AIML430/COMP309: ML Tools and Techniques
Lecture 8: Exploratory Data Analysis

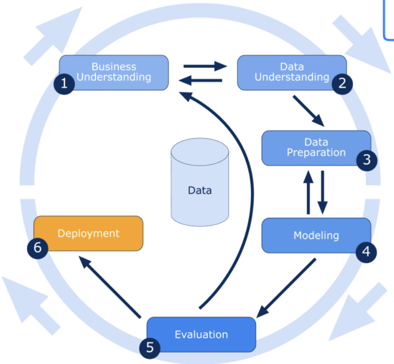
Ali Knott
School of Engineering and Computer Science, VUW

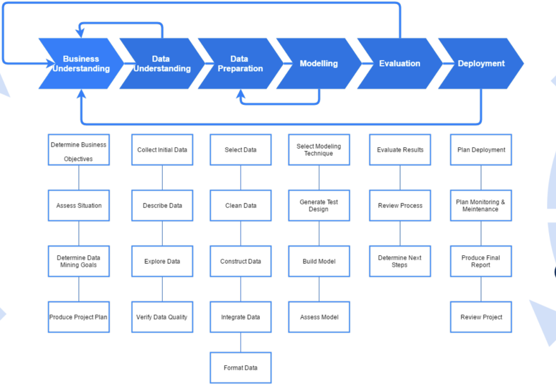
What we're doing this week

We'll introduce a second data analytics paradigm, called Exploratory Data Analysis (EDA).

- Today: an introduction to EDA.
  - With a focus on some new visualisation methods.
- Tomorrow: a survey of Python tools for EDA.
  - With some Python background we skipped in Weeks 1-2...
  - With some very useful Python tricks...
- Thursday: EDA demos in Python and Orange.

But first—what's EDA got to do with CRISP-DM??





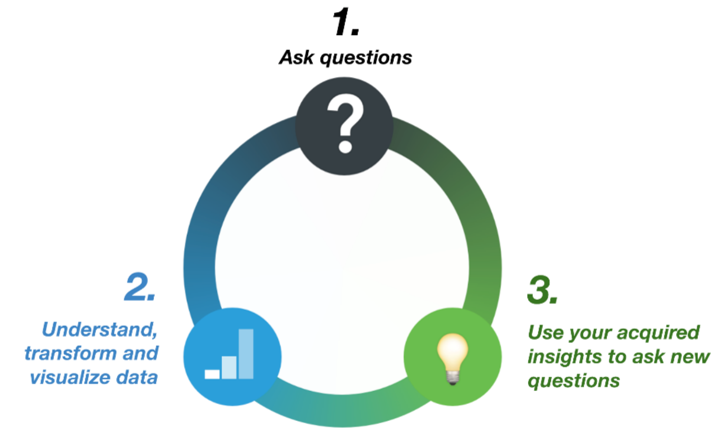
It kind of overlaps with these bits...

- But it's really just a method of its own—with several new techniques.

EDA

EDA is structured as a cycle, like CRISP-DM.

- But the cycle here is all about *understanding and manipulating your data*.



Step 2: 'Understand, transform, visualise'...

We'll focus on *Step 2* in this lecture.

- This step involves a collection of useful data analysis *methods*.

There's some useful structure here—

EDA methods are either graphical or non-graphical.

- We'll look at lots of examples of each.

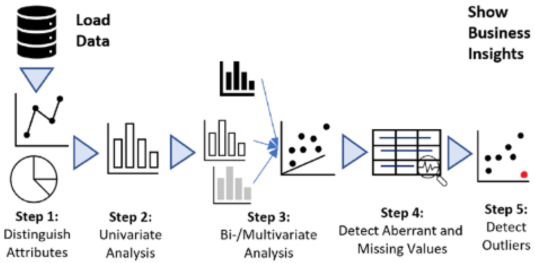
EDA Methods are either univariate or multivariate.

- Univariate methods analyse just *one* feature...
- Multivariate methods analyse relations between *several* features.
  - Bivariate methods analyse exactly two features.

