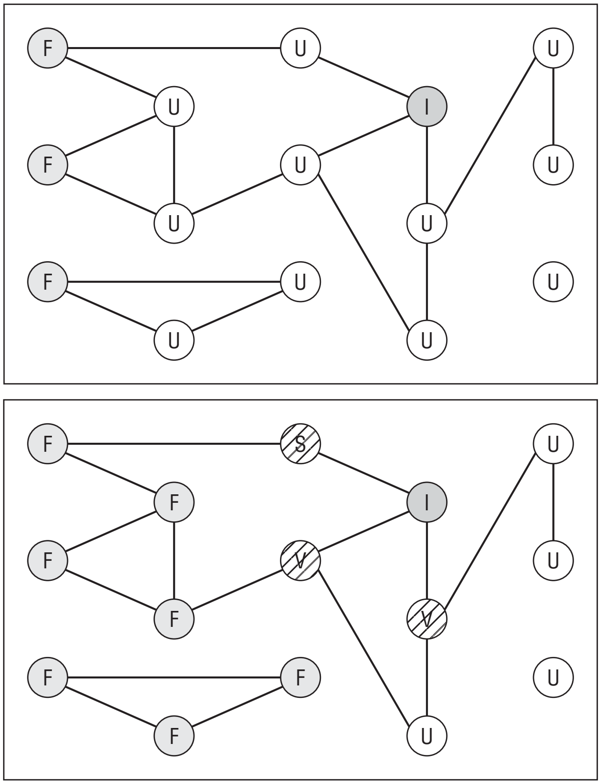


Figure 16-9: A call graph for 15 numbers and 19 calls.



This is the initial call graph with short calls removed and with nodes labeled as “fax,” “unknown,” and “information.”

Nodes connected to the initial fax machines are assigned the “fax” label.

Those connected to “information” are assigned the “voice” label.

Those connected to both are “shared.”

The rest are “unknown.”

Figure 16-10: Applying the graph-coloring algorithm to the call graph shows which numbers are fax numbers and which are shared.

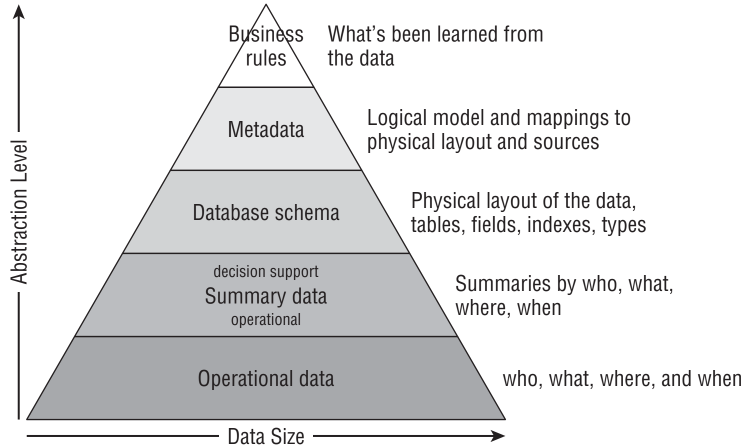


Figure 17-1: A hierarchy of data and its descriptions helps users navigate around a data warehouse. As data gets more abstract, it generally gets less voluminous.

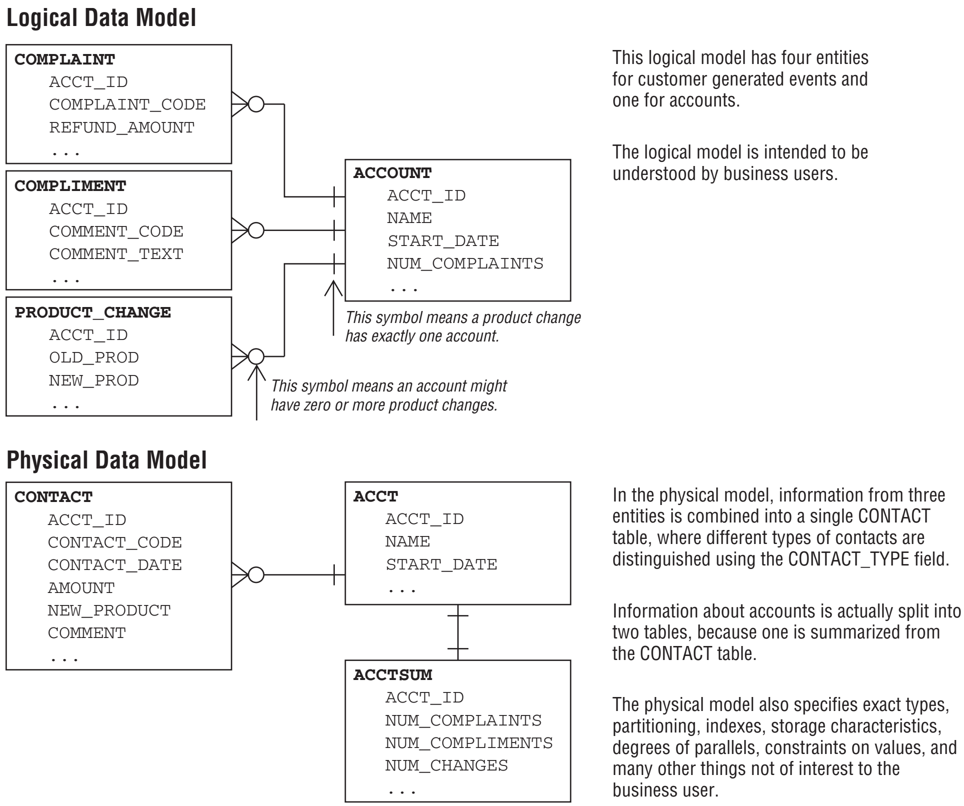
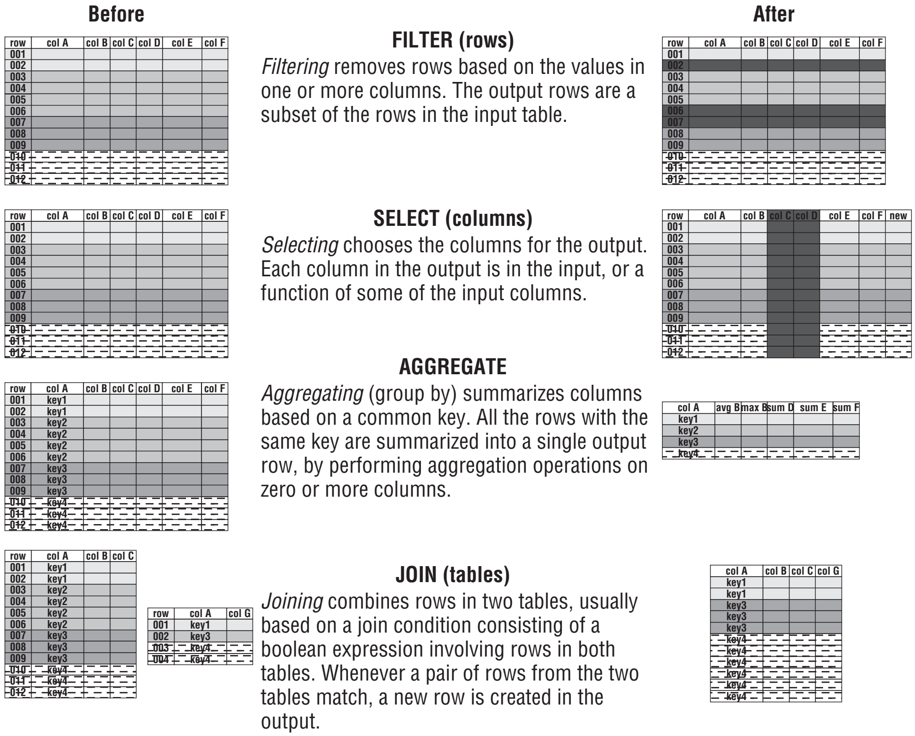
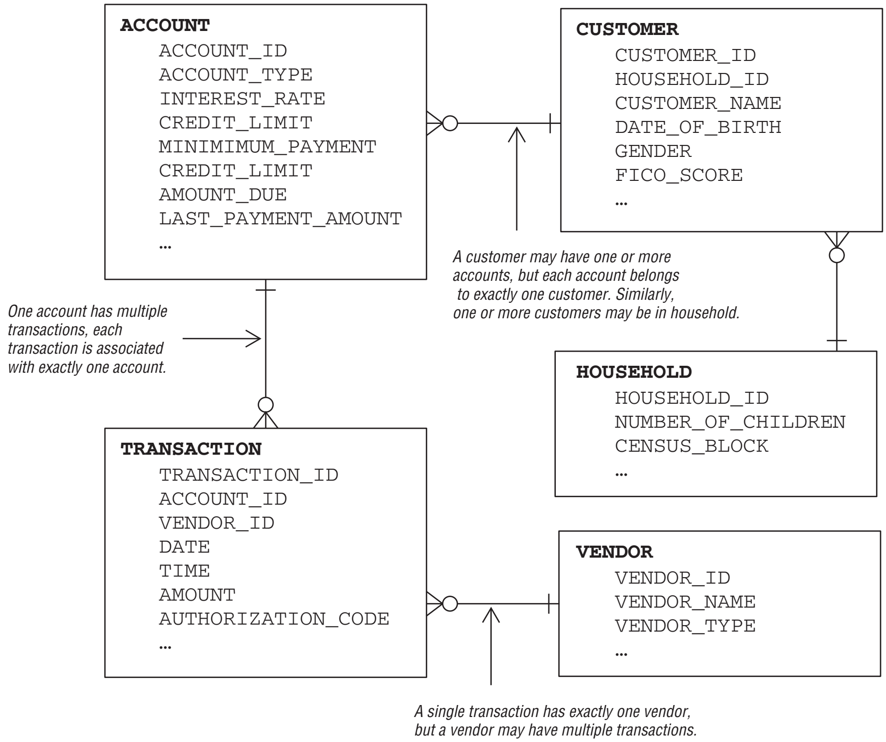


Figure 17-2: The physical and logical data models may not be similar to each other.



Relational databases have four major querying operations.

Relational Databases



An ER diagram can be used to show the tables and fields in a relational database. Each box shows a single table and its columns. The lines between the boxes show relationships, such as 1-many, 1-1, and many-to-many. Because each table corresponds to an entity, this is called a physical model.

Sometimes, the physical model of a database is very complicated. For instance, the TRANSACTION table might actually be split into a separate table for each month of transactions, to facilitate backup and restore processes.

An entity relationship diagram describes the layout of data for a simple credit card database.

Database Structure

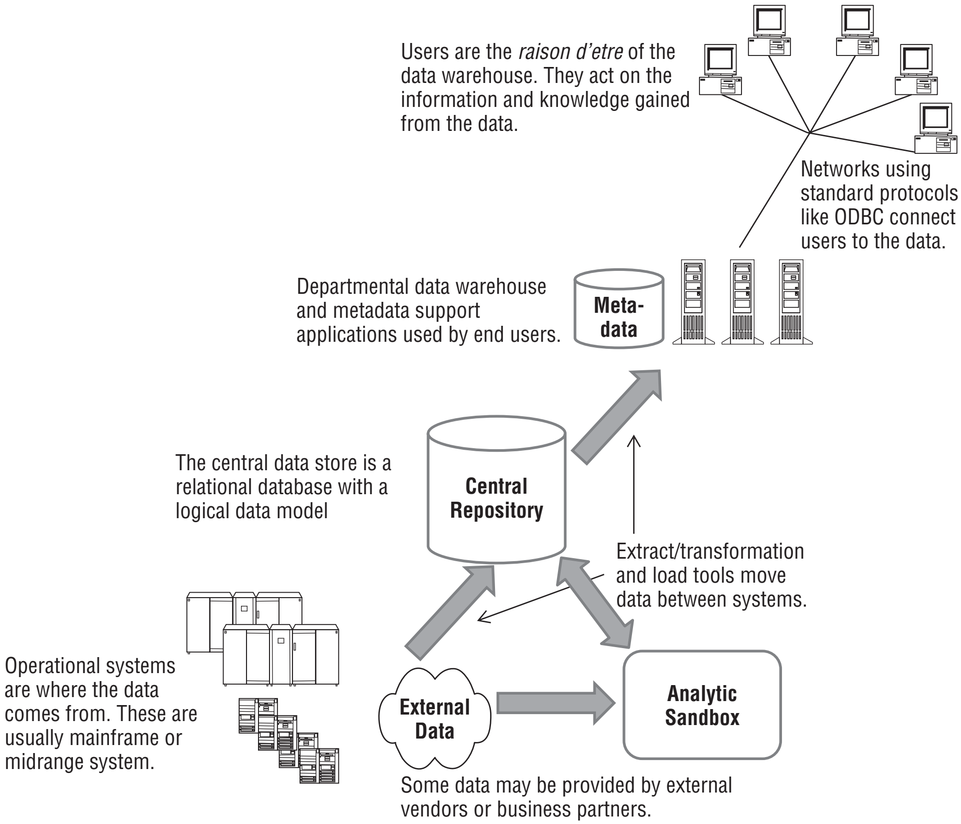
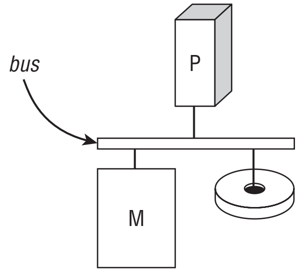
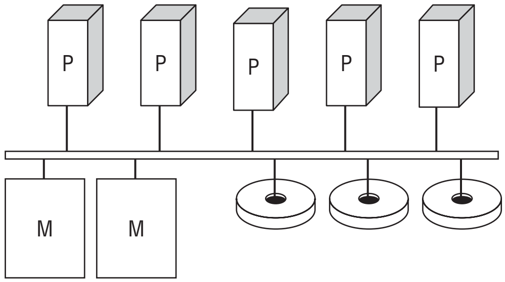


Figure 17-3: The multitiered approach to data warehousing includes a central repository, data marts, analytic sandboxes, end-user tools, and tools that connect all these pieces together.



