Chapter 3: Scatterplots & Scaling

Scatterplot

- relates 2 variables x and y.
- shows a large number of x and y pairs at once, in such a way that they can be compared.

• Checking data:
- What is the story behind the data?
- Does the data make sense?
- Is some data missing or obviously incorrect?
- Are there obvious questions that we want to ask?

• Are there outliers?
- Are outliers errors, or are they the most interesting & important cases?

We may have questions about the relationship between x & y:
• Is there a relationship? What type of relationship could there be?
- Is it linear? How accurate is it?
- Is it curved? Can we find a 'law'?
• What are sensible scalings to use?
• If there are different groups of data,
- do they have the same distribution?

Example: Diamonds – Data checking
• Scatterplot: Price vs weight in carat of 54,000 diamonds.
1) What can we say from the plot?
- Both weight & prices are positive, Good.
- Price increases with weight; seems nonlinear; higher variance at higher weights.
2) Try log-log plot?

- Prices appear censored above 20,000.
  - Where have these prices gone?

- No obvious errors.
- Data is heavily overplotted; we cannot see density by eye.
- Given the scatter, the outliers don't seem unreasonable.
  Spot-check them?
- Weights concentrate just at round-numbers (1, 1.5, 2):
  - seems reasonable diamonds are cut to size,
    perhaps these weights are preferred.

- Probably other factors affect price as well
  - since there is so much variation in price at a given weight.
- There are other variables on the data set that need to be investigated.

Try a log-log plot

- If you were starting to work with this data, there would be many questions to ask.
- The log-log plot looks fairly linear, although there is much variation about the trend.
  - Is price proportional to a power of weight (with other factor relevant)?

1. High prices are missing? (Censored?)
2. Prices just above 1 carat seem to go suddenly higher:
   - Is this why stones cut to certain weights?
3. What are these outliers? Large stones of low quality?
4. The bottom edge of the points looks fairly straight, or is it an illusion from overplotting?
5. What is this gap in the prices?
6. Low weights are in discrete values: values are rounded?

Scaling

- Re-scaling: Why?
  - To spread data out, to improve visibility.
    - don't have all data bunched together on a blob in one corner of the plot.
  - Finding a "law" by finding a straight line (or, better, a horizontal line).
  - Plotting data relative to a "reference comparison" (a "law" we don't believe, but which makes a useful point of comparison).
- Before doing any machine learning or linear modelling,
  - rescale input variables to compact distributions

Logarithmic re-scaling

We employ our knowledge of logarithms to simplify plotting the relation between one variable & another.

- Is a variable always positive?
- Are we more interested in percentage changes than in absolute changes?
- Then try log (variable)

Example:

- If $x = 10$ cm, then $y = \log_{10}(x) = \log_{10}(10) = 1$ cm.

- Monetary amounts (incomes, customer value, account or purchase size)
  - are some of the most commonly encountered sources of skewed distributions in data science applications.

- Scaling: lin-log ($y = Ae^{bx}$)

$y = Ae^{bx}$ $\quad\quad \log y = bx + \log A$

$\log y = \log Ae^{bx} = \log A + \log e^{bx} = \log A + bx$.

- If the reference comparison is exponential, try lin-log.
  - Watch out for additive constants.

- $y = e^{nx}$ → Exponential Graphs.

Scaling: log-log

- shows powers as straight lines

$y = Ax^b$ $\quad \log y = b \log x + \log A$

$\log y = \log Ax^b \Rightarrow \log y = \log A + \log x^b \Rightarrow \log A + b \log x$

- very general.
- watch out for additive constants...

Scaling: general principles?

- Identify reference comparison.

Example: Should relationship be linear? $f(y) = g(x)$?

- exponential?
- logarithmic? $y = 1/x$?

- Should one term be a constant?

- Transform data so that reference comparison is as simple as possible.
  - you may need to do this in several steps.

Scaling to a reference comparison

- If it would be 'natural' that $f(y) = g(x)$ for two functions f & g, then
  - plot $f(y)$ Vs $g(x)$ or $f(y) - g(x)$ Vs x,

or - plot $f(y)/g(x)$ Vs x which is more appropriate.

- If it would be 'natural' that $y = g(x) + \text{delta}(x)$
  - for some natural relationship g & some interesting different delta (x)
  - then plot $(y - g(x))$ Vs x.