EDA data analysis methods

	Univariate	Multivariate
Non-Graphical	**Categorical Variable:** tabular representation of frequency **Quantitative Variable:** • Location (mean, median) • Shape and Spread • Modality • Outliers …	**One Categorical Variable and One Quantitative Variable:** Standard univariable non-graphical statistics for the quantitative variable separately for each level of the categorial variable **Two and more Quantitative Variable:** • Correlation, • Covariance, • …
Graphical	**Categorical Variable:** Bar Chart **Quantitative Variable:** • Histogram • Boxplot • …	**One Categorical Variable and One Quantitative Variable:** • Side-by-side Boxplots **Two and more Categorical Variable:** • Grouped Bar Chart **Two and more Quantitative Variable:** • Scatter plot, Correlation Heatmap, Pairplot **…**

There's a standard order for EDA methods...



- We'll go through these one by one...
- Note that EDA still connects with 'business-oriented' analytics!

Step 1: ‘Distinguish attributes’. . .

‘Attributes’ are just the features in your dataset.

- The analysis here is just to identify the different *types* of feature.

Step 1: Possible feature types

Categorical features place items into *discrete categories*.

- Sometimes there's an ordering to categories.
  - E.g. 'big', 'medium', 'small'
- Sometimes categories are unordered.
  - E.g. color of a car (blue, red, green).

Numerical features place items on numerical scales.

- Sometimes these are whole integers (e.g. number of children)
- Sometimes they can be any 'real number'.
  - E.g. length
  - E.g. price.

Step 2: Univariate analyses (non-graphical)

For a categorical variable, we can create a table, showing frequency of different possible values.

For a quantitative variable, there are a few things to compute.

- The shape of its distribution.
  - Is it normally distributed, or something else?
  - We'll discuss skew and kurtosis below…
- The modality of its distribution.
  - Unimodal (one bump)?
  - Bimodal (two bumps)?
  - Multimodal (more than one bump)?
- If it's normal, we can compute its mean and standard deviation.
- If not, the median and/or mode may be useful.

Step 2: Univariate analyses (graphical)

Background: there are several standard ways of visualising data.

- Some can be used for either univariate or multivariate analyses.
- Some aren't suitable for multivariate, because they can't show *relationships* between variables very well.

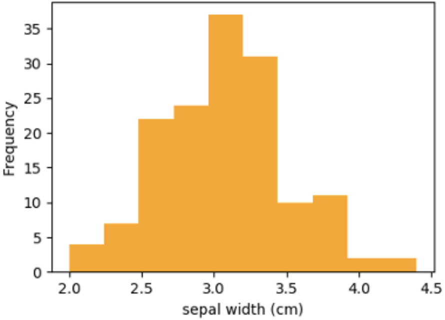
Background: visualisations can be used to show four different things:

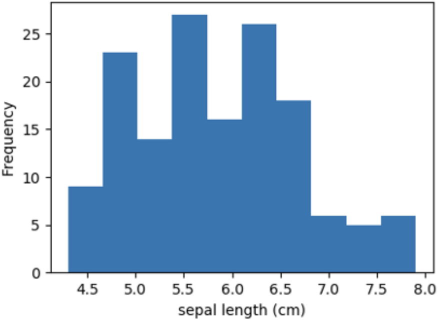
- Distribution (of a single variable, or of multiple variables)
- Comparison (between variables, of of one variable over time)
- Composition (to show the *parts* that make up a whole)
- Relationships (between variables).

Visualising distribution of a quantitative variable

Histograms are a standard way to do this.

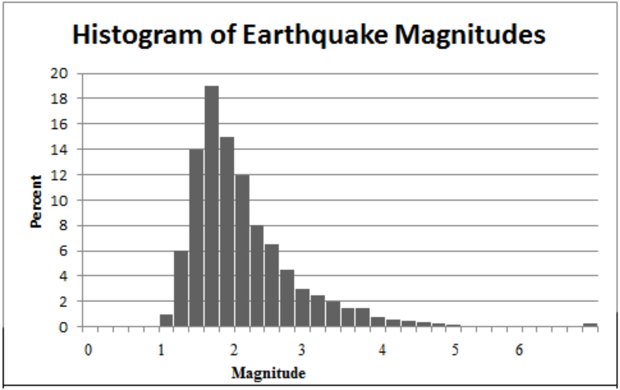
- Break variable values into 'bins' of equal ranges...
- Then plot a bar chart showing frequency for each bin.