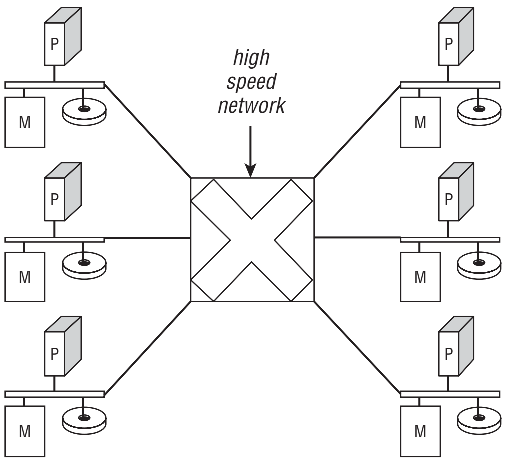
Uniprocessor

A simple computer follows the architecture laid out by Von Neumann. A processing unit communicates to memory and disk over a local bus. (Memory stores both data and the executable program.) The speed of the processor, bus, and memory limits performance and scalability.



SMP

The symmetric multiprocessor (SMP) has a shared-everything architecture. It expands the capabilities of the bus to support multiple processors, more memory, and a larger disk. The capacity of the bus limits performance and scalability. SMP architecture usually max out with fewer than 20 processing units.



MPP

The massively parallel processor (MMP) has a shared-nothing architecture. It introduces a high-speed network (also called a switch that connects independent processor/memory/disk components. MPP architectures are very scalable but fewer software packages can take advantage of all the hardware.

Parallel computers build on the basic Von Neumann uniprocessor architecture. SMP and MPP systems are scalable because more processing units, disk drives, and memory can be added to the system.

Processors

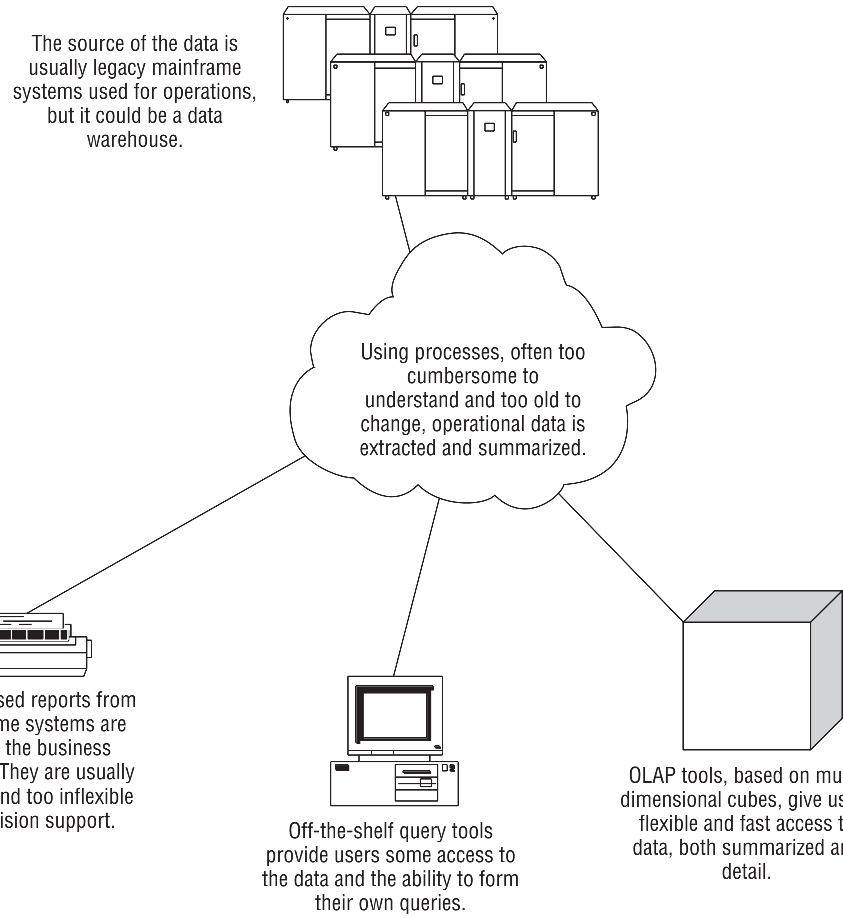


Figure 17-4: Reporting requirements on operational systems are typically handled the same way they have been for decades. Is this the best way?

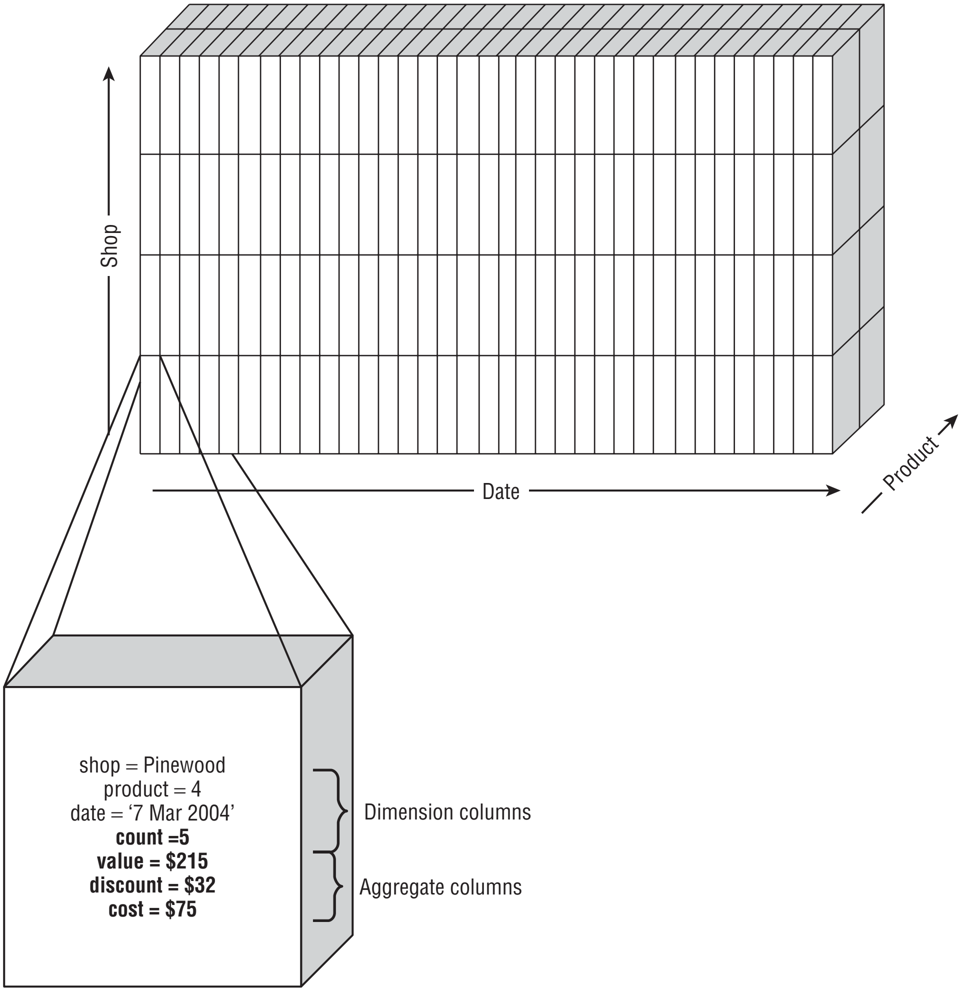


Figure 17-5: The cube used for OLAP is divided into subcubes. Each subcube contains the key for that subcube and summary information for the data falls into that subcube.

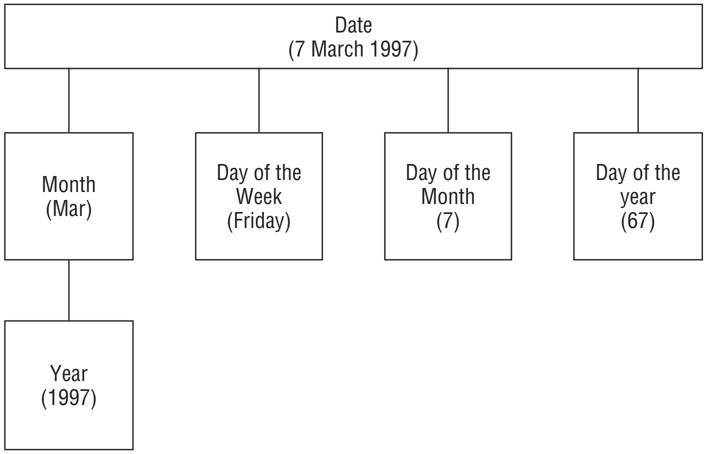


Figure 17-6: Dates have multiple hierarchies.

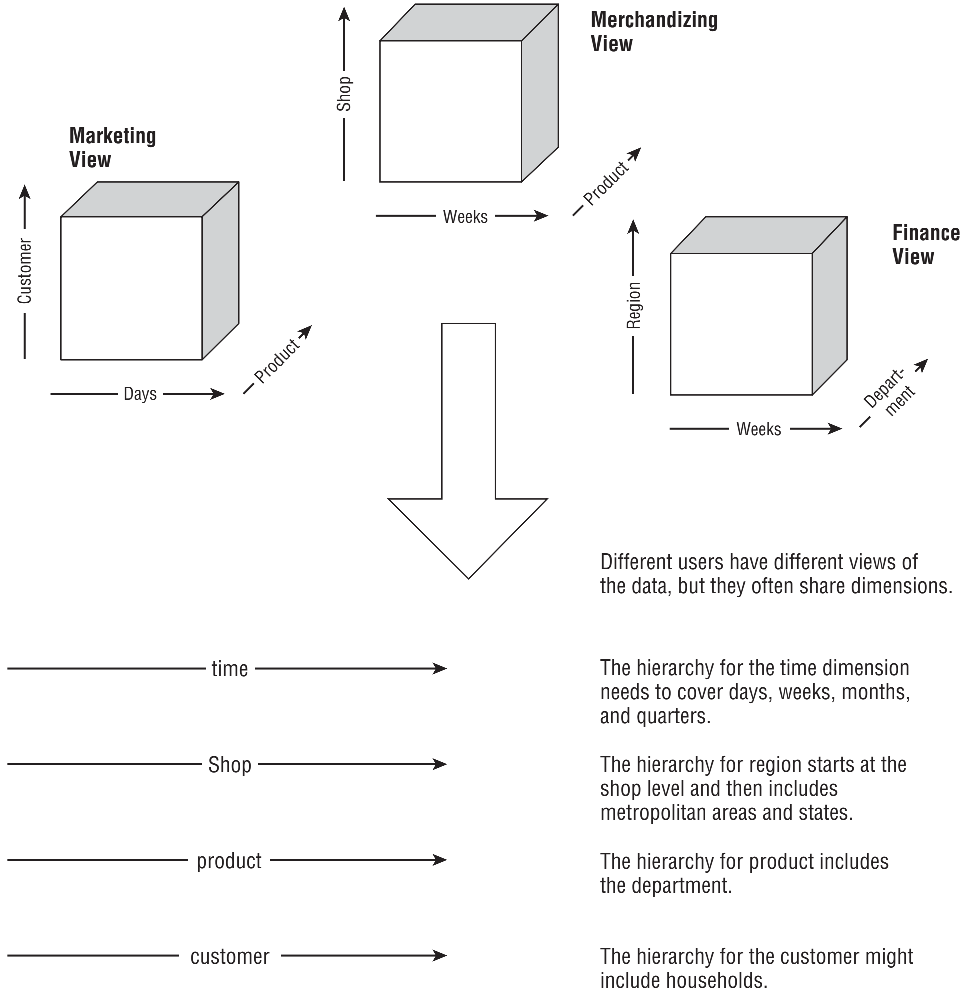


Figure 17-7: Different views of the data often share common dimensions. Finding the common dimensions and their base units is critical to making data warehousing work well across an organization.

