

Data Exploration & Visualisation

John Pinney
November 2019

Exploratory Data Analysis

- Distributions

- Outliers, errors, missing data

- Variation and covariation

- Asking questions about data

- Clustering

Exploratory Data Analysis

What is exploratory data analysis?

- An approach to data analysis that focuses on summarising the main characteristics of data.
- Often employs visualisation methods.
- Can be used to help formulate hypotheses.
- May reveal unexpected patterns in the data.

Example data set

A 1987 study of different types of glass for criminological investigation.[1]

Types of glass

- 1 building windows (float processed)
- 2 building windows (non-float processed)
- 3 vehicle windows (float processed)
- 4 vehicle windows (non-float processed)
- 5 containers
- 6 tableware
- 7 headlamps

The [float process](#) for making very flat glass sheets was invented in the 1950s

Problems with Spreadsheets...

Fragile analysis: easy to introduce errors and hard to detect them.

e.g. Excel has long been known to mangle gene names! [2]

Analysis done with spreadsheets can also be **difficult to reproduce**.

Orange is one approach to making data analysis more robust and reproducible.

<https://orange.biolab.si/>

The application is based on a **visual programming** paradigm: you construct a **workflow** by chaining together different **widgets**.

Orange contains a wide range of widgets for

- data handling
- visualisation
- machine learning

Widget outputs can be assembled into [reports](#) and exported to PDF or HTML.

CSV (comma separated values) files are tables where columns are delimited by ,

A header row gives the name for each column, e.g.

```
type,RI,Na,Mg,Al,Si,K,Ca,Ba,Fe
3,1.51655,13.41,3.39,1.28,72.64,0.52,8.65,0,0
2,1.51851,13.2,3.63,1.07,72.83,0.57,8.41,0.09,0.17
1,1.51742,13.27,3.62,1.24,73.08,0.55,8.07,0,0
1,1.52213,14.21,3.82,0.47,71.77,0.11,9.57,0,0
2,1.53125,10.73,0,2.1,69.81,0.58,13.3,3.15,0.28
```

Task

Export each lab's data from Excel to a separate CSV file.

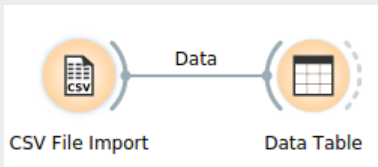
Importing CSV

Task

Use a **CSV File Import** widget to load the data for Lab 1.

Make sure that the **type** column is treated as a categorical variable.

View the data as a **Data Table**.



Task

Use a **Distributions** widget to look at histograms for each column.

Which variables appear to be normally distributed?

What happens when the data are split by **type**?

Combining data sets

Task

Use two more **CSV File Imports** to load the data for Labs 2 and 3.

Use a **Concatenate** widget to combine data from all three labs into one data set.

Are there any differences in the data from each lab?

Use the *Report* icon to make a note of anything interesting.

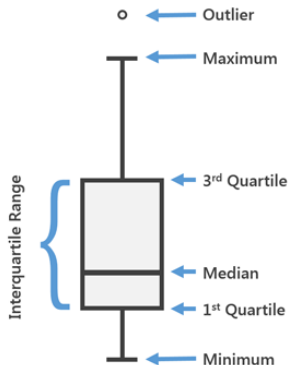
Dealing with errors

Task

Insert a **Select Rows** widget to take care of suspicious outliers.

Insert a **Select Columns** widget to eliminate any columns that appear to contain systematic errors.

Box plots



Task

Use the **Box Plot** widget to compare variable distributions between **types**.

Which variables have significantly different means between **types**?

Scatter plots

Task

Add a **Scatter Plot** widget to look at **covariation** between variables.

Use *Find Informative Projections* to find the pairs of variables that maximise separation between **types**.

Use a **Correlations** widget to suggest other interesting projections of the data.

Predicting RI from the other features?

We may identify continuous-valued features (e.g. **RI**) that we would like to predict from the other features.