The ‘shape’ of a quantitative variable distribution

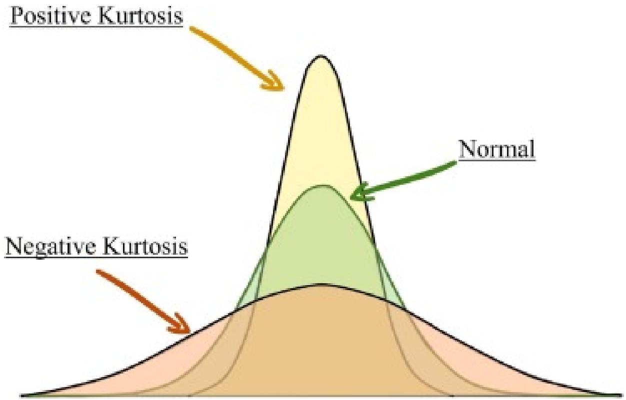
Distributions can be skewed, either left or right. . .



The ‘shape’ of a quantitative variable distribution

Distributions can also be ‘sharper’ or ‘flatter’ than a normal distribution.

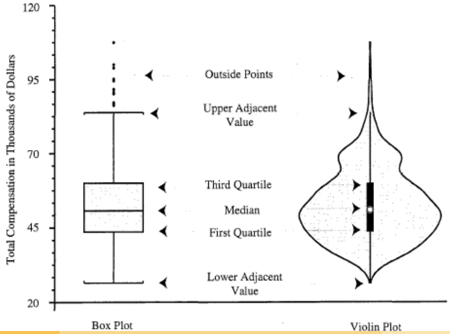
- We use the term kurtosis to describe this.



Box and violin plots

A box plot visualises some values from 'nonparametric' statistics.

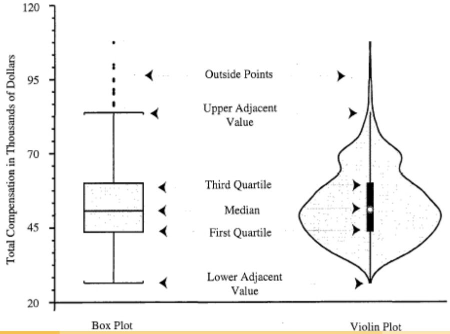
- We *rank* the data, and split into 4 'quartiles', holding 25% each.
- The 'box' shows the range for the two middle quartiles.
  - Also called the 'interquartile range' ('IQR').
- The median value (the middle one in the ranked list) is also shown.
- The 'whiskers' extend to the 1.5 IQR values in each direction.



Box and violin plots

A violin plot shows a function that *models* the distribution of the data.

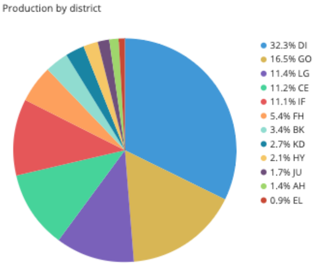
- The model is a kernel density estimation—you don't need to know the details!
- A box-and-whisker plot is often overlaid on a violin plot.

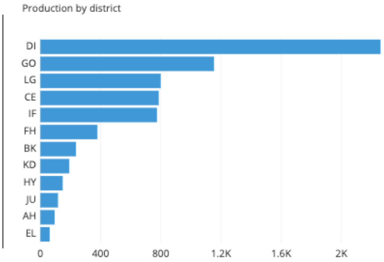


Visualising distribution of a categorical variable

Here, we're mainly interested in what proportion of items fall into each class.

- Pie charts are one possibility.
- Bar charts are another.



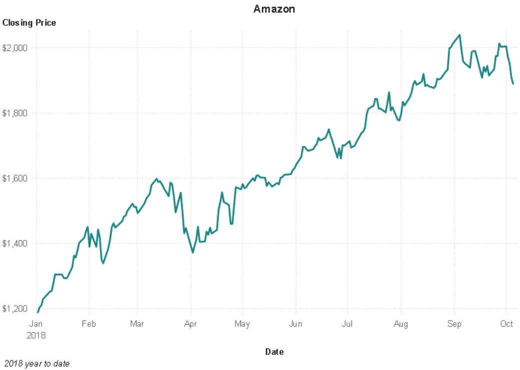


Bar charts are better for showing small differences between categories.