Dept	DESCRIPTION
01	CORE FRAGRANCE
02	MISCELLANEOUS
05	GARDENS
06	BRIDAL
10	ACCESSORIES

SKU	Description	Dept
0001	V NECK TEE	70
0002	PANTYHOSE	65
0003	TUXEDO PJ	60
0004	NOVELTY T SHIRT	70
0005	VELOUR JUMPSUIT	76

Color	Description
01	BLACK
02	IVORY
03	TAYLOR GREEN
04	STILETTO
05	BLUE TOPAZ

Shop	SKU	Color	Date	Count	Sales	Cost	Returns
0001	0001	01	000001	5	$50	$20	0
0001	0002	02	000001	12	$240	$96	0
0001	0002	03	000001	4	$80	$32	1
0001	0002	04	000001	12	$240	$96	0
0001	0003	09	000001	19	$85	$19	2
0001	0003	01	000001	5	$25	$5	0
0150	0001	01	000001	31	$310	$134	2

Date	Year	Month	Day
000001	1997	01	01
000002	1997	01	01
000003	1997	01	01
000004	1997	01	01
000005	1997	01	01

Shop	Reg	State	City	Sq Ft
0001	J	CA	San Francisco	3,141
0007	A	MA	Central Boston	1,026
0034	E	FL	Miami	5,009
0124	H	MN	Minneapolis	1,793
0150	B	NY	New York City	6,400

Date	Hol?
000001	Y
000002	N
000003	N
000004	N
000005	N

Date	DoW
000001	Wed
000002	Thu
000003	Fri
000004	Sat
000005	Sun

Reg	Name
A	Northeast
B	New York/NJ
C	Mid Atlantic
D	North Central
E	Southeast

Figure 17-8: A star schema looks more like this. Dimension tables are conceptually nested, with more than one dimension table for a given dimension.

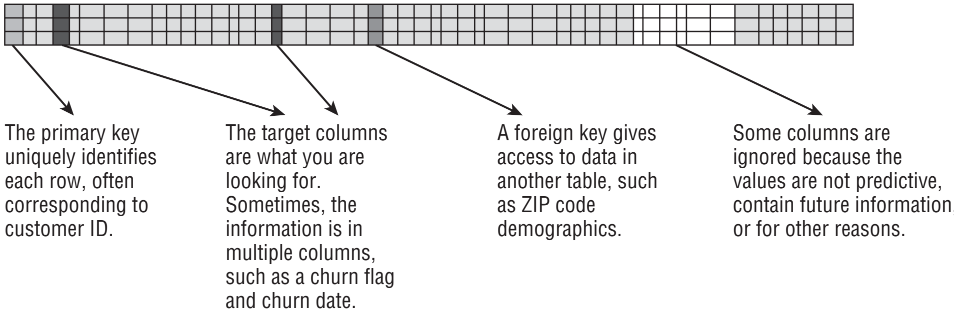


Figure 18-1: The fields of a customer signature have various roles.

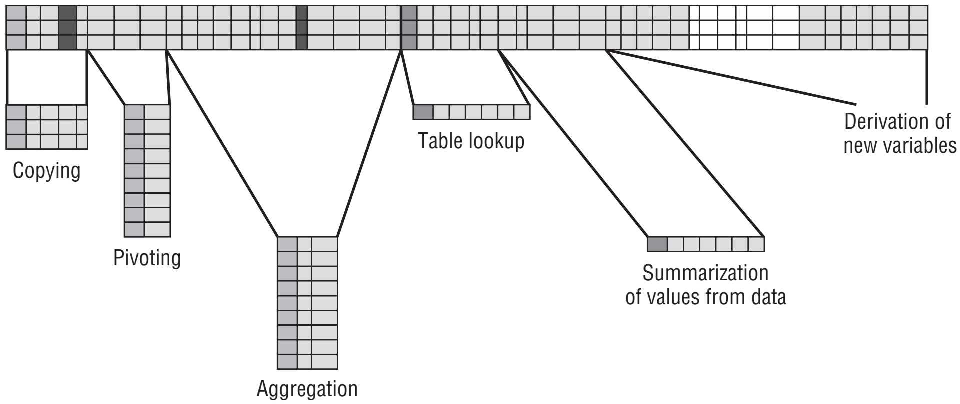


Figure 18-2: Data from most sources must be transformed in various ways before it can be incorporated into the signature.

M13	M12	M11	M10	M09	M08	M07	M06	M05	M04	M03	M02	M01
									$100.10	$943.06	$105.20	$129.12

Customer ID	Date	Amount
271828	2010-08-10	$21.40
314159	2010-08-10	$129.12
	. .	
271828	2010-07-10	$21.40
314159	2010-07-10	$105.20
	. .	
271828	2010-06-10	$21.40
314159	2010-06-10	$943.06
	. .	
271828	2010-05-10	$21.40
314159	2010-05-10	$100.10

Figure 18-3: Vertical data must be pivoted to insert it into the customer signature

Table 18-1: Customer Occupation and Income

OCCUPATION	AGE	INCOME
Database Administrator	50	$92,000
Flight Attendant	32	$42,240
High School Teacher	45	$64,500
Database Administrator	47	–
Letter Carrier	41	$36,500
Bus Driver	58	$24,000
College Professor	41	$73,300
Barista	22	–
Yoga Instructor	28	$15,500

Table 19-1: An Enumeration of the States

STATE	CODE
Alabama	1
Alaska	2
Arizona	3
Arkansas	4
California	5
Colorado	6
Connecticut	7
Delaware	8
Florida	9
Georgia	10
Hawaii	11
Idaho	12
…	…
…	…
…	…

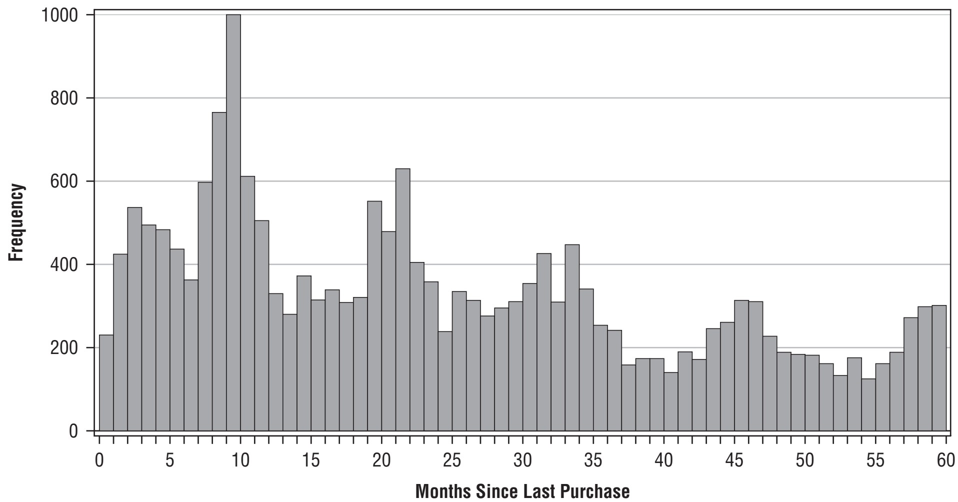


Figure 19-1: Most customers have made a purchase within the last two years.

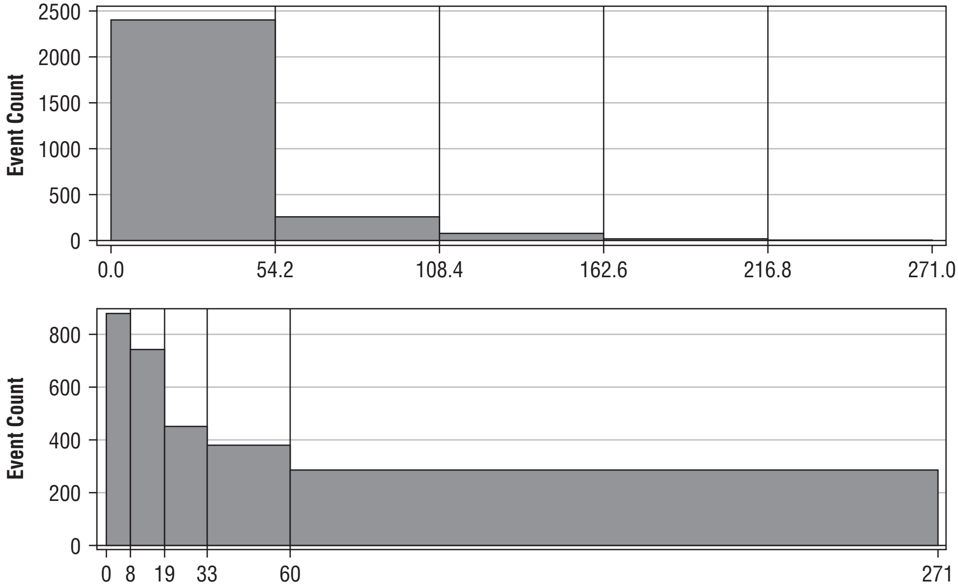


Figure 19-2: Quantiles are generally more useful than equal-width bins.

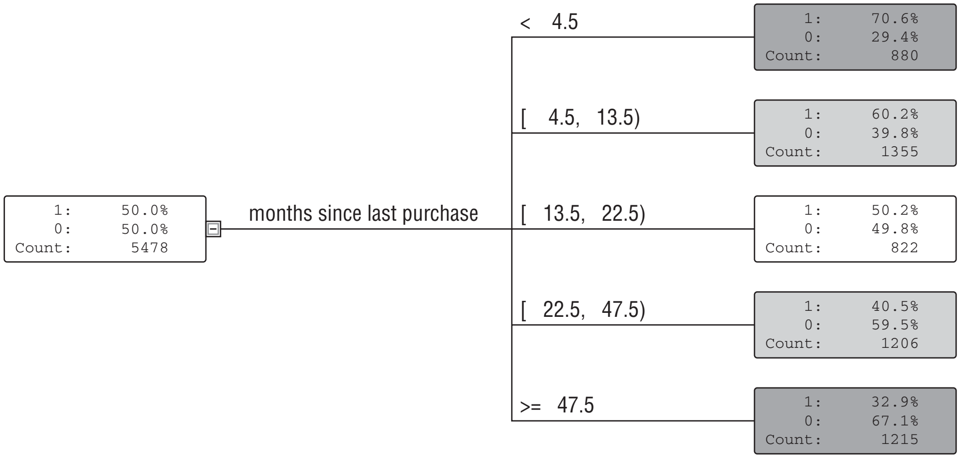


Figure 19-3: A decision tree with a single input variable provides supervised binning.

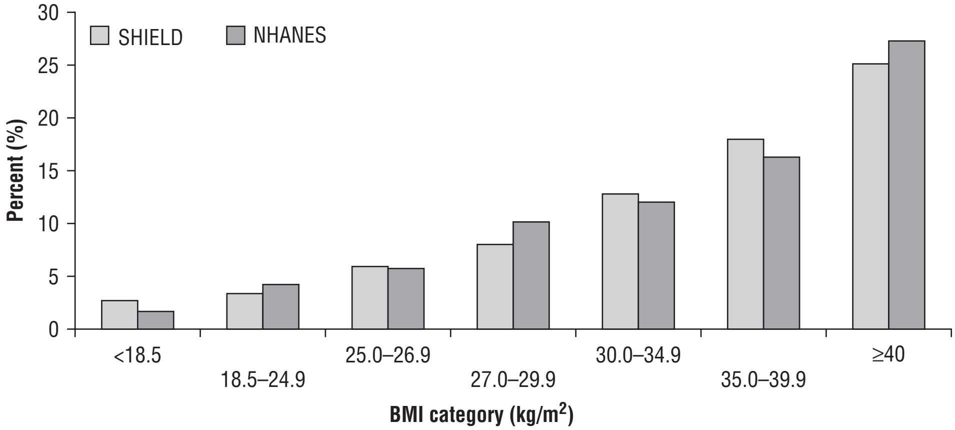


Figure 19-4: Type II diabetes is strongly correlated with body mass index.

$$OBP = \frac{H + BB + HBP}{AB + BB + HBP + SF}$$

Equation 29

$$T_{wc} = 35.74 + 0.6215T_a - 35.75V^{0.16} + 0.4275T_aV^{0.16}$$

Equation 30

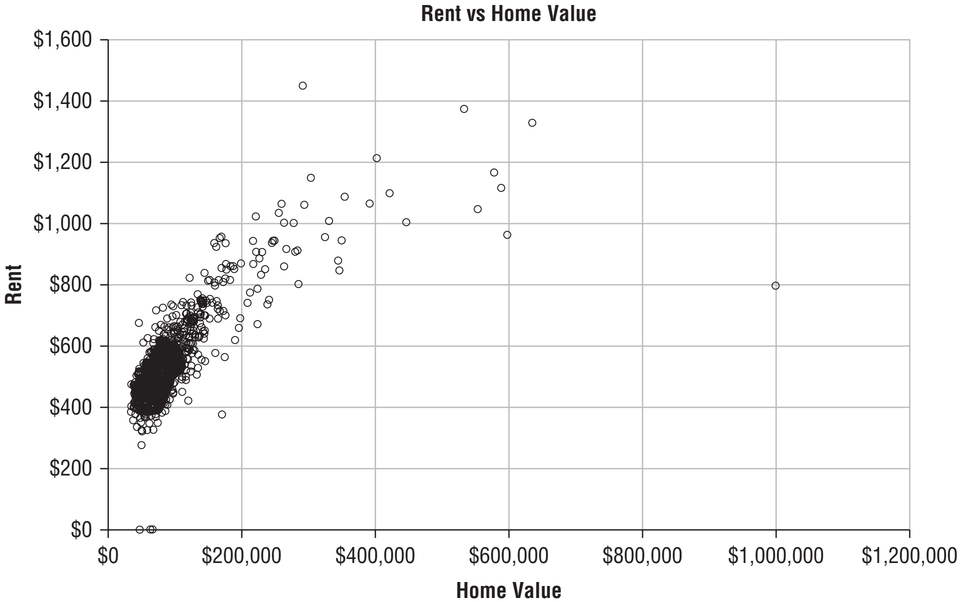


Figure 19-5: As median home value increases, so does median rent.

