7

Datasets

7.1 Tips

Source: Bryant, P. G. and Smith, M. A. (1995), *Practical Data Analysis: Case Studies in Business Statistics*, Richard D. Irwin Publishing, Homewood, IL.

Number of cases: 244
Number of variables: 8
Description: Food servers' tips in restaurants may be influenced by many factors, including the nature of the restaurant, size of the party, and table locations in the restaurant. Restaurant managers need to know which factors matter when they assign tables to food servers. For the sake of staff morale, they usually want to avoid either the substance or the appearance of unfair treatment of the servers, for whom tips (at least in restaurants in the United States) are a major component of pay.

In one restaurant, a food server recorded the following data on all customers they served during an interval of two and a half months in early 1990. The restaurant, located in a suburban shopping mall, was part of a national chain and served a varied menu. In observance of local law the restaurant offered seating in a non-smoking section to patrons who requested it. Each record includes a day and time, and taken together, they show the server's work schedule.

Variable	Explanation
obs	Observation number
totbill	Total bill (cost of the meal), including tax, in US dollars
tip	Tip (gratuity) in US dollars
sex	Sex of person paying for the meal (0=male, 1=female)
smoker	Smoker in party? (0=No, 1=Yes)
day	3=Thur, 4=Fri, 5=Sat, 6=Sun
time	0=Day, 1=Night
size	Size of the party

Primary question: What are the factors that affect tipping behavior?

Data restructuring: A new variable tiprate = tip/totbill should be calculated.

Analysis notes: This dataset is fabulously simple and yet fascinating. The original case study fits a traditional regression model, using tiprate as a response variable. The only important variable emerging from this model is size: As size increases, tiprate decreases. The reader may have noticed that restaurants seem to know about this association, because they often include a service charge for larger dining parties. (There has been at least one lawsuit regarding this service charge.) Here, this association explains only 2% of all the variation in tip rate — it is a very weak model! There are many other interesting features in the data, as described in this book.

Data files:

`tips.csv`, `tips.xml`

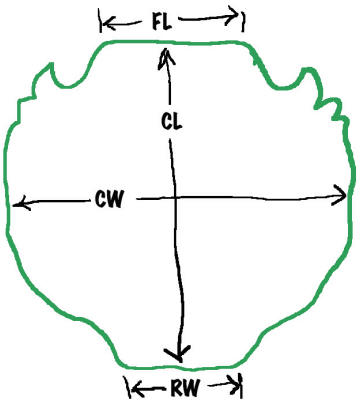
7.2 Australian Crabs

Source: Campbell, N. A. & Mahon, R. J. (1974), A Multivariate Study of Variation in Two Species of Rock Crab of genus *Leptograpsus*, *Australian Journal of Zoology* **22**, 417–425. The data was first brought to our attention by Venables & Ripley (2002) and Ripley (1996).

Number of rows: 200
Number of variables: 8

Description: Measurements on rock crabs of the genus *Leptograpsus*. One species *L. variegatus* had been split into two new species, previously grouped by color, orange and blue. Preserved specimens lose their color, so it was hoped that morphological differences would enable museum specimens to be classified. There are 50 specimens of each sex of each species, collected on site at Fremantle, Western Australia. For each specimen, five measurements were made, using vernier calipers.

Variable	Explanation
species	orange or blue
sex	male or female
index	1–200
frontal lip (FL)	length, in mm
rear width (RW)	width, in mm
carapace length (CL)	length of midline of the carapace, in mm
carapace width (CW)	maximum width of carapace, in mm
body depth (BD)	depth of the body; for females, measured after displacement of the abdomen, in mm



Primary question: Can we determine the species and sex of the crabs based on these five morphological measurements?

Data restructuring: A new class variable distinguishing all four groups would be useful.

Analysis notes: All physical measurements on the crabs are strongly positively correlated, and this is the main structure in the data. For this reason, it may be helpful to sphere the data and use principal components instead of raw variables in any analysis. Despite this strong association, there are a lot of differences among the four groups. Species can be perfectly distinguished by physical characteristics, and so can the sex of the larger crabs. In previous analyses, the measurements were logged, but we have not found this to be necessary.

Data files:

`australian-crabs.csv`, `australian-crabs.xml`

7.3 Italian Olive Oils

Source: Forina, M., Armanino, C., Lanteri, S. & Tiscornia, E. (1983), Classification of Olive Oils from their Fatty Acid Composition, in Martens, H. and Russwurm Jr., H., eds, Food Research and Data Analysis, Applied Science Publishers, London, pp. 189–214. It was brought to our attention by Glover & Hopke (1992).