This is a **regression modelling** task, which is a type of machine learning.

Orange provides widgets to train and test linear regression models, but this is beyond the scope of today's workshop.

Imputation

Scatter plots based on features can be informative by themselves. However, to explore patterns in high-dimensional data, we usually need to find projections onto axes that are *linear combinations* of features.

To do this, we want every data point to have a value for every feature. Missing data will cause problems for most analysis methods.

Instead of dropping data points, one solution is to **impute** values where they are missing in the table.

Imputation

There are a variety of ways to impute missing data.

The simplest approach is to insert the mean value of the variable - this should ensure that the imputed value does not bias any downstream analysis.

However, imputation will change the distribution of data, so we should be careful about drawing strong conclusions from a data set containing imputed values.

Task

Add an **Impute** widget to deal with missing values.

Principal Component Analysis finds a set of orthogonal directions in the (normalised) data space that maximise variance.

This is an *unsupervised* analysis, as it does not depend on the data labels.

Sometimes PCA can be a helpful part of exploratory data analysis, but it is not the only way to look for patterns in the data.

PCA is *not* a clustering method.

Task

Add a **PCA** widget to calculate the first three principal components.

You will need to insert a **Select Columns** widget to set **type** as the *target* variable.

Look at the transformed data and the weights for each component.

FreeViz is a *supervised* approach to finding an informative 2D projection of a high-dimensional data set [3].

It adjusts the weights for each feature so as to maximise the separation between the classes of the target variable. You can also move the feature vectors manually.

FreeViz can be an intuitive way to find a subset of features that explain the differences between classes, i.e. to perform **feature selection**.

Task

Use a **FreeViz** widget to find a good projection for separating types.

You should find that type 2 glass appears quite heterogeneous. Take it out of the data set temporarily by inserting a **Select Rows** widget.

Now use FreeViz to find features that can separate

- float from non-float glass
- headlamps from other non-float glass
- tableware from containers

Predicting type from the other features?

We have identified a few simple rules for distinguishing between some of the **types**.

In machine learning, predicting a categorical variable from a set of features is called **classification**. This is a *supervised* task.

Assuming there is sufficient information in the data provided, a good classification method will find ways to distinguish classes reliably, even when they are not linearly separable in the original feature space.

Orange provides widgets to train and test a variety of classification models, but this is beyond the scope of today's workshop.

k-means clustering

Clustering is an *unsupervised* task that tries to divide the dataset into subgroups that contain similar data points. It can be a useful element of exploratory data analysis.

k-Means clustering operates directly on the feature space. This means that we do not need to compute distances between data points.

k-means is a *heuristic* method, which means that it may not always produce the same result.

k-means clustering

Task

Use a **k-Means** widget to find clusters in the data set.

How many clusters appear to be present?

Which features can be used to separate clusters?

How do the clusters correspond to **type**?

Your turn

Explore your own data

Task

Apply these exploratory data analysis techniques to your own tabular data set and prepare a report on your findings.

No suitable data of your own? The **Datasets** widget has lots of examples to play with.

Choose a data set that

- has at least 5 variables.
- has more instances than variables.
- is not tagged *synthetic* or *image analytics*.

Tips

Columns must correspond to variables and need variable names as a header row.




Categorical variables need to be identified during CSV import.

Use **Data Table** and **Distributions** for an initial sanity check.

Use **Select Rows** to work with a subset of the data.

Use **Select Columns** to specify the target variable.

Use **Impute** to fill in missing data if needed.

-  <https://archive.ics.uci.edu/ml/datasets/Glass+Identification>
-  Ziemann M, Eren Y El-Osta A. Gene name errors are widespread in the scientific literature. *Genome Biol* 17:177 (2016)
-  Demšar J, Leban G Zupan B. FreeViz—An intelligent multivariate visualization approach to explorative analysis of biomedical data. *J Biomed Inform* 40:661-671 (2007)