Why visualisation?

- Goal 1: Have a data set that I want to understand. This is called exploratory data analysis.
- Today's lecture.
- Goal II: Want to display data (i.e., for publication)
  - Will save this for later lecture (if time)
  - Find or display relationships in the data
  - This is a prelude to model building (what is most important to model?)
  - Major goal is inter-ocular impact

Visualisations that we won’t be interested in

For an interesting perspective on this difference, see: Gelman and Unwin. Infovis and statistical graphics: Different goals, different looks (with discussion). Journal of Computational and Graphical Statistics. 2013

[source: Wikipedia]
### Summaries

- **Mean**: 27.7
- **Std Dev**: 9.5

**Sample mean**

$$\bar{x} = \frac{1}{N} \sum_{i} x_i$$

**Sample standard deviation**

$$s_x = \sqrt{\frac{1}{N-1} \sum_{i} (x_i - \bar{x})}$$

- **Min**: 0.00
- **1Q**: 21.7
- **Median**: 28.0
- **3Q**: 33.6
- **Max**: 57.3

### Histograms

Median and quartiles

- **Min**: 0.00
- **1Q**: 21.7
- **Median**: 28.0
- **3Q**: 33.6
- **Max**: 57.3

**Outliers in histograms**

Blood pressure data set

UCI ML repository says no missing data (well, for 20 years it did) [Source: Padhraic Smyth]

**Class-Conditional Histograms**

- **Positive** (diabetes)
- **Negative**

Alternative: Box plot

- **Quartile**
- **Median**
- **Extreme data**

Maybe for only 2 groups, graphs not necessary. For more visual comparisons, can be helpful.
Effect of bin size

More misleading histograms

Data: US Post Codes

[Source: Padhraic Smyth]
Bivariate data

Numerical bivariate summaries

Data are \((x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\)

Sample covariance:
\[
s_{xy} = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \bar{y})(x_i - \bar{x})
\]

Sample correlation:
\[
\rho_{xy} = \frac{s_{xy}}{s_x s_y}
\]

where as before
\[
\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i
\]
\[
\bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i
\]
\[
s_x = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2}
\]
\[
s_y = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (y_i - \bar{y})^2}
\]

Dangers of correlation

Scatterplots

[Anscombe, 1973]
Overplotting

samples from bivariate normal

also: notice the axes!

For our purposes, note:

- Use of colour to add a categorical variable
- Without this colour would not have seen these two outliers
- Use of y=x line to add the eye

96,000 bank loan applicants

[Source: Hand, Manilia, and Smyth]
To fix overplotting, could consider:

- Jittering points
- Subsampling points (i.e., plot only 10%)
- Averaging (if this makes sense)
- Add trend lines (e.g., quantile lines)

**Time Series**

Examples
- Financial data
- Network traffic
- Energy usage
- Human traffic
- Building occupancy

Visualization tricks include:
- Smoothing
  - (running mean, median)
  - Repeated multiples

**Transformations**

Consider powers, logs.
Occasionally reciprocals (e.g., rates).
Also square root

This fit is from loess (local linear regression).
Example Transformation

Why log log here? Hint: Imagine a spherical cow

[Source: William Cleveland, Visualizing Data]

Three-Dimensional Data

- Generally hard
- 3-D plots are not usually useful
- Usually better to use colour on a 2-D plot
- Or show multiple 2D plots for each value of third variable

High-Dimensional Data

Wait, what if you have categorical data?

Tools here include:

- Colour
- Contingency tables
- Multiple plots (e.g., class-conditional histograms)

Two main options:

- Project the data down to 2-D
  - Many techniques
    - Principal Components Analysis (IAML, MLPR)
    - Multidimensional scaling
    - Modern nonlinear methods: t-SNE, LLE, Isomap, Eigenmaps
  - Problem: Sometimes this will obscure high-D structure and nonlinear structure
- Another option: Scatterplot matrix (see next)
What are you looking for?

- Anomalies. If something looks weird, figure out why. It could be an error in your data.
- Learn from your data but do not trust it! (Not completely.)
- Relationships. Hypothesis-based visualization. What relationships do you expect to exist? Can you see them?
- Use visualization to inform models and vice versa
  - e.g., Can help with feature construction, e.g., whether a relationship is "really" nonlinear
- Fancy 3D graphs … meh
- These techniques also useful for the outputs of learning!

If you really like this stuff

- Tukey, Exploratory Data Analysis
- Bill Cleveland, Visualizing Data
- Edward Tufte, all books