Number of rows: 572
Number of variables: 10

Description: This data consists of the percentage composition of fatty acids found in the lipid fraction of Italian olive oils. The data arises from a study to determine the authenticity of an olive oil.

Variable	Explanation
region	Three "super-classes" of Italy: North, South, and the island of Sardinia
area	Nine collection areas: three from the region North (Umbria, East and West Liguria), four from South (North and South Apulia, Calabria, and Sicily), and two from the island of Sardinia (inland and coastal Sardinia).
palmitic, palmitoleic, stearic, oleic, linoleic, linolenic, arachidic, eicosenoic	fatty acids, % × 100

Primary question: How do we distinguish the oils from different regions and areas in Italy based on their combinations of the fatty acids?

Data restructuring: None needed.

Analysis notes: There are nine classes (areas) in this data, too many to easily classify. A better approach is to take advantage of the hierarchical structure in the data, partitioning by region before starting.

Some of the classes are easy to distinguish, but others present a challenge. The clusters corresponding to classes all have different shapes in the eight-dimensional data space.

Data files:

`olive.csv`, `olive.xml`

7.4 Flea Beetles

Source: Lubischew, A. A. (1962), On the Use of Discriminant Functions in Taxonomy, Biometrics **18**, 455–477.

Number of rows: 74
Number of variables: 7

Description: This data contains physical measurements on three species of flea beetles.

Variable	Explanation
species	*Ch. concinna*, *Ch. heptapotamica*, and *Ch. heikertingeri*
tars1	width of the first joint of the first tarsus in microns
tars2	width of the second joint of the first tarsus in microns
head	the maximal width of the head between the external edges of the eyes in 0.01 mm
aede1	the maximal width of the aedeagus in the fore-part in microns
aede2	the front angle of the aedeagus (1 unit = 7.5 degrees)
aede3	the aedeagus width from the side in microns

Primary question: How do we classify the three species?

Data restructuring: None needed.

Analysis notes: This straightforward dataset has three very well separated elliptically shaped clusters. It is fun to cluster the data with various algorithms and see how many get the clusters wrong.

Data files:

`flea.csv`, `flea.xml`

7.5 PRIM7

Source: First used in Friedman, J. H. & Tukey, J. W. (1974), A Projection Pursuit Algorithm for Exploratory Data Analysis, *IEEE Transactions on Computing C* **23**, 881–889. Originally from Ballam, J., Chadwick, G. B., Guiragossian, G. T., Johnson, W. B., Leith, D. W. G. S., & Moriyasu, K. (1971), Van hove analysis of the reaction $\pi^- p \longrightarrow \pi^-\pi^-\pi^+ p$ and $\pi^+ p \longrightarrow \pi^+\pi^+\pi^- p$ at 16 gev/c*, *Physics Review D* **4**(1), 1946–1966.

Number of cases: 500
Number of variables: 7

Description: This data contains observations taken from a high-energy particle physics scattering experiment that yielded four particles. The reaction $\pi_b^+ p_t \rightarrow p\pi_1^+\pi_2^+\pi^-$ can be described completely by seven independent measurements. Below, $\mu^2(A, B, \pm C) = (E_A + E_B \pm E_C)^2 - (P_A + P_B \pm P_C)^2$ and $\mu^2(A, \pm B) = (E_A \pm E_B)^2 - (P_A \pm P_B)^2$, where E and P represent the particle's energy and momentum, respectively, as measured in billions of electron volts. The notation $(p)^2$ represents the inner product P/P. The ordinal assignment of the two π^+'s was done randomly.

Variable	Explanation
X1	$\mu^2(\pi^-, \pi_1^+, \pi_2^+)$
X2	$\mu^2(\pi^-, \pi_1^+)$
X3	$\mu^2(p, \pi^-)$
X4	$\mu^2(\pi^-, \pi_2^+)$
X5	$\mu^2(p, \pi_1^+)$
X6	$\mu^2(p, \pi_1^+, -p_t)$
X7	$\mu^2(p, \pi_2^+, -p_t)$

Primary question: What are the clusters in the data?

Data restructuring: None needed, although it is helpful to sphere the data to principal component coordinates before using projection pursuit.

Analysis notes: The case study illustrates the strength of our graphical methods in detecting sparse structure in high-dimensional space. It is a stunning look at uncovering a very geometric structure in high-dimensional space. A combination of interactive brush controls and motion graphics reveals that the points lie on a structure comprising connected low-dimensional pieces: a two-dimensional triangle, with two linear pieces extending from each vertex (Cook et al. 1995). Various graphical tools specifically facilitated the discovery of the structure: Plots of low-dimensional projections of the seven-dimensional object allowed discovery of the low-dimensional pieces, highlighting allowed the pieces to be recorded or marked, and animating many projections into a movie over time allowed the pieces to be reconstructed into the full shape. The data was 20 years old by the time these visual methods were applied to it, and the structure is known by physicists.