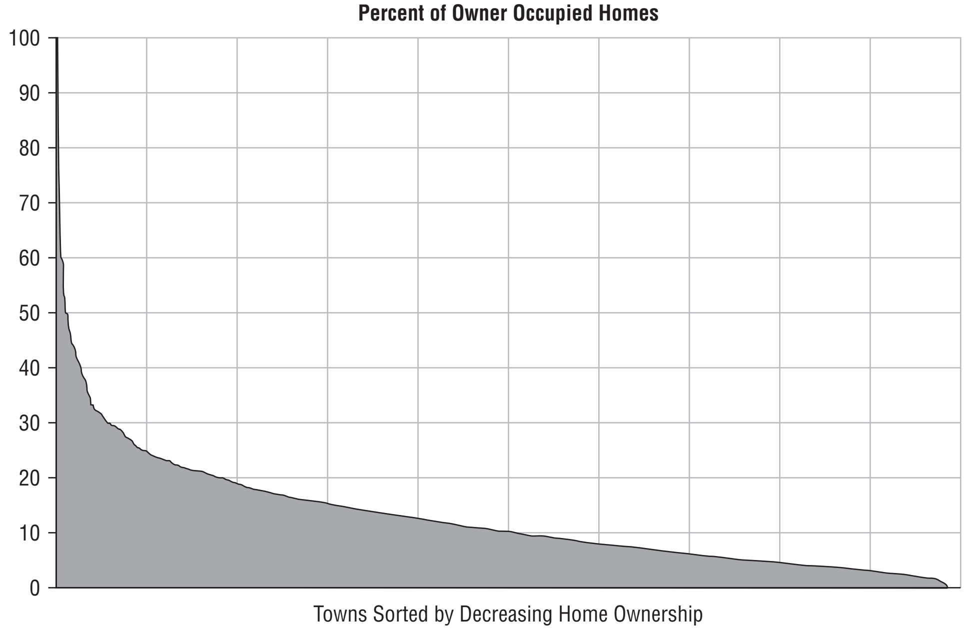


Figure 19-6: Although there are a few towns where no one pays rent, in most towns, many households do pay rent.

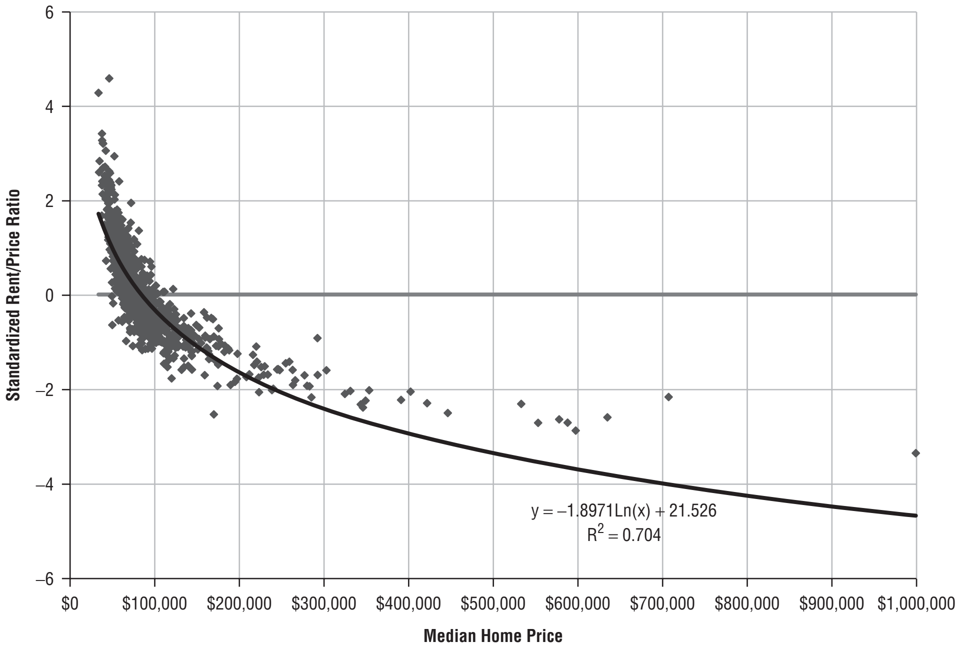


Figure 19-7: Most of the variability in the rent-to-home price ratio is in towns with lower median home prices.

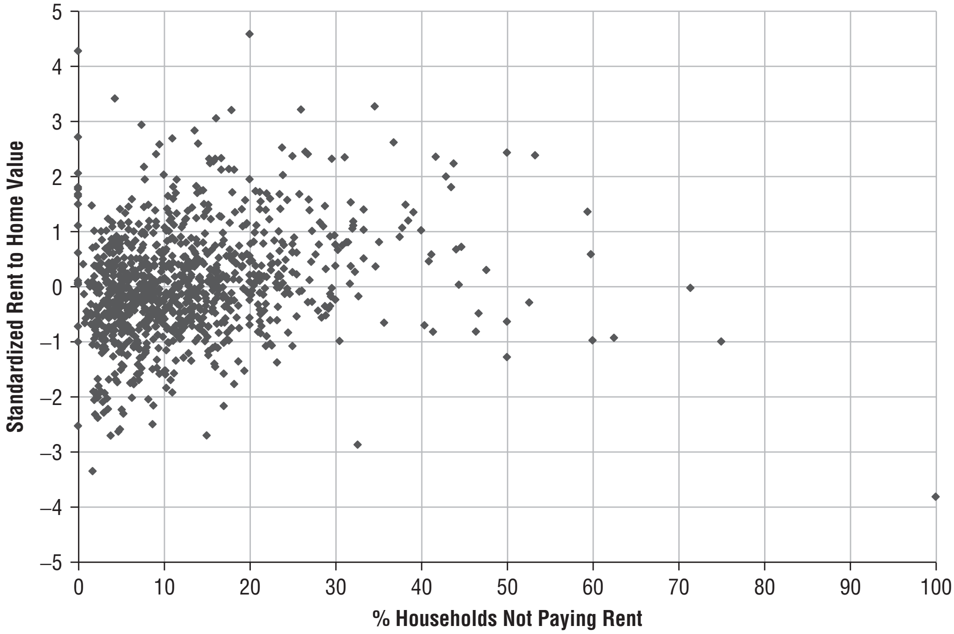


Figure 19-8: No relationship appears to exist between the percentage of owners and the median rent–to–median home value ratio.

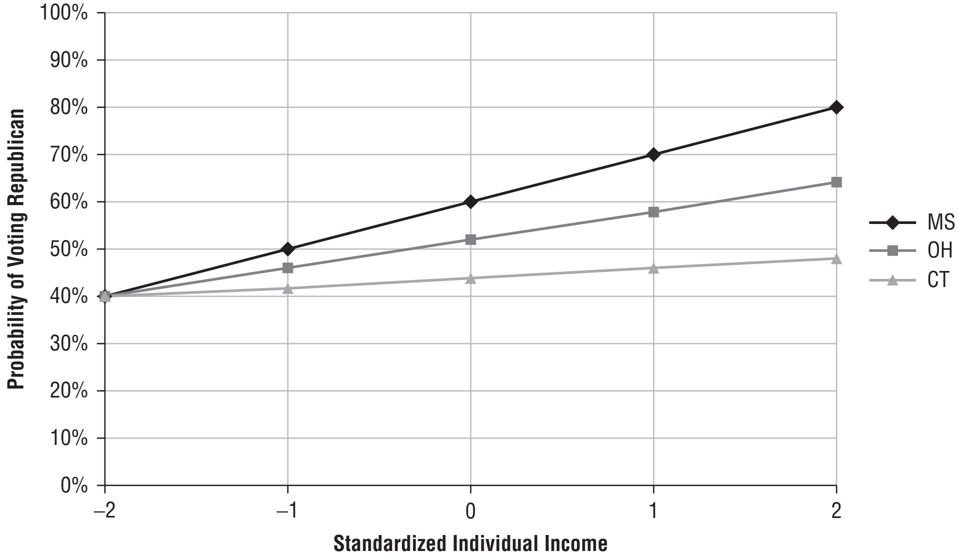


Figure 19-9: Probability of voting Republican as a function of income for several states.

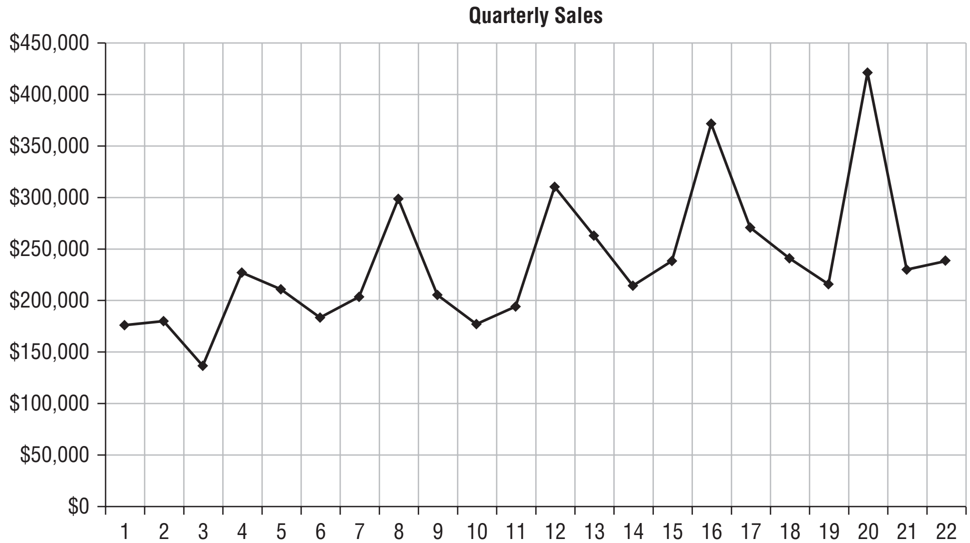


Figure 19-10: Quarterly sales for a small retailer show seasonal fluctuations but an overall increase over time.

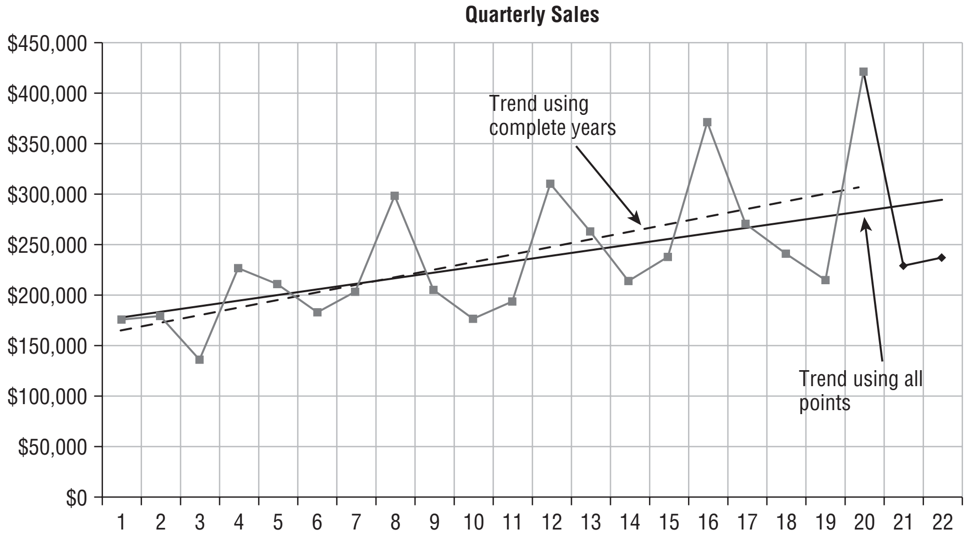


Figure 19-11: The growth trend can be captured by the slope of a best-fit line.