Example: Consider data on weight & heights of a sample of people. A natural reference comparison is that

$\text{weight} = A\,\text{height}^3$, for some constant A.

so, plot for example, $(\text{weight}/\text{height}^3)$ Vs height.

Which logarithm?

- Logs base 10: $10^{\log_{10}(x)} = x$
  - conventional, can see powers of 10.
- Logs base e: (natural logs): $e^{\log(x)} = x$
  - small changes (e.g. +0.05) are roughly equal to changes by the same percentage (5%)

- Logs base 2: $2^{\log_2(x)} = x$
  - Can be more intuitive than logs base 10: a 'doubling time' is a smaller 'unit' than a time to increase by a factor of 10.

Converting between bases of logs

$x = e^{\ln(x)} = (e^{\ln(2)})^{\ln(x)/\ln(2)} = 2^{\frac{\ln(x)}{\ln(2)}}$

So, $\log_2(x) = \frac{\ln(x)}{\ln(2)}$

Note that: $e = 2.718...$, $\ln(e) = 1$, so $\ln(2) < 1$, so $\log_2(x) > \ln(x)$, as it should be.

Re-scaling: ratios

- Ratios can be tricky.
- Reciprocal:
  - Is your ratio the right way up for your purpose?
  - e.g. relating car weight to miles per gallon?

- If you are fitting a linear predictive model using machine learning,
  - much better to try to predict gallons/mile than miles/gallon.)

- Reciprocals:
  - Consider the price-earnings (PE) ratio
    - Often quoted about companies).
    - This is ratio of the price of one share to the profits (earnings) per share.
  - A company that has low-priced shares & makes good profits
    - has a low PE ratio
  - A company that makes little profit
    - has a high PE ratio.
  - A company that is making a loss
    - has (by convention) no PE ratio.

So, plot earnings/price, not price/earnings

- Reciprocals are not always the answer: a financial example.

1) Market Capitalisation = number of shares X share price
   - this is what the market thinks a company is worth

2) Book Value of a company = Value of what the company owns
   - this is the value of company if it were stopped & everything sold.

- Mkt Cap / Book value is a widely quoted financial ratio
  - high for a start-up company with no assets
  - low for a railway company with lots of land & trains
  - is typically wrong to use: book value can be close to zero.

- Book / MktCap also typically wrong: this too can become high.
- log scaling wrong too: if book value very low, even doubling it doesn't matter.
- In this case try: Book / (MktCap + Book)
  - This is guaranteed to be between 0 & 1.

Re-scaling: Probabilities

- What if we are plotting probabilities:
  - some of them may be very low (e.g. 1/1000, 1/10,000),
  - others close to 1 (0.9, 0.999) etc
- P=1/1000 can be very different from p=1/10,000
  - we may care a lot.
- p=0.99 may be very different from p=0.9999
- In this case, try

$$\log(p/(1-p))$$

Combinations of variables

- Natural "concepts" may be combinations of the variables given.

Example: In "mat cars" dataset, we have
- hp: this is power of engine (energy/second)
- wt: weight of the car
- qsec: seconds to cover a fixed distance (1/4 mile)

One question to ask is whether hp - the rated horsepower
- really corresponds to the energy that
- can be put into accelerating the car.

Let's see:
- kinetic energy $1/2\, mv^2$ is proportional to $wt/qsec^2$
- energy produced by engine
  - should be proportional to $hp * qsec$

- So, if all were equally efficient at converting
  - their rated power (hp) into kinetic energy
  - then $wt/(qsec^3\, hp)$ should be the same for all of them.
  - Is it?
- The reference assumption
  - is that kinetic energy/(hp* time) will be constant
- Aggreement is pretty good - all within a factor of 2.

Another Example: Consider a dataset of people's heights & weights.

- Body mass index (BMI) = weight / height^2 ...

  (Given as $kg/m^2 = 703 * lb/inches^2$).

- BMI should be (roughly) constant for people of different heights & ages. ...

- Plot height / age / any other characteristic Vs BMI
  - to see how people's 'shape' varies.

Summary.

- Plotting data can reveal a story, and raise questions.
- Think carefully about what relationships you should expect between variables.
- Transform variables to give a meaningful value, with a compact range of values.