Visualising a single variable over time: Line charts

Sometimes a dataset includes one variable at a succession of times.

- That's several different variables, but with a particular relationship!
- We can use a line chart to show how a variable changes over time.



NB we use lines for *numerical* variables, and bars for *categorical* ones!

Step 3: Multivariate analyses

If you have a categorical variable and some other variable V, you can do a univariate analys of V *for each category*.

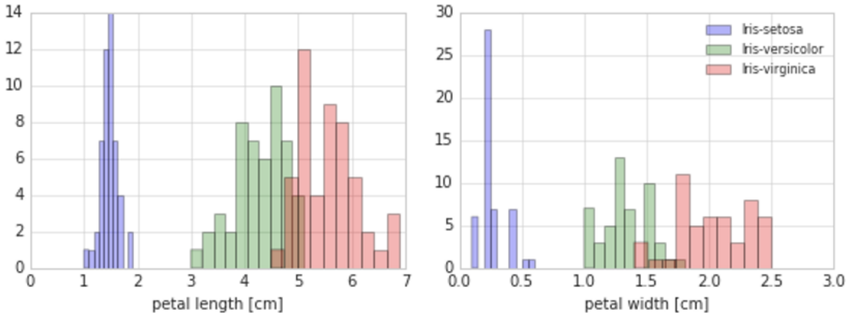
- That includes visual and nonvisual analyses.

Step 3: Multivariate analyses

If you have a categorical variable and some other variable V, you can do a univariate analys of V *for each category*.

- That includes visual and nonvisual analyses.

If V is quantitative, you could show *overlaid histograms*:

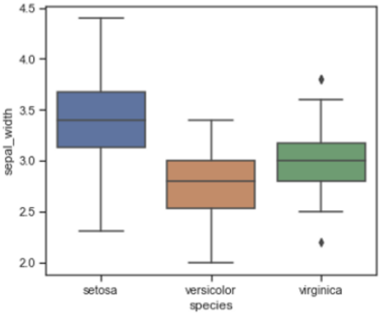


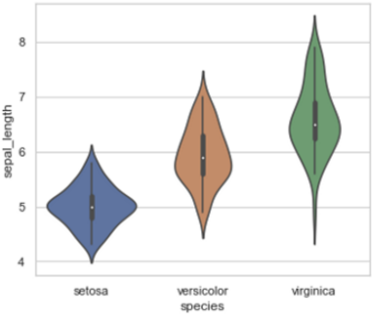
Step 3: Multivariate analyses

If you have a categorical variable and some other variable V, you can do a univariate analys of V *for each category*.

- That includes visual and nonvisual analyses.

If V is quantitative, you could show *multiple box or violin plots*:





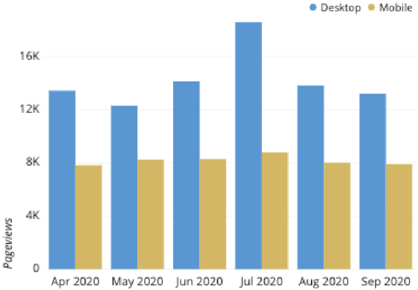
Step 3: Multivariate analyses

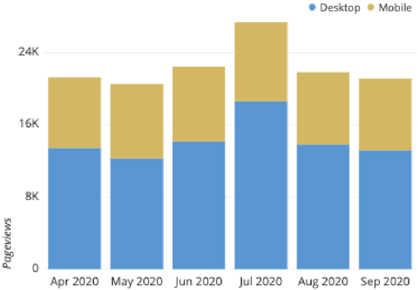
If you have a categorical variable and some other variable V, you can do a univariate analys of V *for each category*.

- That includes visual and nonvisual analyses.

If V is categorical, you could show a *grouped bar chart*.

- If your features measure components of some whole, you can use a *stacked bar chart*.