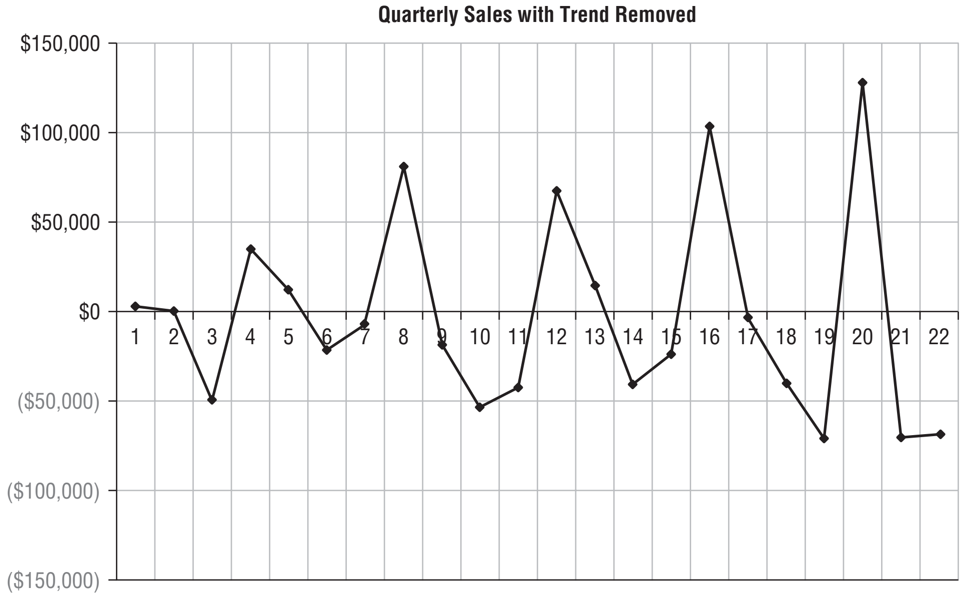


Figure 19-12: After removing trend, capturing the effect of seasonality is easier.

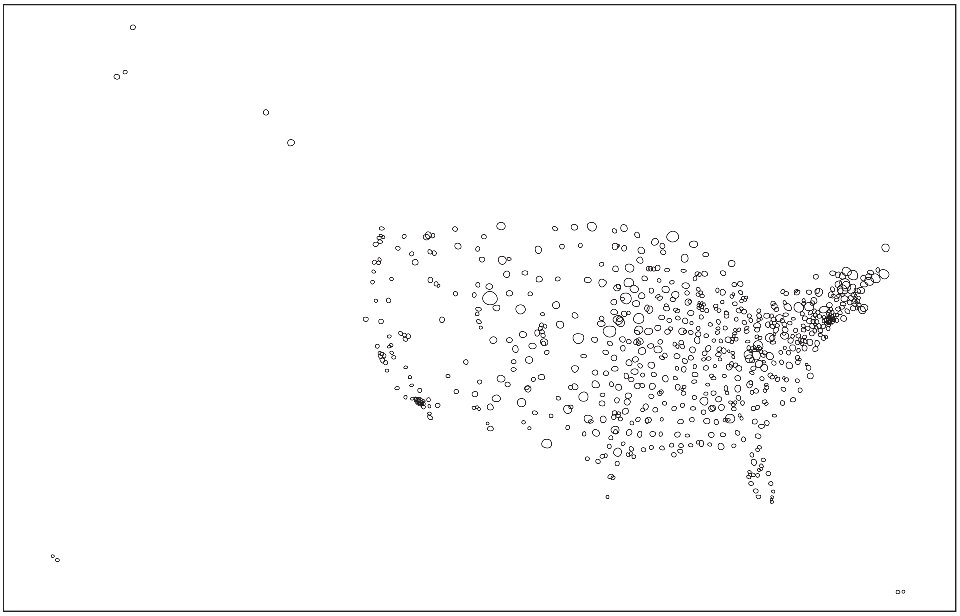


Figure 19-13: Product penetration by ZIP code.

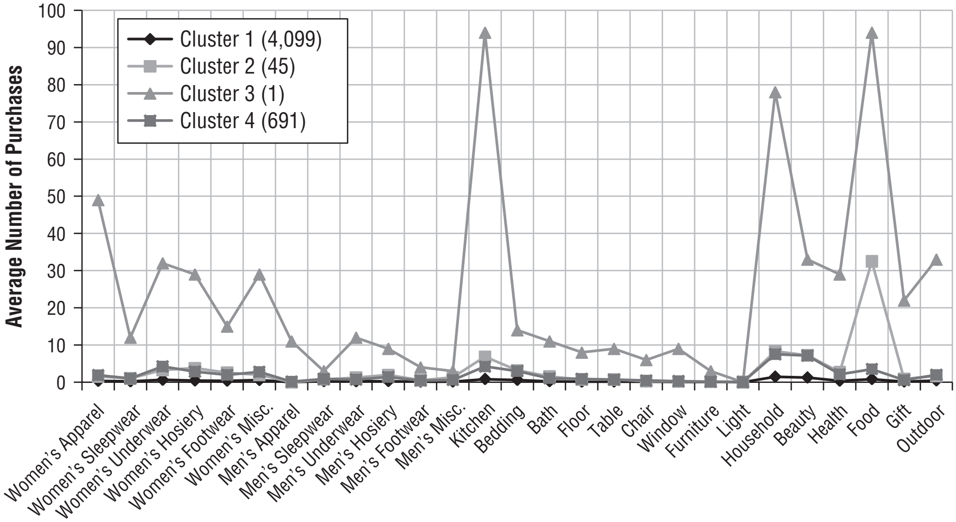


Figure 20-1: Because of sparse data, these four clusters are uninteresting, segmenting the customers into groups based on how much they have purchased. The one customer in Cluster 3 has made many purchases. The many customers in Cluster 1 have probably made only one purchase each.

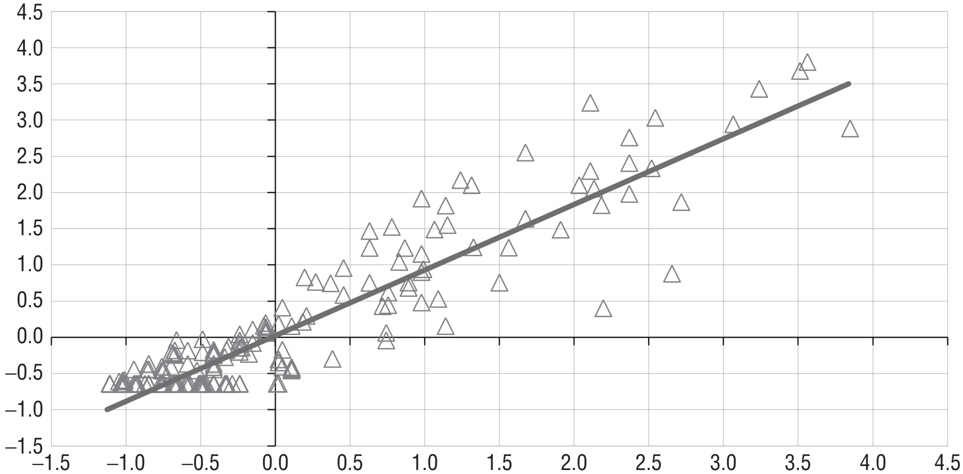
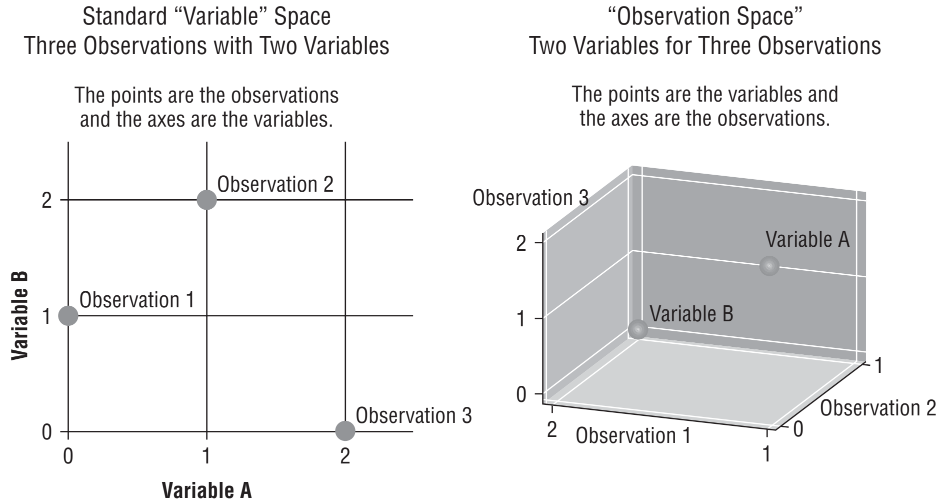


Figure 20-2: This data looks sparse in two dimensions, because of the many areas where there is no data. However, it is not sparse along just the X-axis.



In the variable space, each observation is shown as a point (as shown on the top), with the axes representing variables. In the observation space, each point represents a variable, with the axes representing observations.

Variable vs Observation Space

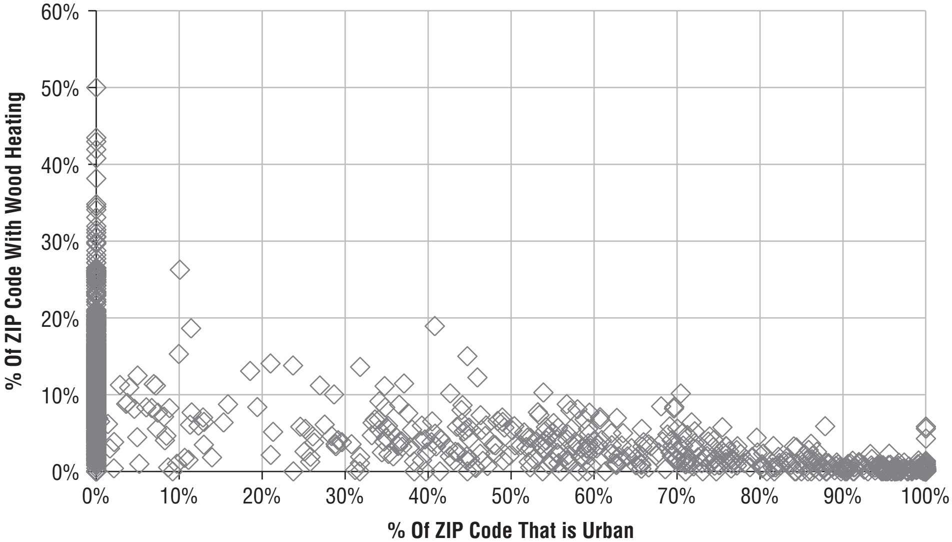


Figure 20-3: The relationship between the proportion of a ZIP code that is urban and the proportion of homes heated primarily by wood shows a partial linear relationship. The relationship between the two factors is quite different depending on whether or not the population is entirely rural.

Table 20-1: Exponential Growth of the Number of Combinations Needed for Exhaustive Selection

NUMBER OF VARIABLES	NUMBER OF COMBINATIONS
2	3
3	7
4	15
5	31
10	1,023
20	1,048,575
30	1,073,741,823
40	1,099,511,627,775
50	1,125,899,906,842,623

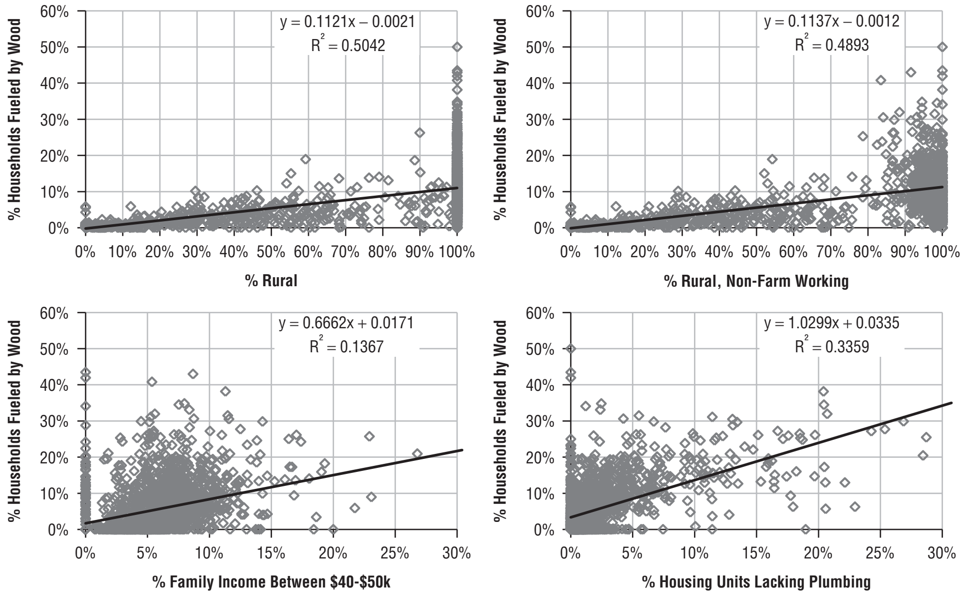


Figure 20-4: The scatter plots in this figure are for four different input variables. The best input variable is the one on the upper left, because it has the largest R^2 value.