Data files:

`prim7.csv`, `prim7.xml`

7.6 Tropical Atmosphere-Ocean Array (TAO)

Source: The data from the array, along with current updates, can be viewed on the web at `http://www.pmel.noaa.gov/tao`.

Number of cases: 736
Number of variables: 8

Description: The El Niño/Southern Oscillation (ENSO) cycle of 1982–1983, the strongest of the century, created many problems throughout the world. Parts of the world such as Peru and the United States experienced destructive flooding from increased rainfall, whereas countries in the western Pacific experienced drought and devastating brush fires. The ENSO cycle was neither predicted nor detected until it was near its peak, which highlighted the need for an ocean observing system to support studies of large-scale ocean-atmosphere interactions on seasonal-to-interannual time scales.

This observing system was developed by the international Tropical Ocean Global Atmosphere (TOGA) program. The Tropical Atmosphere Ocean (TAO) array consists of nearly 70 moored buoys spanning the equatorial Pacific, measuring oceanographic and surface meteorological variables critical for improved detection, understanding and prediction of seasonal-to-interannual climate variations originating in the tropics, most notably those related to the ENSO cycles.

The moorings were developed by the National Oceanic and Atmospheric Administration's (NOAA) Pacific Marine Environmental Laboratory (PMEL). Each mooring measures air temperature, relative humidity, surface winds, sea surface temperatures, and subsurface temperatures down to a depth of 500 meters, and a few of the buoys measure currents, rainfall, and solar radiation.

The TAO array provides real-time data to climate researchers, weather prediction centers, and scientists around the world. Forecasts for tropical Pacific Ocean temperatures for one to two years in advance can be made using the ENSO cycle data. These forecasts are possible because of the moored buoys, along with drifting buoys, volunteer ship temperature probes, and sea level measurements.

Variable	Explanation
year	1993 (a normal year), 1997 (an El Niño year), for November, December, and the January of the following year.
latitude	0°, 2°S, 5°S only.
longitude	110°W, 95°W only.
sea surface temp (SST)	measured in °C, at 1 m below the surface
air temp (AT)	measured in °C, at 3 m above the sea surface.
humidity (Hum)	relative humidity, measured 3 m above the sea surface.
uwind	east–west component of wind, measured 4 m above sea surface: positive means the wind is blowing toward the east.
vwind	north–south component of the wind, measured 4 m above sea surface: a positive sign means that the wind is blowing toward the north.

Primary question: Can we detect the El Niño event, based on sea surface temperature? What changes in the other observed variables occur during this event?

Data restructuring: This subset comes from a larger dataset extracted from the web site mentioned above. That data runs from March 7, 1980 to December 31, 1998, from −10°S to 10°N, and from 130°E to 90°W. There are 178,080 recorded measurements, with time, latitude, longitude, and five atmospheric variables for each record.

Longitude is measured with east as positive and west as negative units, with the prime meridian in Greenwich, UK, at 0°. The buoys are moored in the Pacific Ocean on the other side of the globe, with measurements on either side of the International Date Line (−180° = 180°)! This makes it very difficult to plot the data using numerical scales. We added new categorical variables marking the moored location of the buoys, making it easier to plot the spatial coordinates. These variables also make it easier to identify the buoys, because they tend to break free of the moorings and drift occasionally.

Analysis notes: One hurdle in working with this data is the large number of missing values. The missingness needs to be explored as a first step, and missing values need to be imputed before an analysis.

The larger data is cumbersome to work with, because of the missing values and the spatiotemporal context, but it has some interesting features. Plotting the latitude and longitude reveals that some buoys tend to drift, quite substantially at times, and that they are eventually retrieved and reattached to the moorings!

There is a massive El Niño event in the last year of this larger subset, 1997–1998, and it is visible at some locations when time series of sea surface temperature are plotted. Smaller El Niño events are visible at several other years. Changes in other variables are noticeable during these events too, particularly in one of the wind components.

Data files:

`tao.csv, tao.xml`	Small subsets mostly used in the book.
`tao-full.csv, tao-full.xml`	The full data, not commonly used in the book, but included for data context.

7.7 Primary Biliary Cirrhosis (PBC)

Source: Distributed with Fleming & Harrington, *Counting Processes and Survival Analysis*, Wiley, New York, 1991, and available from `http://lib.stat.cmu.edu/datasets`. A description of the clinical background for the trial and the covariates recorded here is in Chapter 0, especially Section 0.2. It was originally from the Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver conducted between 1974 and 1984. A more extended discussion can be found in Dickson et al., Prognosis in primary biliary cirrhosis: model for decision making, *Hepatology* **10**, 1–7 (1989) and in Markus et al., Efficiency of liver transplantation in patients with primary biliary cirrhosis, *New England Journal of Medicine* **320**, 1709–1713 (1989).

