Knowledge from data

Main questions are: **Why? What? Enough? Will it? How?**

- Purpose for collecting data - what hypothesis/question we seek to answer from the data.
- The variables or attributes of each case or item. Rates versus absolute values. Primary and surrogate variables.
- Response or dependent variable. Independent variables.
- The cases/items in the data set. The size of the data set - how many cases or items. Is sample size is enough to infer for the population?
- Sufficiency - can the hypothesis/question be answered using the available data.
- Gathering data. - available, anecdotal, observational, experiment.
- The type of each variable or attribute - continuous, discrete, categorial, ordinal.
- The measure of each variable like units of measurement. Normalization.

Examples

- Ranking colleges or universities (e.g. IITs).
- Ranking cities.
- GDP growth rates.

Terminology[2]

- Data capture: available, anecdotal, observational, experimental. Sample surveys, census.
- Population, sample, voluntary response sample, simple random sample, stratified random sample, multi-stage random sample. Undercoverage and non-response.
- Experiment: Experimental unit, treatment, response variable(s), explanatory variables (factors), levels of factors, outcomes (variables measured to compare treatments), confounding variables, treatment group, control group, bias, randomization, chance variation, mitigate chance variation, double blind, matched pair design, block design.

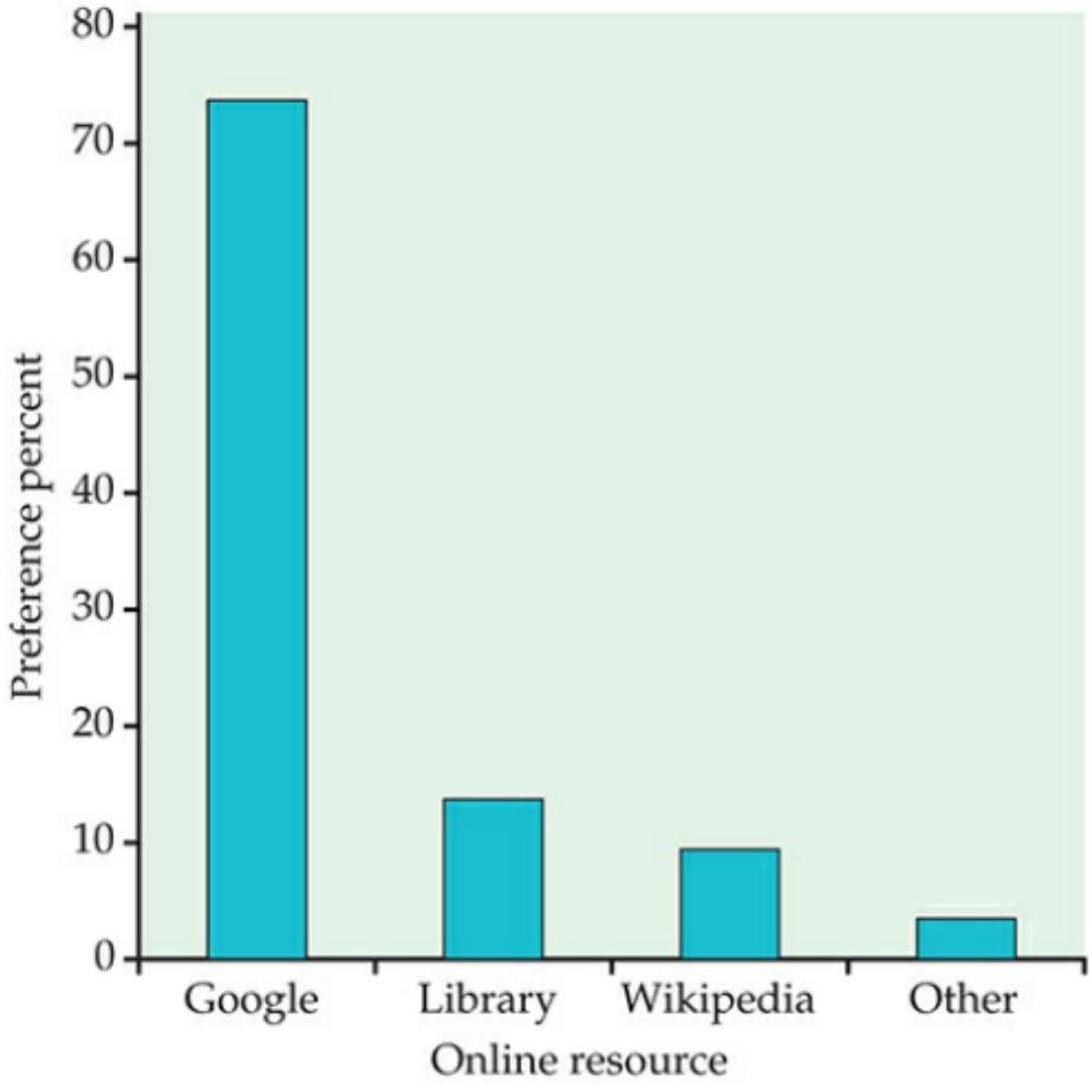
[2] Look at the text for definitions discussed in the class.