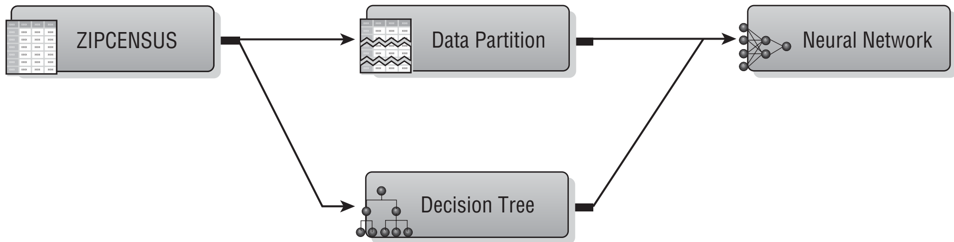


Figure 20-5: This picture shows an SAS Enterprise Miner diagram that uses a decision tree node to select variables for a neural network node. The data "flows" across the top part of the diagram from the source, through the partitioning node, to the neural network. The data also goes to the decision tree node, which builds the tree and passes the variables used to the neural network.

Table 20-2: A Comparison of the Variables Chosen by a Decision Tree and Forward Regression

VARIABLES CHOSEN BY REGRESSION		VARIABLES CHOSEN BY DECISION TREE	
VARIABLE	IMPORTANCE	VARIABLE	IMPORTANCE
hhuoplumbinglacking	1.000	longitude	1.000
pruralnonfarm	0.968	hhuoplumbinglacking	0.992
longitude	0.612	prural	0.641
latitude	0.350	hhuoplumbingcomplete	0.352
hhperson2fnonfamily	0.310	hhumedianyear	0.228
faminc010_015	0.288	latitude	0.196

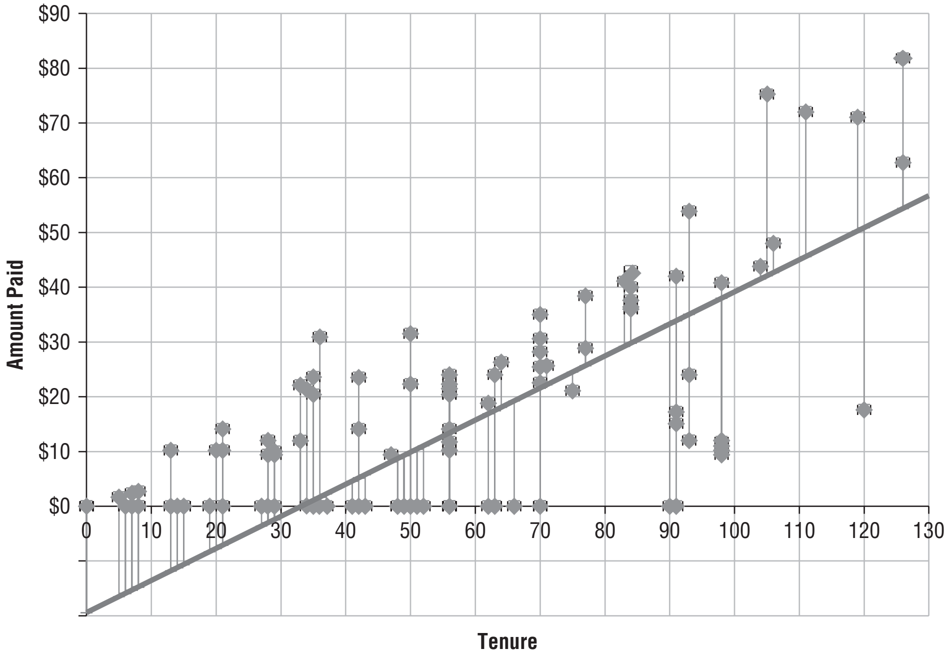


Figure 20-6: The best-fit line minimizes the sum of the squares of the vertical distances from the data points to the line.

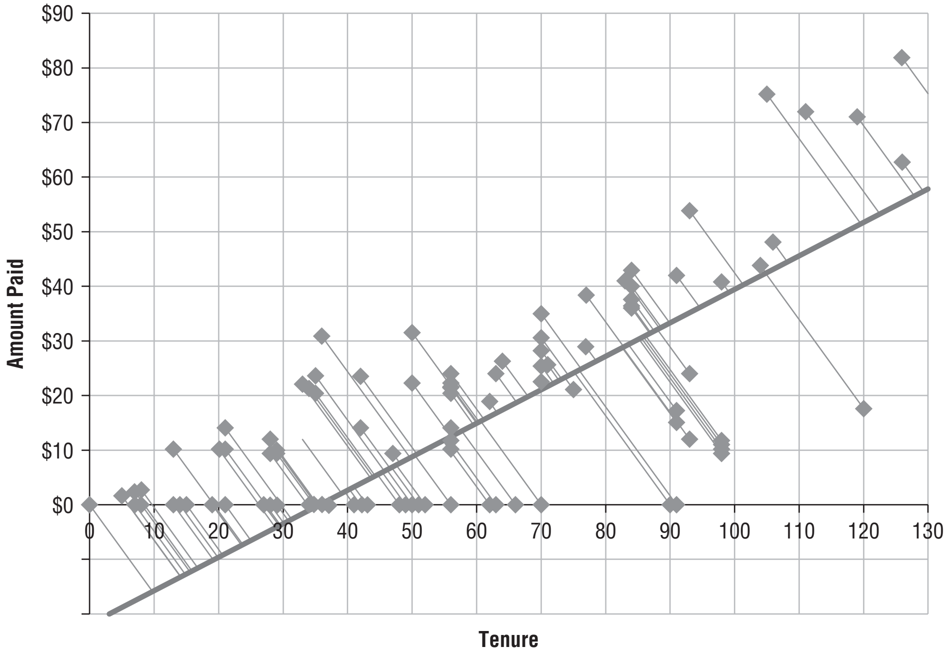


Figure 20-7: The first principal component is the line that minimizes the sum of the squares of the distances from each point to the line.

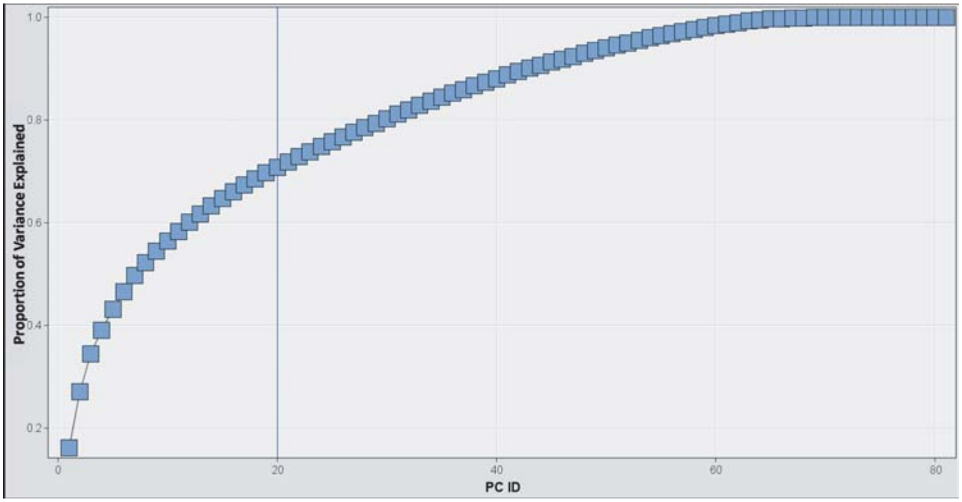


Figure 20-8: A scree plot shows the amount of information included in the first *n* principal components.

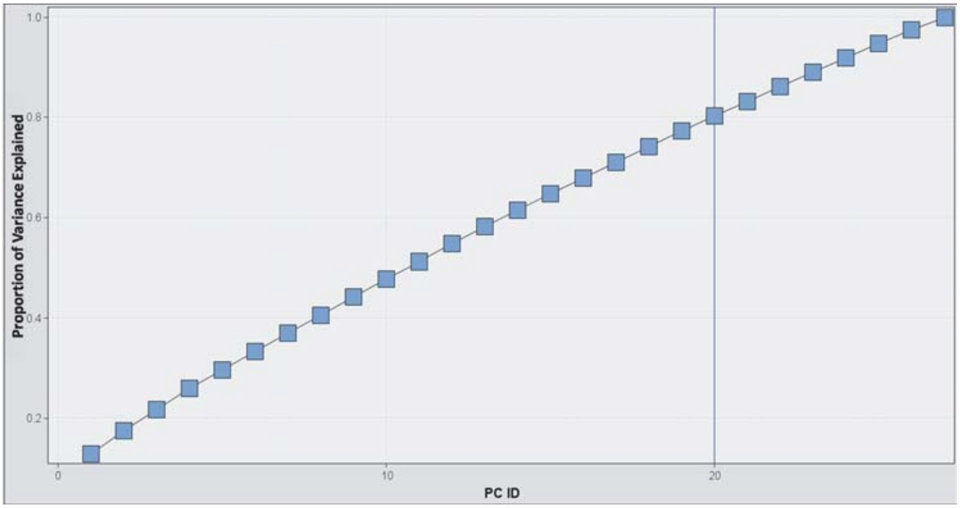


Figure 20-9: In this example with 27 flags, the scree plot for the principal components is not steep, indicating that the principal components do not efficiently capture the information in the original data.

Table 20-3: Coefficients for the First Principal Component for the Education Variables, Weighted by ZIP Code and Weighted by Population

VARIABLE	DESCRIPTION	COEFFICIENT UNWEIGHTED	COEFFICIENT WEIGHTED
Popedunone	No Education	−0.1609	−0.2313
Popedunohs	No High School	−0.3526	−0.3601
Popedusomehs	Some High School	−0.3415	−0.3970
Popeduhsgrad	High School Graduate	−0.3402	−0.3202
popedusomecol	Some College	0.2016	0.1163
Popeduassoc	2-Year College Degree	0.2010	0.1770
Popedubach	4-Year College Degree	0.4701	0.4444
Popedumast	Master's Degree	0.4206	0.4221
popeduprofdoct	Doctorate or Professional Degree	0.3720	0.3690

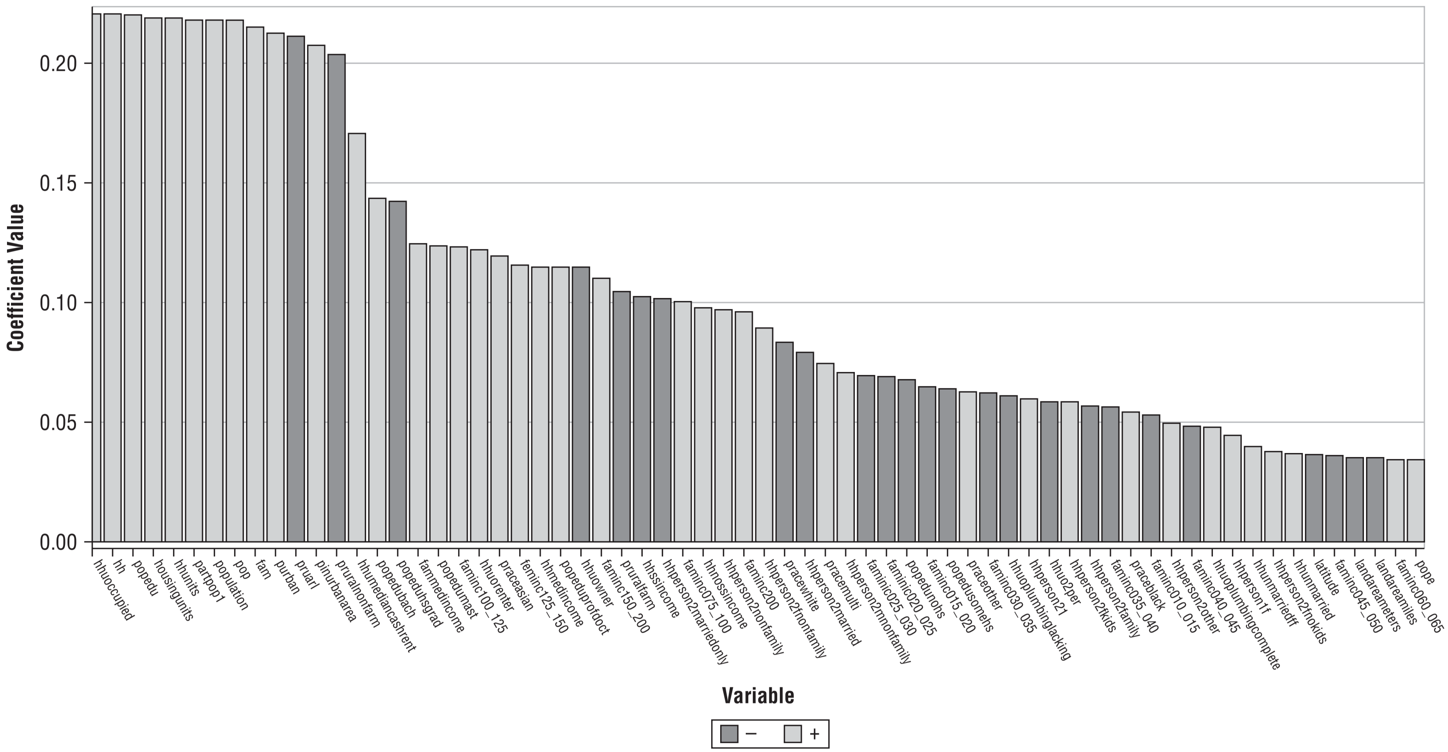


Figure 20-10: The most important variables in the first principal component are on the left, with the height of the bars showing each variable's contribution (the chart continues to the right, with variables that have smaller coefficients falling off the chart). Lighter shading is positive; darker shading is negative.