Number of cases: 312
Number of variables: 20

Description: A total of 424 PBC patients, referred to the Mayo Clinic during that ten-year interval, met eligibility criteria for the randomized placebo controlled trial of the drug D-penicillamine. The first 312 cases in the dataset refer to subjects who participated in the randomized trial; they contain largely complete data. The additional 112 subjects did not participate in the clinical trial but consented to have basic measurements recorded and to be followed for survival; they are not represented here.

Variable	Explanation
id	
fu.days	number of days between registration and the earlier of death, transplantation, or study analysis time in July 1986
status	status is coded as 0=censored, 1=censored due to liver tx, 2=death
drug	1=D-penicillamine, 2=placebo
age	in days
sex	0=male, 1=female
ascites	presence of ascites: 0=no 1=yes
hepatom	presence of hepatomegaly: 0=no 1=yes
spiders	presence of spiders: 0=no 1=yes
edema	presence of edema: 0=no edema and no diuretic therapy for edema; .5 = edema present without diuretics, or edema resolved by diuretics; 1 = edema despite diuretic therapy
bili	serum bilirubin in mg/dl
chol	serum cholesterol in mg/dl
albumin	in gm/dl
copper	urine copper in μg/day
alk.phos	alkaline phosphatase in U/l
sgot	SGOT in U/ml
trig	triglycerides in mg/dl
platelet	platelets per cubic ml/1,000
protime	prothrombin time in seconds
stage	histologic stage of disease

Primary question: How do the different drugs affect the patients?

Data restructuring: Only records corresponding to patients that were in the original clinical trial were included in this data. The remaining records had too many systematic missing values.

Analysis notes: Handling missing values is an interesting exercise in this data, and experimenting with data transformations.

Data files:

`pbc.csv`

7.8 Spam

Source: This was data collected at Iowa State University (ISU) by the 2003 Statistics 503 class.

Number of cases: 2,171
Number of variables: 21

Description: Every person monitored their email for a week and recorded information about each email message; for example, whether it was spam, and what day of the week and time of day the email arrived. We want to use this information to build a spam filter, a classifier that will catch spam with high probability but will never classify good email as spam.

Variable	Explanation
isuid	Iowa State U. student id (1–19)
id	email id (a unique message descriptor)
day of week	sun, mon, tue, wed, thu, fri, sat
time of day	0–23 (only integer values)
size.kb	size of email in kilobytes
box	yes if sender is in recipient's in- or outboxes (i.e., known to recipient); else no
domain	high-level domain of sender's email address: e.g., .edu, .ru
local	yes if sender's email is in local domain, else no; local addresses have the form xx@yy.iastate.edu
digits	number of numbers (0-9) in the sender's name: e.g., for lottery2003@yahoo.com, this is 4.
name	"name" (if first and last names are present), "single" (if only one name is present), or empty
capct	% capital letters in subject line
special	number of non-alphanumeric characters in subject
credit	yes if subject line includes one of mortgage, sale, approve, credit; else no
sucker	yes if subject line includes one of the words earn, free, save; else no
porn	yes if subject line includes one of nude, sex, enlarge, improve; else no
chain	yes if subject line includes one of pass, forward, help; else no
username	yes if subject includes recipient's name or login; else no
large.text	yes if email is HTML® and includes test for large font, defined as size = +3 or size = 5 or higher; else no
spampct	probability of being spam, according to ISU spam filter.
category	extended spam/mail category: "com," "list," "news," "ord"
spam	yes if spam; else no

Primary question: Can we distinguish between spam and "ham?"

Data restructuring: A *lot* of work was done to prepare this data for analysis! It is now quite clean, and no restructuring should be needed.

Analysis notes: The ISU mail handlers examine each email message and assign it a probability of being spam. Commonly used mail readers can use this information to file email directly into the trash, or at least to a special folder. It will be interesting to compare the results of a spam filter built on our collected data with results of the university's algorithm. (The university's algorithm was classifying a lot of email from the university president as spam for a short period!) Another aside is that there is a temporal trend to spam, which seems to be more frequent at some times of day and night. We have also seen that some users get more spam than others.

Careful choice of variables is needed for building the spam filter. Only those that might be automatically calculated by a mail handler are appropriate.

There are some missing values in the data due to differences between mail handlers and the availability of information about the emails.

Spammers evolve their attacks quickly, and the recognizable signs of spam of 2003 no longer exist in 2006 spam. For example, all spam now arrives with complete Caucasian-style name fields, and messages are embedded in images rather than plain text.