Displaying data graphically

- Bar graph - typically categories vs counts.
- Pie charts - circle is carved into sectors proportional to counts. Total has to be 100%. Typically for categories.
- Stem plots - for quantitative data. Trimming and splitting.
- Histograms.
- Scatter plots. Two variables.

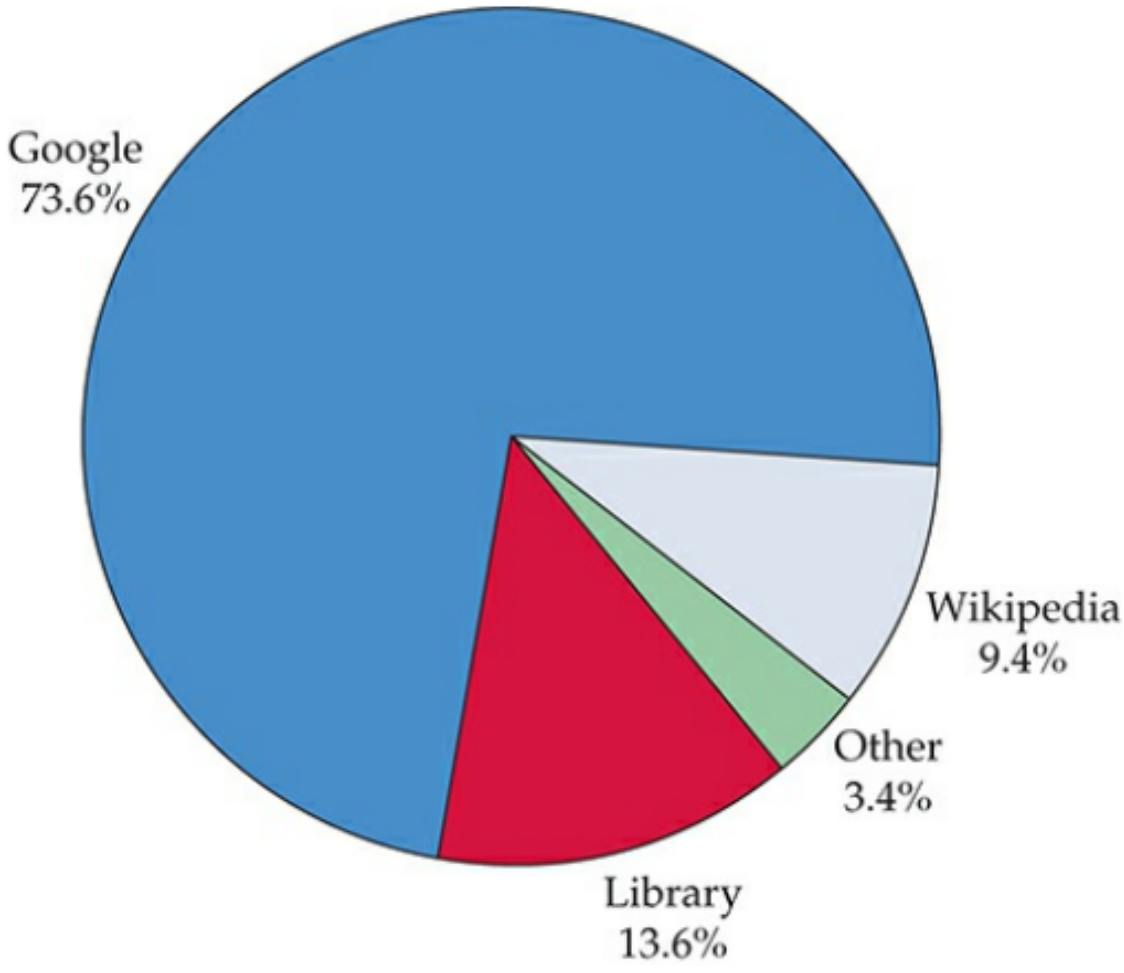
Bar graph

Online resource usage by students.