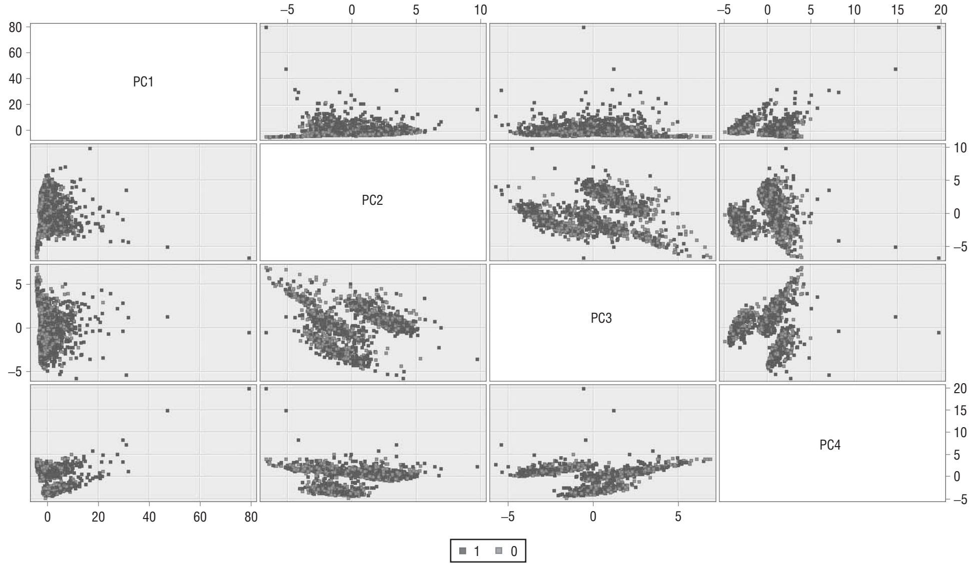


Figure 20-11: These scatter plots show the data along the first four principal components, plotted pairwise. The two charts on the upper left, for instance, are the scatter plot using the first two principal components.

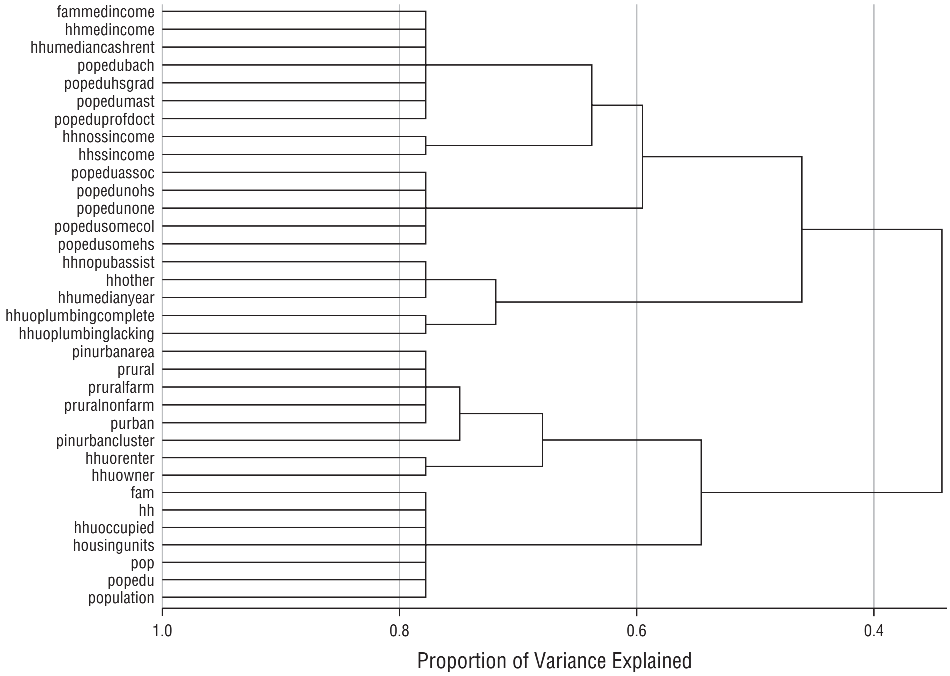


Figure 20-12: This tree shows an example of variable clustering for some of the census variables. The variables at the top, for instance, all indicate highly educated wealthy regions (or, equivalently, poorly educated, impoverished ones).

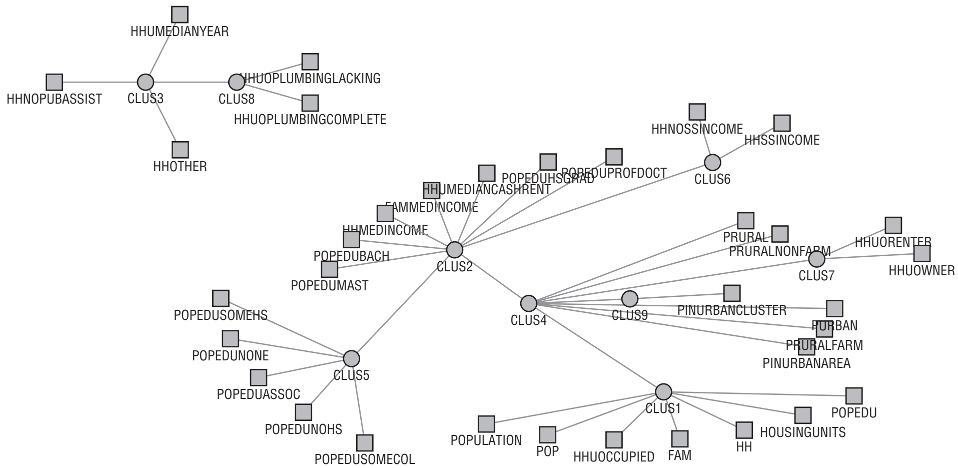


Figure 20-13: This cluster plot shows an alternative way of looking at the variable clusters as nodes in a graph that show relationships among the variables.

Table 20-4: Example of Data for Six ZIP Codes

ZIPCODE	LANDAREAMILES	HHMEDINCOME	HHNOPUBASSIST	NOHHDIPLOMA	COLDEGREE
10011	0.6	$61,986	98.5%	3.5%	68.6%
33158	3.1	$118,410	99.3%	2.6%	60.6%
33193	13.7	$39,990	96.5%	10.5%	19.7%
55343	8.3	$44,253	97.0%	3.3%	38.1%
94518	5.6	$64,429	95.7%	4.7%	32.3%
98053	32.4	$96,028	99.4%	2.0%	57.8%

Table 20-5: Correlation Matrix for Five Variables Used for Variable Clustering Example

	LANDAREAMILES	HHMEDINCOME	HHNOPUBASSIST	NOHHDIPLOMA	COLDEGREE
landarea-miles	1.000	−0.129	−0.012	0.019	−0.075
hhmedin-come	−0.129	1.000	0.327	−0.433	0.679
hhnopub-assist	−0.012	0.327	1.000	−0.129	0.163
nohhdi-ploma	0.019	−0.433	−0.129	1.000	−0.492
coldegree	−0.075	0.679	0.163	−0.492	1.000

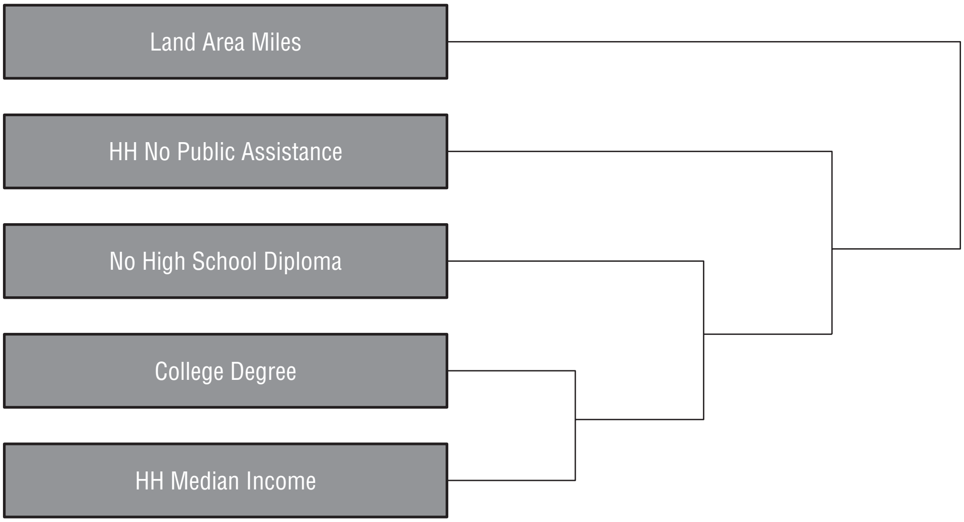


Figure 20-14: Tree structure for variables clustered using correlation and principal components.

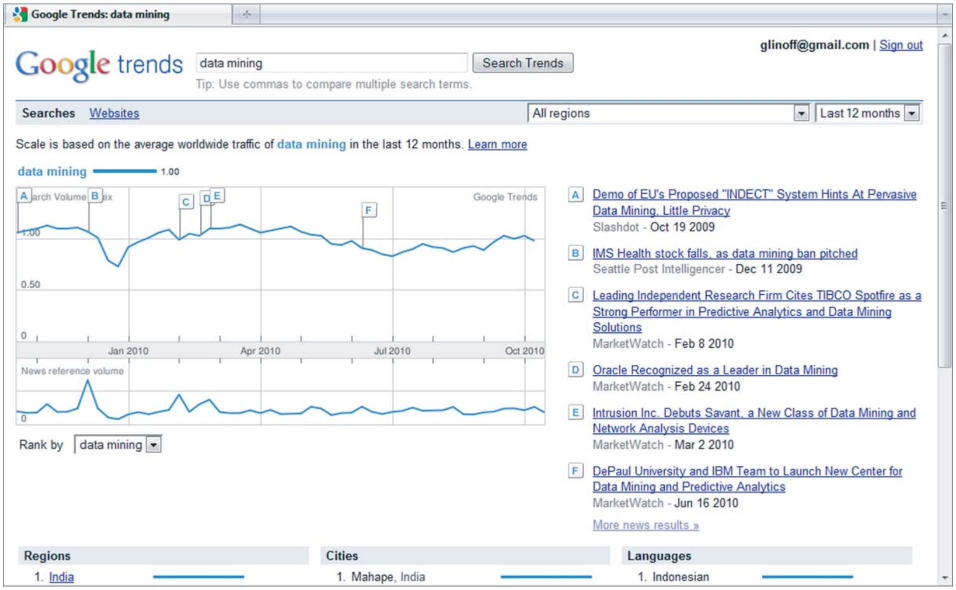


Figure 21-1: Google trends provides information about the popularity of search terms over time.

Table 21-1: Counts of Unique Terms and Total Words for Translations of the Bible[2]

LANGUAGE	UNIQUE TERMS	TERM COUNT
English	12,335	789,744
French	20,428	812,947
Spanish	28,456	704,004
Russian	47,226	560,524
Arabic	55,300	440,435

[2]Bader B. and Chew P, 2010. "Algebraic Techniques for Multilingual Document Clustering." In *Text Mining Applications and Theory*, page 23. (Michael W. Berry and Jacob Kogan, eds.). Wiley.