Data files:

`spam.csv, spam.xml`

7.9 Wages

Source: Singer, J. D. & Willett, J. B. (2003), *Applied Longitudinal Data Analysis*, Oxford University Press, Oxford, UK. It is a subset of data collected in the National Longitudinal Survey of Youth (NLSY) described at `http://www.bls.gov/nls/nlsdata.htm`.

Number of subjects: 888
Number of variables: 15
Number of observations, across all subjects: 6,402

Description: The data was collected to track the labor experiences of male high-school dropouts. The men were between 14 and 17 years old at the time of the first survey.

Variable	Explanation
id	1–888, for each subject.
lnw	natural log of wages, adjusted for inflation, to 1990 dollars.
exper	length of time in the workforce (in years). This is treated as the time variable, with t_0 for each subject starting on their first day at work. The number of time points and values of time points for each subject can differ.
ged	when/if a graduate equivalency diploma is obtained.
black	categorical indicator of race = black.
hispanic	categorical indicator of race = hispanic.
hgc	highest grade completed.
uerate	unemployment rates in the local geographic region at each measurement time.

Primary question: How do wages change with workforce experience?

Data restructuring: The data in its original form looked as follows, where time-independent variables have been repeated for each time point:

id	lnw	exper	black	hispanic	hgc	uerate
31	1.491	0.015	0	1	8	3.215
31	1.433	0.715	0	1	8	3.215
31	1.469	1.734	0	1	8	3.215
31	1.749	2.773	0	1	8	3.295
31	1.931	3.927	0	1	8	2.895
31	1.709	4.946	0	1	8	2.495
31	2.086	5.965	0	1	8	2.595
31	2.129	6.984	0	1	8	4.795
36	1.982	0.315	0	0	9	4.895
36	1.798	0.983	0	0	9	7.400
36	2.256	2.040	0	0	9	7.400
36	2.573	3.021	0	0	9	5.295
36	1.819	4.021	0	0	9	4.495
36	2.928	5.521	0	0	9	2.895
36	2.443	6.733	0	0	9	2.595
36	2.825	7.906	0	0	9	2.595
36	2.303	8.848	0	0	9	5.795
36	2.329	9.598	0	0	9	7.600

It was restructured into two tables of data. One table contains the time-independent measurements identified by subject id, and the other table contains the time-dependent variables:

id	black	hispanic	hgc
31	0	1	8
36	0	0	9

id	lnw	exper	uerate
31	1.491	0.015	3.215
31	1.469	1.734	3.215
31	1.749	2.773	3.295
31	1.931	3.927	2.895
31	1.709	4.946	2.495
31	2.086	5.965	2.595
31	2.129	6.984	4.795
36	1.982	0.315	4.895
36	1.798	0.983	7.400
36	2.256	2.040	7.400
36	2.573	3.021	5.295
36	1.819	4.021	4.495
36	2.928	5.521	2.895
36	2.443	6.733	2.595
36	2.825	7.906	2.595
36	2.303	8.848	5.795
36	2.329	9.598	7.600

Analysis notes: Singer & Willett (2003) use this data to illustrate fitting mixed linear models to ragged time indexed data. The analysis reports that the average growth in wages is about 4.7% for each year of experience. There is no difference between whites and Hispanics, but a big difference from blacks. The model uses a linear trend (on the log wages) to follow these patterns. The within-variance component of the model is significant, which indicates that the variability for each person is dramatically different. It does not tell us, however, in what ways people differ, and which people are similar.

The data is fascinating from a number of perspectives. Although on average wages tend to increase with time, the temporal patterns of individual wages vary dramatically. Some men experience a decline in wages over time, others a more satisfying increase, and yet others have very volatile wage histories. There is also a strange pattern differentiating the wage histories of black men from whites and Hispanics.

Data files:

`wages.xml`

7.10 Rat Gene Expression

Source: X. Wen, S. Fuhrman, G. S. Michaels, D. B. Carr, S. Smith, J. L. Barker & R. Somogyi (1998), Large-scale temporal gene expression mapping of central nervous system development, in *Proceedings of the National Academy of Science* **95**, pp. 334–339, available on the web from `http://pnas.org`.

Number of cases: 112
Number of variables: 17

Description: This small subset of data is from a larger study of rat development using gene expression. The subset contains gene expression for nine developmental times, taken by averaging several replicates and normalizing the values using the maximum value for the gene.