Pie chart

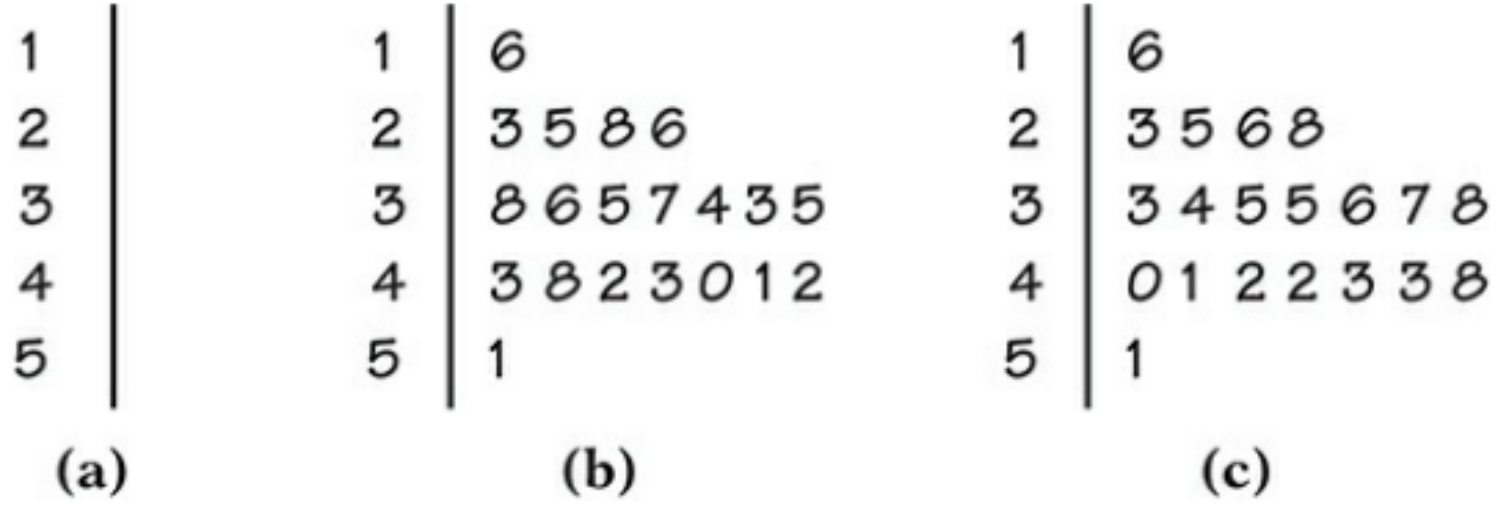
Online resource usage by students as pie chart.



Stem plot

Vitamin D levels

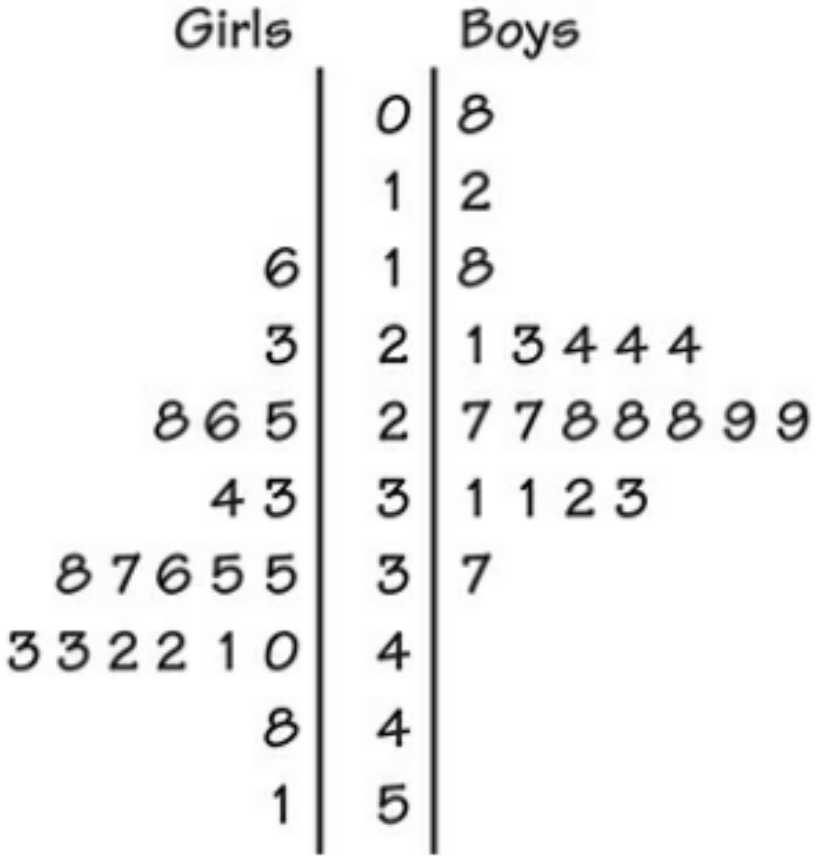
16	43	38	48	42	23	36	35	37	34
25	28	26	43	51	33	40	35	41	42