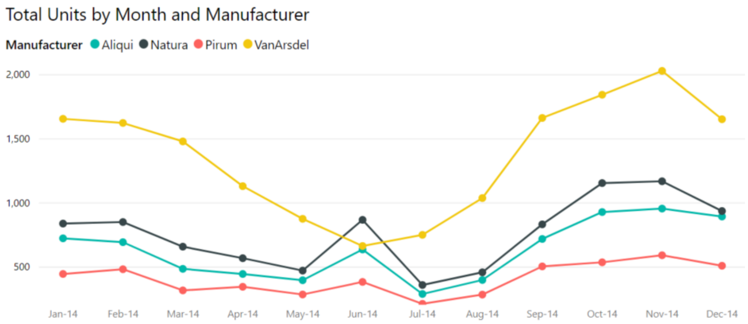


Step 3: Multivariate analyses

If you have a categorical variable and some other variable V, you can do a univariate analys of V *for each category*.

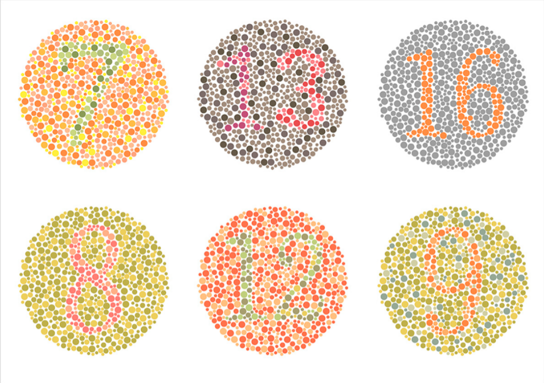
- That includes visual and nonvisual analyses.

For multiple time series, you can show a line graph with several lines:



Aside: Make your visualisations *accessible*!

Some of your viewers will have colourblindness, of various different types.



You can find some good advice on colourblind-friendly palettes here.

- Simple advice: don't use red & green togetheror yellow & blue!

Non-visual multivariate analyses

If you have a set of quantitative variables, two key nonvisual measures are correlation and covariance.

- These both operate on *pairs* of variables.
- But if you have n dimensions, you can build a *matrix* of values, running correlation/covariance on *each pair*.

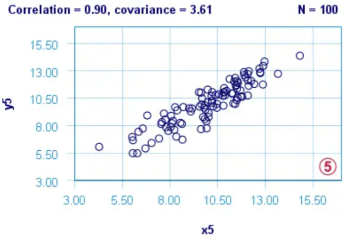
Remember:

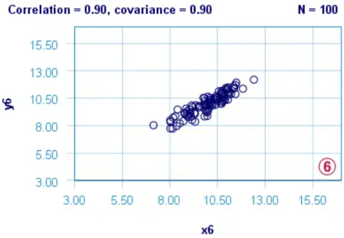
- PCA works by computing the covariance matrix for the data...
- Regression analyses don't work well if input variables are highly correlated!
  - We called that 'the problem of collinearity'...
- If you have collinear inputs, PCA is a good idea!

Correlation and covariance analyses

Correlation and covariance both measure how well the relationship between two variables can be modelled with a straight line. (And the slope of the best line.)

- Covariance provides an absolute number, that grows as points spread...
- Correlation (Pearson's correlation coefficient) *normalises* covariance by the spread of the data, so it lies between 1 and -1.





Some statistical multivariate analyses

You can also run an analysis of variance (ANOVA), to identify *effects* of input variables on a quantitative output variable.

- This identifies 'main effects', due to individual variables...
- And cases where input variables 'interact' to influence the output.

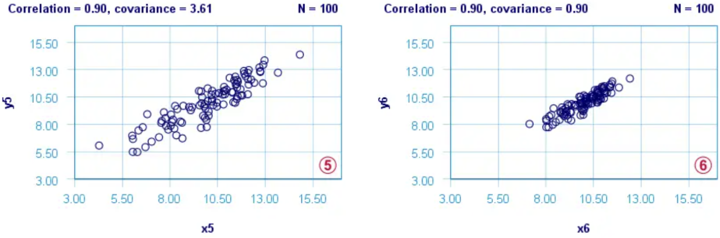
A t-test can test a *hypothesis* that some manipulation influences some output variable.

- It tells you how likely the output variable values were obtained 'by chance'.

Graphical multivariate analyses

For two or three dimensions, we can use a scatter plot.