Variable	Explanation
E11	11-day-old embryo
E13	13-day-old embryo
E15	15-day-old embryo
E18	18-day-old embryo
E21	21-day-old embryo
P0	at birth
P7	at 7 days
P14	at 14 days
A	adult
Class1 (*Class2*)	Functional classes representing expert's best guess: 1 neuro-glial markers *(1 markers)*, 2 neurotransmitter metabolizing enzymes *(2 neurotransmitter receptors, 3 GABA-A receptors, 4 glutamate receptors, 5 acetylcholine receptors, 6 serotonin receptors)*, 3 peptide signaling *(7 neurotrophins, 8 heparin-binding growth factors, 9 insulin/IGF)*, 4 diverse *(10 intracellular signaling, 11 cell cycle, 12 transcription factor, 13 novel/EST, 14 other)*
avcor	average linkage, correlation distance
wardcor	Wards linkage, correlation distance
comcor	complete linkage, correlation distance
avfluor	average linkage, fluorescence distance
wardfluor	Wards linkage, fluorescence distance
comfluor	complete linkage, fluorescence distance

Primary question: Do genes within a functional class have similar gene expression patterns? How does a clustering of genes compare with the functional classes?

Data restructuring: The data has been cleaned and heavily processed. The variables summarizing the cluster analysis were added, but beyond this no more restructuring of the data should be needed.

Analysis notes: The variables are time-ordered so parallel coordinate plots are very useful here. Brushing, particularly automatically from R, to focus on one

functional class, or cluster, at a time is useful to compare patterns of gene expression between groups.

Data files:

`ratsm.csv, ratsm.xml`

7.11 Arabidopsis Gene Expression

Source: The data was collected in Dr. Basil Nikolau's lab at Iowa State University and it is discussed in Cook, Hofmann, Lee, Yang, Nikolau & Wurtele (2007).

Number of cases: 8,297
Number of variables: 9

Description: This data is from a two-factor, single replicate experiment of the following form:

	Treatment added	
	no	yes
Mutant	*M1,M2*	MT1, MT2
Wild type	W1, W2	WT1, WT2

The mutant organism is defective in the ability to synthesize an essential cofactor, which is provided by the treatment.

The data was recorded on Affymetrix GeneChip Arabidopsis Genome Arrays. The raw data was processed using robust median average and quantiles normalization available in the Bioconductor suite of tools (Bioconductor 2006).

Variable	Explanation
Gene ID	Affymetrix unique identifier for each gene. This is used as a label in the data, and for linking between multiple forms of the data.
M1	Mutant, no treatment added, replicate 1
M2	Mutant, no treatment added, replicate 2
MT1	Mutant, treatment added, replicate 1
MT2	Mutant, treatment added, replicate 2
WT1	Wild type, no treatment added, replicate 1
WT2	Wild type, no treatment added, replicate 2
WTT1	Wild type, treatment added, replicate 1
WTT2	Wild type, treatment added, replicate 2

Primary question: Which genes are differentially expressed when the treatment is not added, with special interest in the mutant genotype?

Data restructuring: Two forms are provided in different tables of data so that we can examine the replicate data values in association with the overall variation.

GeneID	M1	M2	MT1	MT2	WT1	WT2	WTT1	WTT2
1								
⋮								
8297								

GeneID	M	MT	WT	WTT
1				
⋮				
8297				
1				
⋮				
8297				

Averages across replicates are added to the short form of the data.

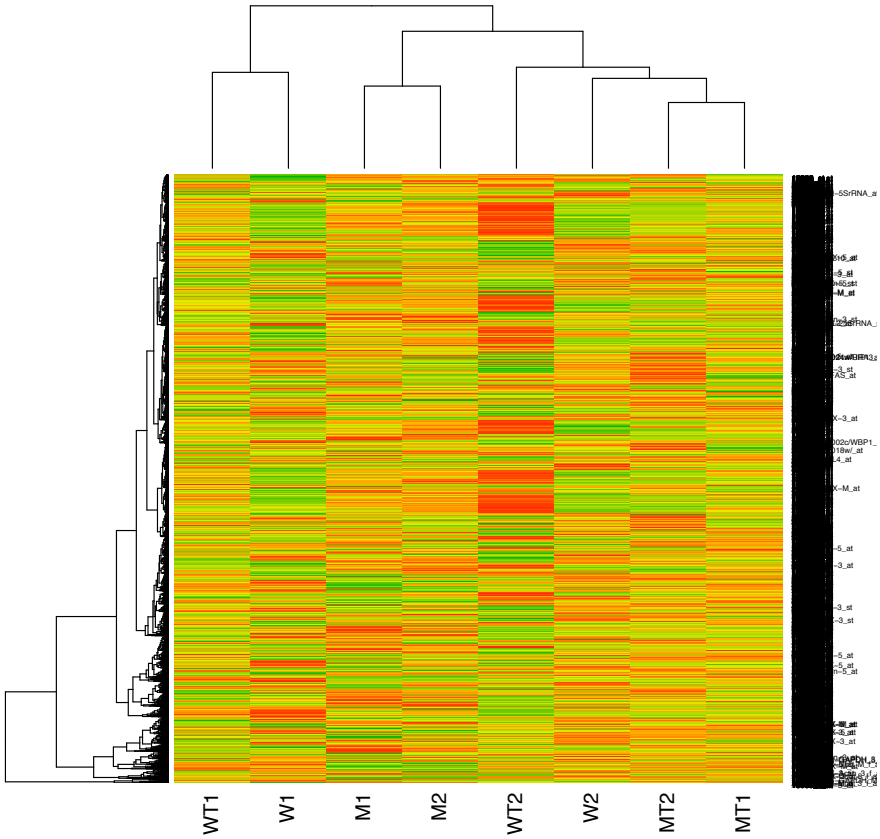
Summaries from ANOVA models fit for each gene in the data are included in the short form. These are useful for helping to detect differentially expressed genes.