(a) **(b)** **(c)**

Stem plot - splitting

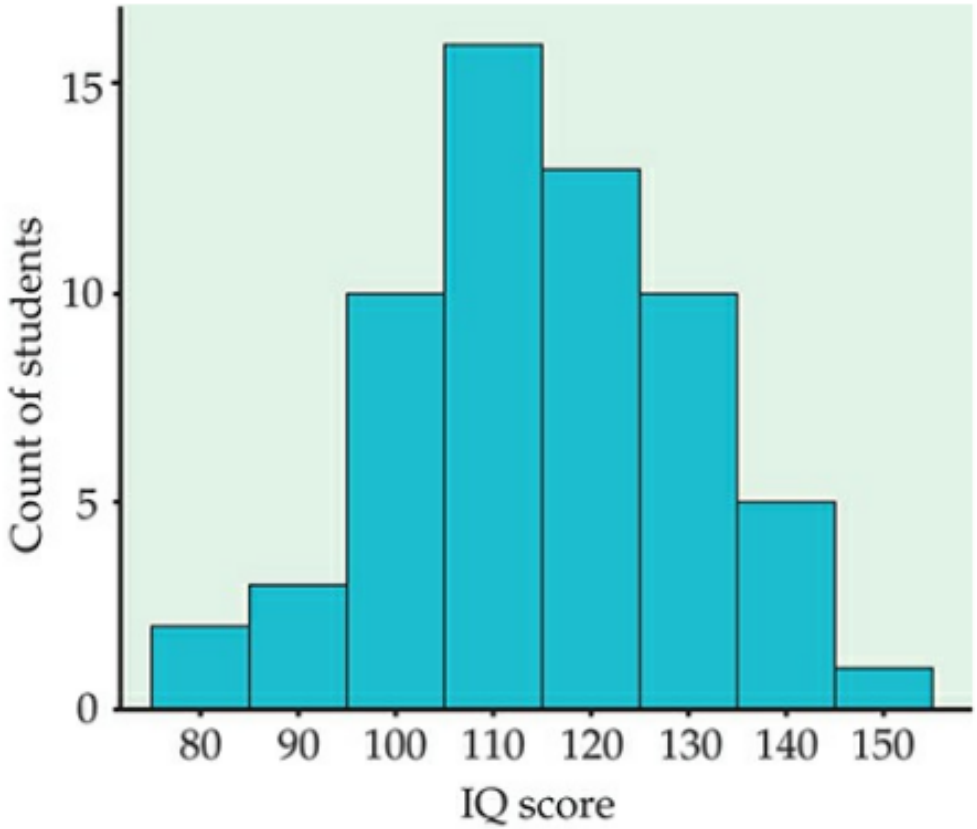
Vitamin D levels - split stem plot



Girls		Boys
	0	8
	1	2
6	1	8
3	2	1 3 4 4 4
8 6 5	2	7 7 8 8 8 9 9
4 3	3	1 1 2 3
8 7 6 5 5	3	7
3 3 2 2 1 0	4	
8	4	
1	5	