Table 21-2: Boycott Stops with Respect to Stop Types

STOP TYPE	TOTAL	BOYCOTT	PERCENT
Editorial Stop	4,893	4,378	89.47%
Vacation	34,678	1,055	3.04%
Other	8,811	349	3.96%
Missing	6,083	292	4.80%
Total		**6,074**	

Table 21-3: Six Types of Codes Used to Classify News Stories

CATEGORY	# CODES	# DOCS	# OCCURRENCES
Government (G/)	28	3,926	4,200
Industry (I/)	112	38,308	57,430
Market Sector (M/)	9	38,562	42,058
Product (P/)	21	2,242	2,523
Region (R/)	121	47,083	116,358
Subject (N/)	70	41,902	52,751

Table 21-4: ClassifiedNeighborsofaNot-Yet-Classifi ed Story

NEIGHBOR	DISTANCE	WEIGHT	CODES
1	0.076	0.924	R/FE,R/CA,R/CO
2	0.346	0.654	R/FE,R/JA,R/CA
3	0.369	0.631	R/FE,R/JA,R/MI
4	0.393	0.607	R/FE,R/JA,R/CA

$$d_{\text{classification}}(A,B) = (1 - \text{score}(A,B)) \,/\, \text{score}(A,A)$$

Equation 30

Table 21-5: Code Scores for the Not-Yet-Classified Story

CODE	1	2	3	4	SCORE
R/CA	0.924	0.654	0.000	0.607	2.185
R/CO	0.924	0.000	0.000	0.000	0.924
R/FE	0.924	0.654	0.631	0.607	2.816
R/JA	0.000	0.654	0.631	0.607	1.892
R/MI	0.000	0.654	0.000	0.000	0.624

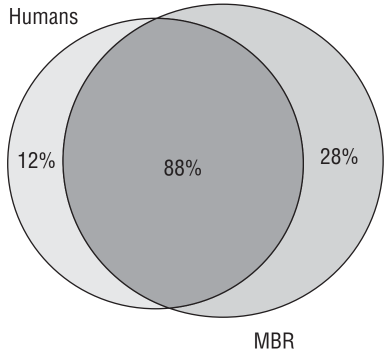


Figure 21-2: A comparison of results by human editors and by MBR on assigning codes to news stories.

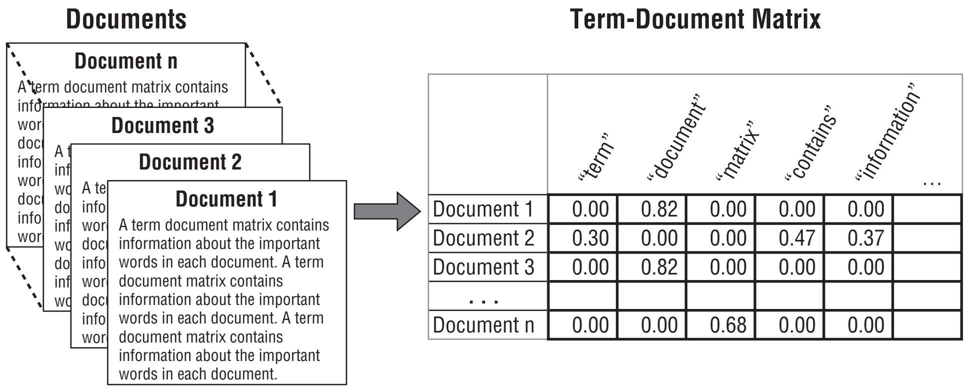


Figure 21-3: A term-document matrix contains information about the important words in each document.

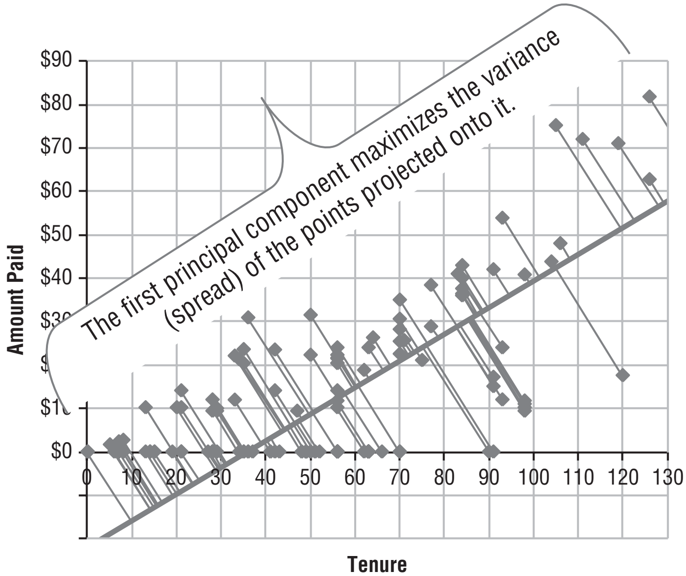
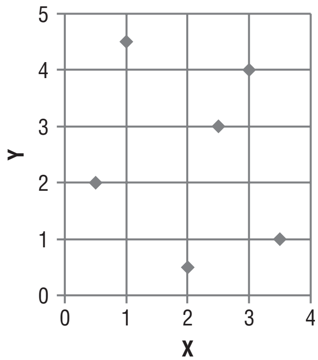
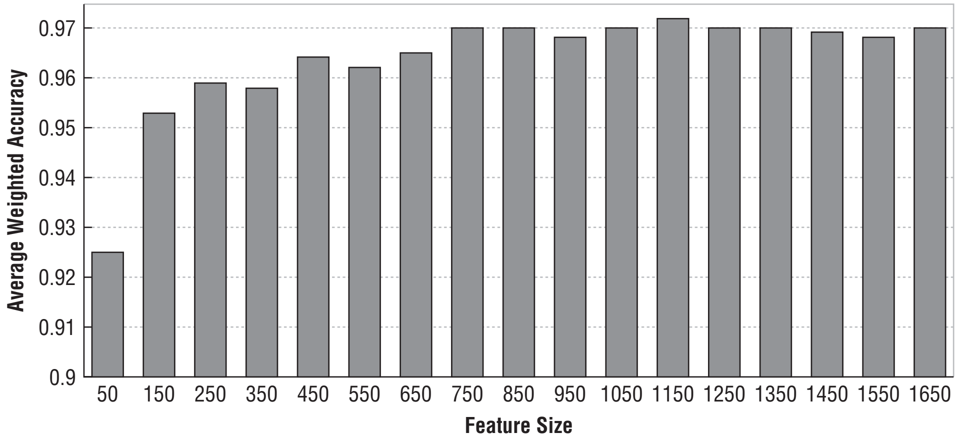


Figure 21-4: The first principal component maximizes the variance of the points on the projected line.



0.5	2.0
2.0	0.5
1.0	4.5
2.5	3.0
3.0	4.0
3.5	1.0

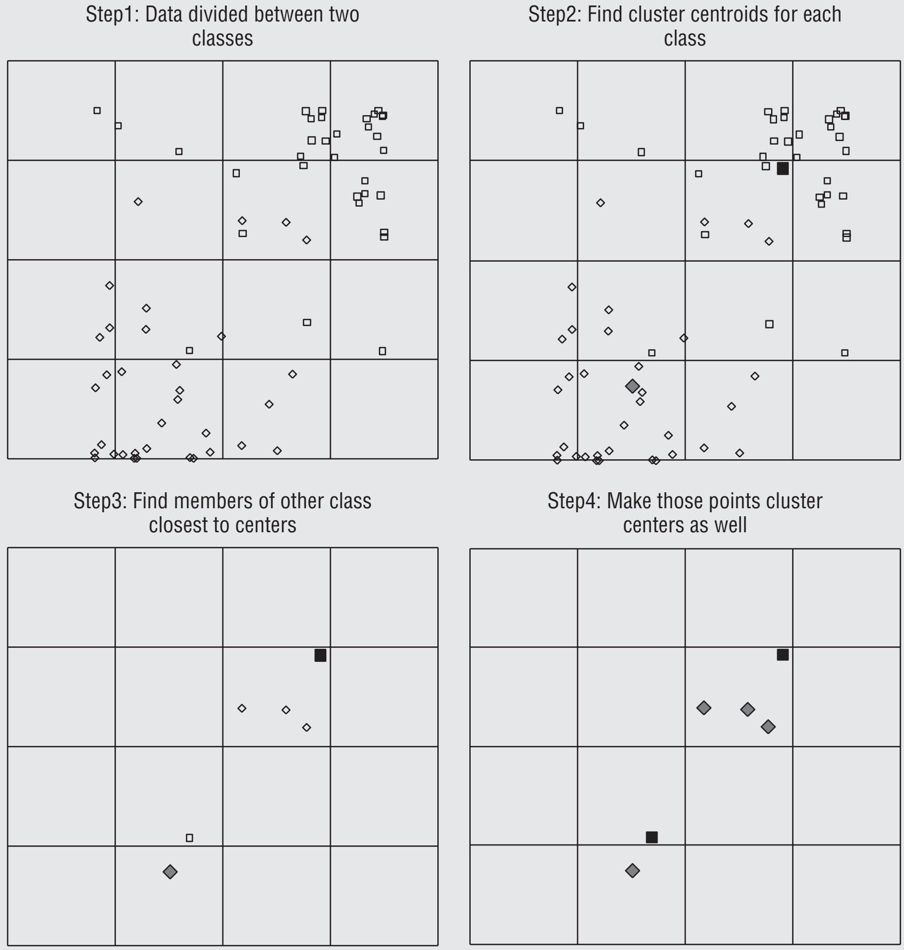
Figure 21-5: One interpretation of a matrix is that it is just a collection of points.



Courtesy of Prof. Eric Jiang, University of San Diego

Figure 21-6: The naïve Bayesian classifier tends to work better as more terms are added, although the improvement plateaus around 750 terms.

AN ALTERNATIVE APPROACH FOR CLASSIFYING E-MAILS (*continued*)



Augmented clustering in LSI adds new centroids, which are members of the other class that are close to the original cluster centroids.

Email Clusters

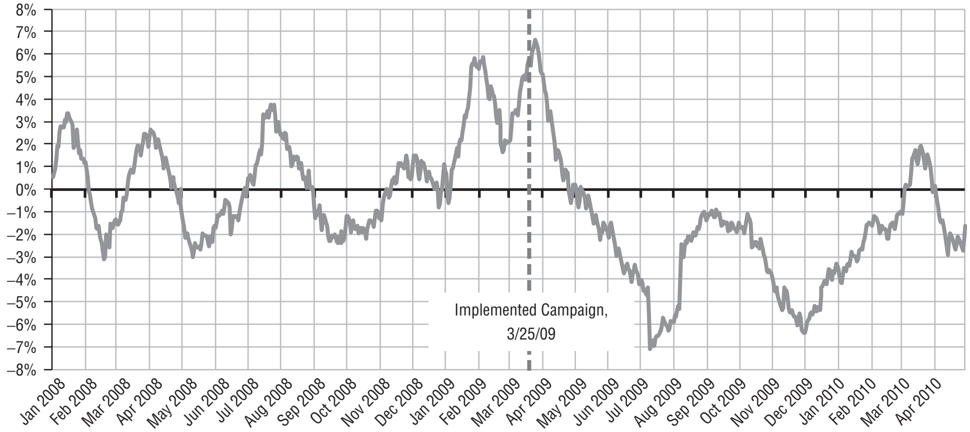


Figure 21-7: The implementation of the new call center interface, inspired by text mining efforts, reduced the average call duration by a noticeable amount.

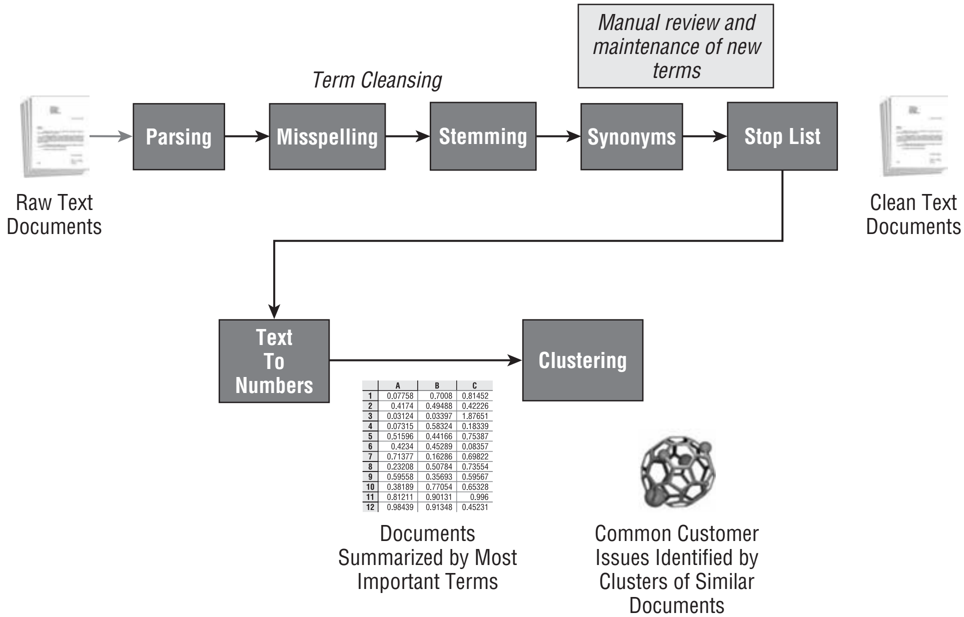


Figure 21-8: The process for building document clusters involves many steps to transform the data into a structure usable for analysis.