Analysis notes: Difference is measured by how the individual gene varies in the replicate and by how all the genes vary in expression value.

We would also hope to see (1) small differences in expression values in the replications, (2) small differences between expression values in wild type with and without the treatment added, and (3) little difference between expression values in the mutant with the treatment and wild type.

It is important to emphasize the difference in analysis of microarray data from many other multivariate data analysis tasks. In microarray data, it is important to find a small number of genes that are behaving differently from others in an understandable way. This task involves both multiple comparisons and outlier detection. From the perspective of a traditional statistical analysis, we are merely dealing with a problem that could be solved by a t-test for comparing means. The drawback is that we have to do a test for every single gene!

Conventionally this type of data is plotted using a heatmap, shown below, but a lot of information can be obtained from linked scatterplots and parallel coordinate plots.



This field of research is evolving rapidly, and data and analysis methods change frequently.

Data files:

`arabidopsis.xml`

7.12 Music

Source: Collected by Dianne Cook.

Number of cases: 62
Number of variables: 7

Description: Using an Apple computer, each track was read into the music editing software Amadeus II, and the first 40-second clip was snipped and saved as a WAV file. (WAV is an audio format developed by Microsoft®, commonly used on Windows but becoming less popular.) These files were read into R using the package `tuneR` (Ligges 2006), which converts the audio file into numeric data. All of the CDs contained left and right channels, and variables were calculated on both channels.

Variable	Explanation
artist	Abba, Beatles, Eels, Vivaldi, Mozart, Beethoven, Enya
type	rock, classical, or new wave
lvar, lave, lmax	average, variance, maximum of the frequencies of the left channel
lfener	an indicator of the amplitude or loudness of the sound
lfreq	median of the location of the 15 highest peak in the periodogram

Primary question: Can we distinguish between rock and classical tracks? Can we group the tracks into a small number of clusters according to their similarity on audio characteristics?

Data restructuring: This dataset is very clean and simplified. The original data contained 72 variables, most of which have been excluded.

Analysis notes: Answers to the primary question might be used to arrange tracks on a digital music player or to make recommendations. Other questions of interest might be:

- Do the rock tracks have different characteristics than classical tracks?
- How does Enya compare with rock and classical tracks?
- Are there differences between the tracks of different artists?

Data files:

`music-sub.csv`, `music-sub.xml`	Subset of data used in this book. The last five tracks in the data (58–62) have the artist and type of music loosely disguised so that they can be used to test classifiers that students built using the rest of the data.
`music-all.csv`, `music-all.xml`	Full datasets, 72 variables, and a few missing values.
`music-clust.csv`, `music-clust.xml`	Subset of data, augmented with results from different cluster analyses
`music-SOM1.xml`, `music-SOM2.xml`	Different SOM models appended to the data.

7.13 Cluster Challenge

Source: Simulated by Dianne Cook.

Number of cases: 250
Number of variables: 5
Description: Simulated data included as a challenge to find the number of clusters.

Primary question: How many clusters in this data?

Data files:

`clusters-unknown.csv`

7.14 Adjacent Transposition Graph

Source: Constructed by Deborah F. Swayne.

Number of cases: 24 nodes and 36 edges in the $n = 4$ adjacent transposition graph; 120 nodes and 240 edges in the $n = 5$ graph.
Number of variables: 3 variables in the $n = 4$ graph; 4 variables in the $n = 5$ graph.
Description: The $n = N$ adjacent transposition graph is generated as follows. Start with all permutations of the sequence 1, 2, ..., N. There are $N!$ such sequences; make each one a vertex in the graph. Connect two vertices by

an edge if one permutation can be turned into the other by transposing two adjacent elements.