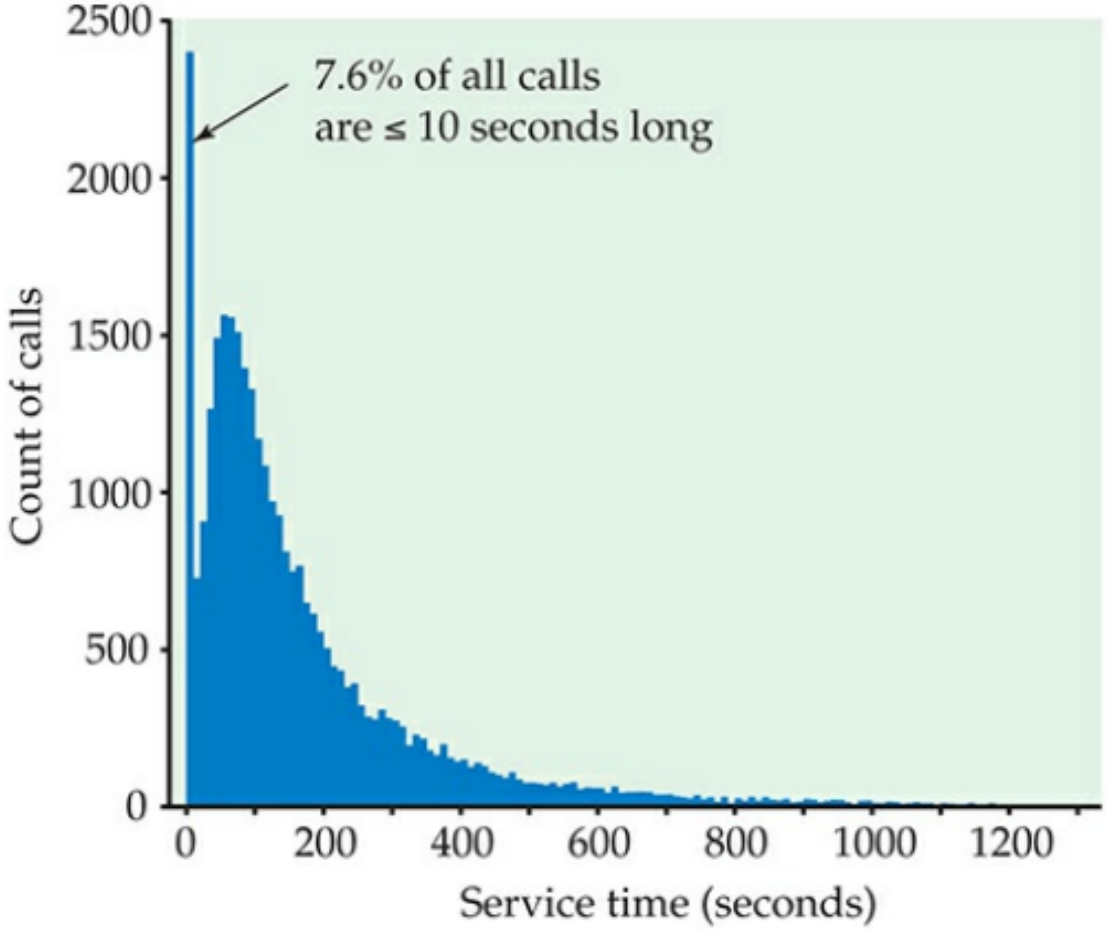
Histogram

IQ data for students



