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.