Principal question: Can a graph layout algorithm be used to arrange the nodes so that it is easy to understand the different permutations of rankings?

Data files:

`adjtrans4.xml`, `adjtrans5.xml`

7.15 Florentine Families

Source:

This data is widely known within the social network community, and is readily available from a number of sources. It was compiled by John Padgett from historical documents such as Kent (1978). The 16 families were chosen for analysis from a much larger collection of 116 leading Florentine families because of their historical prominence. Padgett and Ansell (1993) and Breiger and Pattison (2006) extensively analyzed the data.

We obtained it from the R package `SNAData`, by D. Scholtens (2006), part of the Bioconductor project; Scholtens obtained it from Wasserman & Faust (1994).

Number of cases: 16 nodes; two sets of edges, one with 15 edges and the other with 20.

Number of variables: 3 variables on each node; one on each edge.

Description:

The data include families who were locked in a struggle for political control of the city of Florence in around 1430. Two factions were dominant in this struggle: one revolved around the infamous Medicis, and the other around the powerful Strozzis.

Variable	Explanation
Wealth	Family net wealth in 1427 (in thousands of lira)
NumberPriorates	Number of seats on the civic council held by the family between 1282 and 1344
NumberTies	Total number of business or marriage ties
AveNTies in Business and Marital tables	Average number of business (loans, credits, joint partnerships) or marital ties per family

Primary question: How are the dominant families of old Florence connected to each other?

Data files:

`FlorentineFam.xml`, constructed from `SNAData`.

7.16 Morse Code Confusion Rates

Source: Rothkopf, E. Z. (1957), A Measure of Stimulus Similarity and Error in some Paired-Associate Learning Tasks, *Journal of Experimental Psychology* **53**, 94–101.

Number of pairwise distances: 1,260

Description:

In an experiment, inexperienced subjects were exposed to pairs of Morse codes in rapid order. The subjects had to decide whether the two codes in a pair were identical. The data were summarized in a table of confusion rates.

Confusion rates are similarity measures: Codes that are often confused are interpreted as "similar" or "close." Similarity measures are converted to dissimilarity measures so that multidimensional scaling can be applied.

Morse codes consist of sequences of short and long sounds, which are called "dots" and "dashes" and written using the characters "." and "-". Examples are:

Letter	Code
A	. -
B	- . . .
C	- . - .
D	- . .
E	.

Letter	Code
F	. . - .
G	- - .
H	. . . .
T	-
X	- . . -

Digit	Code
1	. - - - -
2	. . - - -

The codes are of varying length, with the shorter codes representing letters that are more common in English. The digits are all five-character codes.

Variable	Explanation
Length	Length of the code, rescaled to [0,1]
Dashes	Number of dashes
D	Dissimilarity between codes

Primary question: Which codes are similar and often confused with each other?

Data restructuring:

The original data came as an asymmetric 36×36 matrix of similarities, $S_{i,j}, i, j = 1, ..., n$. The values were converted to dissimilarities $D_{i,j}$ and symmetrized, using $D^2_{i,j} = S_{i,i} + S_{j,j} - 2S_{i,j}$. Two variables were derived from the Morse codes themselves, Length and Dashes.

The dissimilarity matrix was reconfigured to conform to GGobi's XML format, a set of $n\ (n-1)\ =\ 1,260$ edges with associated dissimilarity. A second set of 33 edges was added to link similar codes; it is for display only, to aid in interpretation of the configuration, and is not used by the MDS algorithm.

Analysis notes: Start with the edges turned off, and focus on the movement of the points. Add the edges when the layout is complete to understand the structure of the final configuration. Re-start MDS a few times from random starting positions, and compare the resulting configurations.

Data files:

`morsecodes.xml`

7.17 Personal Social Network

Source: Provided by Chris Volinsky and Deborah F. Swayne.

Number of cases: 140 nodes (people) and 203 edges (contacts between people).
Number of variables: two categorical variables for each node; one categorical and two real variables for each edge.

Description: This is a *personal social network*, collected by selecting one person, adding that person's contacts, each contact's contacts, and so on. In its original form, the nodes were telephone numbers and the edges represented calls from one number to another (Cortes, Pregibon & Volinsky 2003), but the privacy of individuals has been protected by disguising the telephone numbers as names and changing the meaning of the original variables.

People:

Variable	Explanation
maritalstat	categorical: married, never married, or other
hours	binary: full time or part time

Contacts:

Variable	Explanation
interactions	a measure of the amount of time spent talking
center triangle	binary; is the point part of a 3-node cycle?
log10(interactions)	base 10 log of interactions

Data files:

`snetwork.xml`