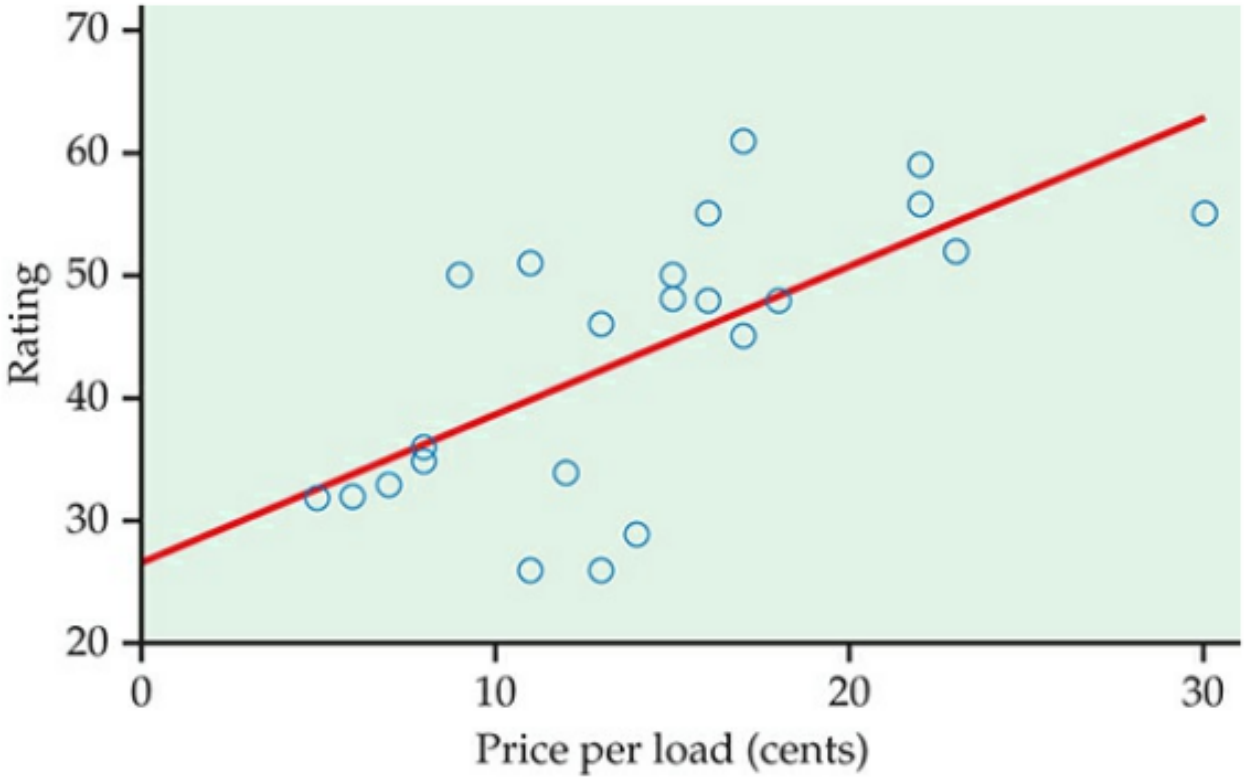
Histogram - outlier

Length of call times at a Bank call centre.