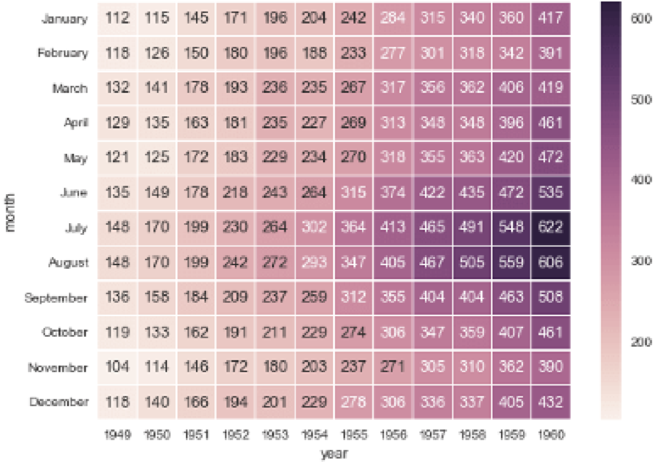
- We just saw some examples!



NB we can use *dimension reduction* to get our data into 2 or 3 dimensions, to help visualise like this.

Heat maps for 2D multivariate visualisations

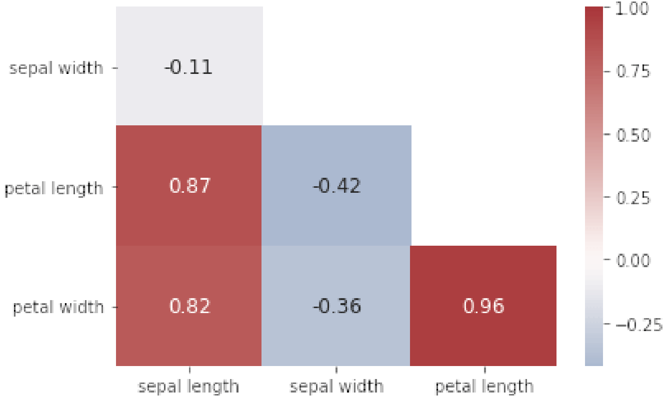
A heat map can be used to show the value of one numerical variable for all values of two other variables. (Normally numerical or ordinal.)



Correlograms

A heat map can also be used to show the correlation coefficient between each pair of numerical features. (Called a correlogram.)

- The map is symmetrical about its leading diagonal, so we only need to show half…



EDA Steps 4 and 5: Data preprocessing

Some concrete methods:

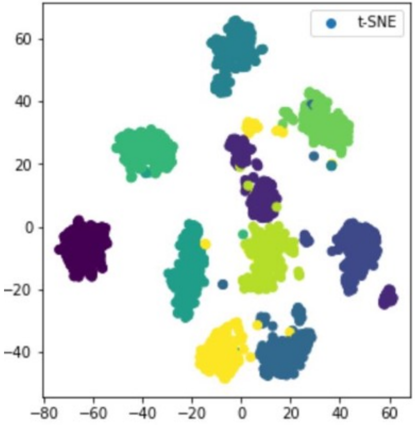
EDA	*Data Preprocessing*
Duplicate Data-Points?	Delete Duplicate Data-Points
Missing Values?	Imputation
Outlier?	Exclude outliers
Highly Correlated Features?	Handling Highly Correlated Features
Low-Variance Features?	Handling Low-Variance Features
Imbalanced Data?	Oversampling, Undersampling
Features vary in their scale?	Feature Scaling
High-dimensional data?	Dimensionality Reduction, Feature selection

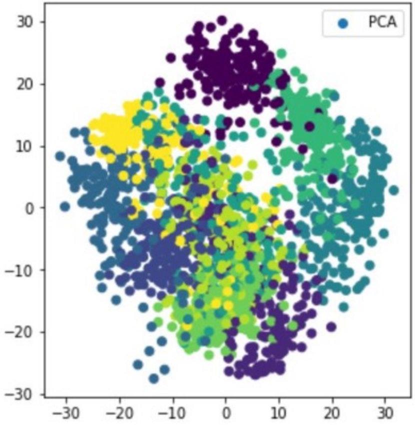
We'll talk more about these in tomorrow's lecture.

Dimensionality reduction

I told you about PCA—but it's also worth knowing about the T-SNE (T-distributed stochastic neighbour embedding) method.

- T-SNE often works better.





Next lecture. . .

A survey of Python tools for EDA.

- With some Python background we skipped in Weeks 1-2. . .
- And some very useful Python tricks!