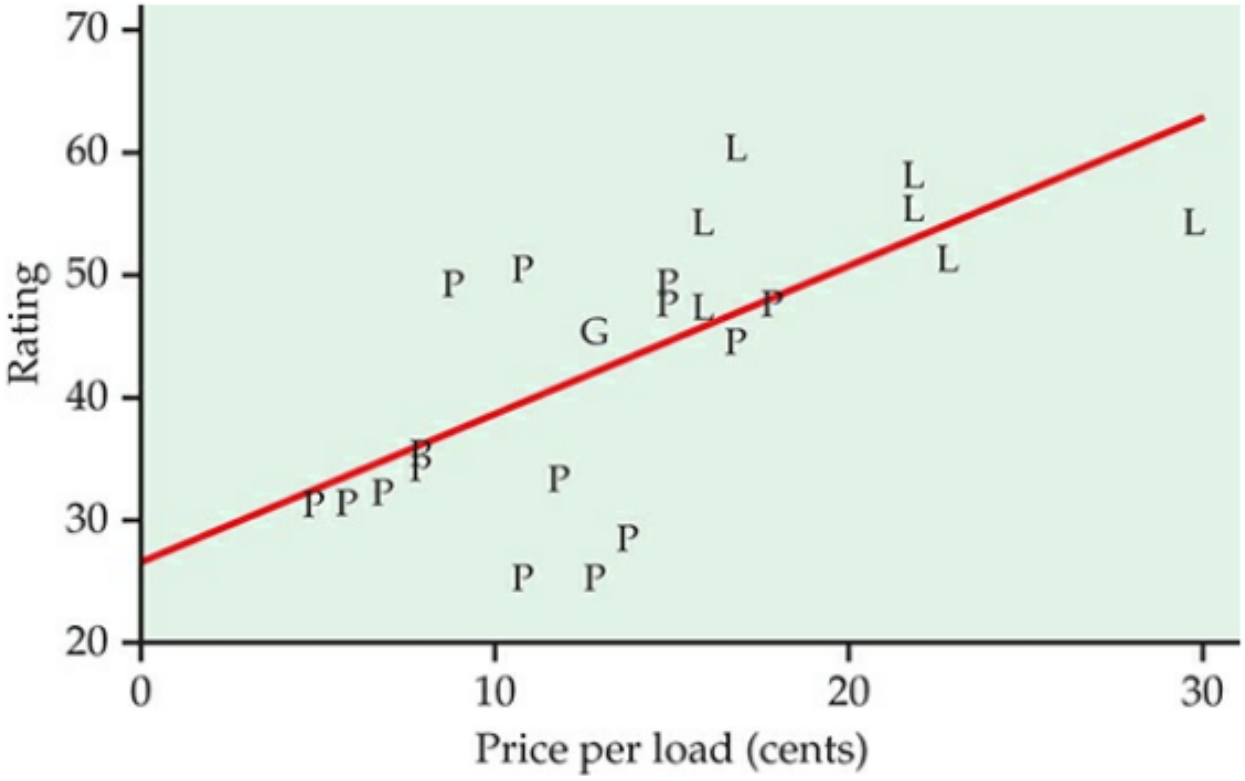
Scatter plot - correlation

Cost per load vs Detergent rating



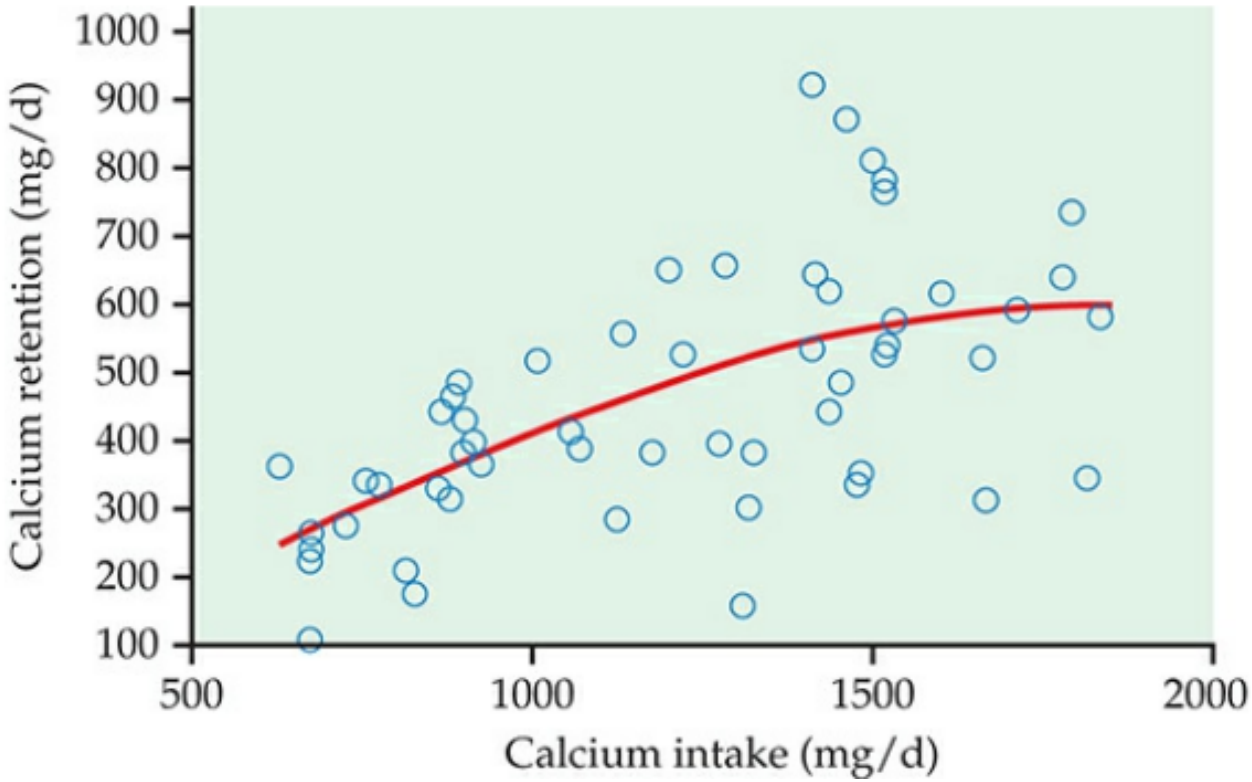
Scatter plot - with category information

Cost per load vs Detergent rating (with Powder, Liquid, Gel form)



Non-linear scatter plot

Calcium given vs Calcium absorbed



Scatter plot after transformation

Calcium given vs Calcium absorbed (log transformation of Y axis)