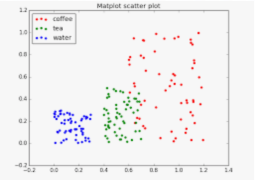
COMP20008
Elements of data processing

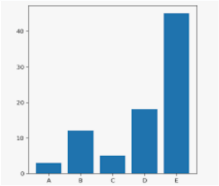
Semester 1 2020

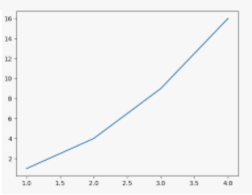
Lecture 8: Advanced data visualisation

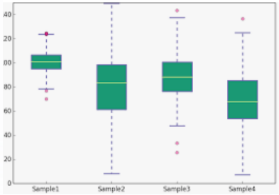
Simple plots for data visualization

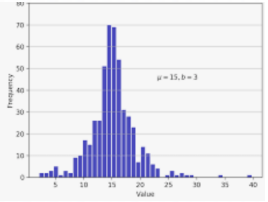


- Basic plots for one- or two- dimensional data
  - Boxplots
  - Histograms
  - Bar plots
  - Scatter plots
  - Line plots









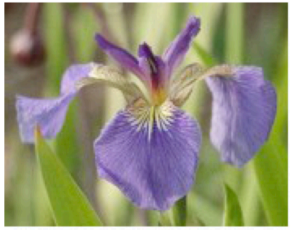
- How to reveal the relationship among several (more than two) dimensions simultaneously?

Example: Iris Flower dataset

- Well known dataset introduced by statistician Ronald Fisher with 150 objects

  https://en.wikipedia.org/wiki/Iris_flower_data_set

- Three flower types (classes):
  - Setosa
  - Virginica
  - Versicolour
- Four features
  - Sepal width and length
  - Petal width and length



Iris setosa

Iris versicolor

Iris virginica

https://www.palass.org/sites/default/files/media/palaeomath_101/article_10/article_10.pdf

Iris dataset

- Extract of Iris data from Wikipedia

Fisher's *Iris* Data

Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	*I. setosa*
4.9	3.0	1.4	0.2	*I. setosa*
4.7	3.2	1.3	0.2	*I. setosa*
4.6	3.1	1.5	0.2	*I. setosa*
5.0	3.6	1.4	0.2	*I. setosa*
5.4	3.9	1.7	0.4	*I. setosa*
4.6	3.4	1.4	0.3	*I. setosa*
5.0	3.4	1.5	0.2	*I. setosa*

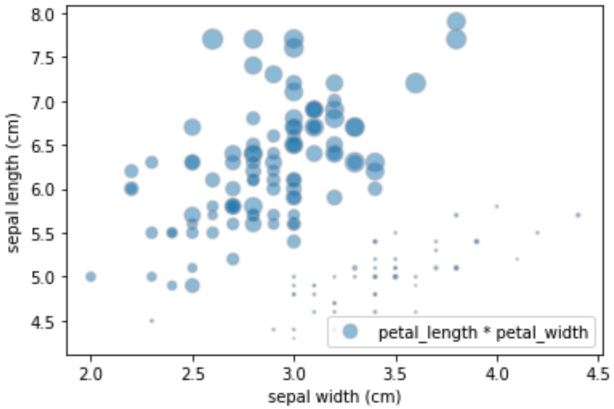
Bubble plots

- A special scatter plot representing 3-dimensional data
- Size of circle around a point indicates the value of the 3rd dimension.

Bubble plots

```
plt.scatter(iris.sepal_width, iris.sepal_length, \
            s = iris.petal_length * iris.petal_width*10, \
            alpha = 0.5, linewidth = 1, edgecolor='darkgrey', \
            label = 'petal_length * petal_width')
plt.xlabel('sepal width (cm)')
plt.ylabel('sepal length (cm)')
plt.legend()
```

```
<matplotlib.legend.Legend at 0x7f9befbea410>
```



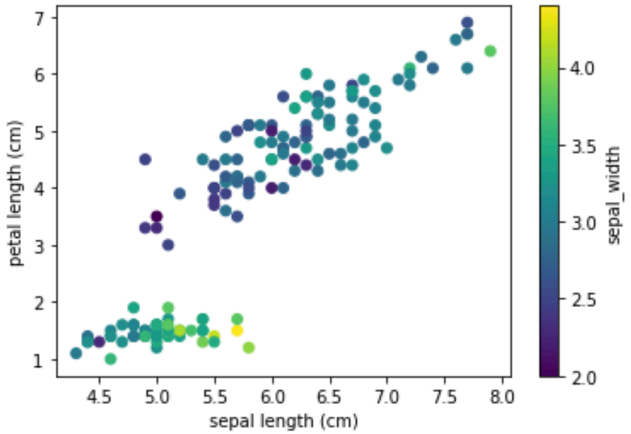
Enhanced scatter plots

- The 2D scatter plot can be enhanced to illustrate relationships among three dimensions.

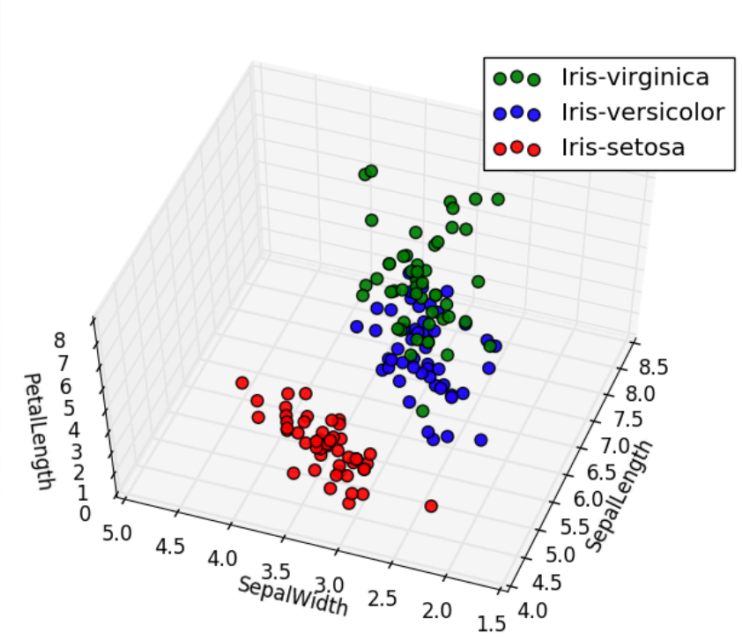
- Use colours for the values of the 3rd dimension.

Enhanced scatter plots

```
plt.scatter(iris.sepal_length, iris.petal_length, c = iris.sepal_width)
plt.xlabel('sepal length (cm)')
plt.ylabel('petal length (cm)')
cbr = plt.colorbar()
cbr.set_label('sepal_width (cm)')
```



3D scatter plots



https://github.com/camila-ud/3D-Scatter-plot

3D scatter plots

Python

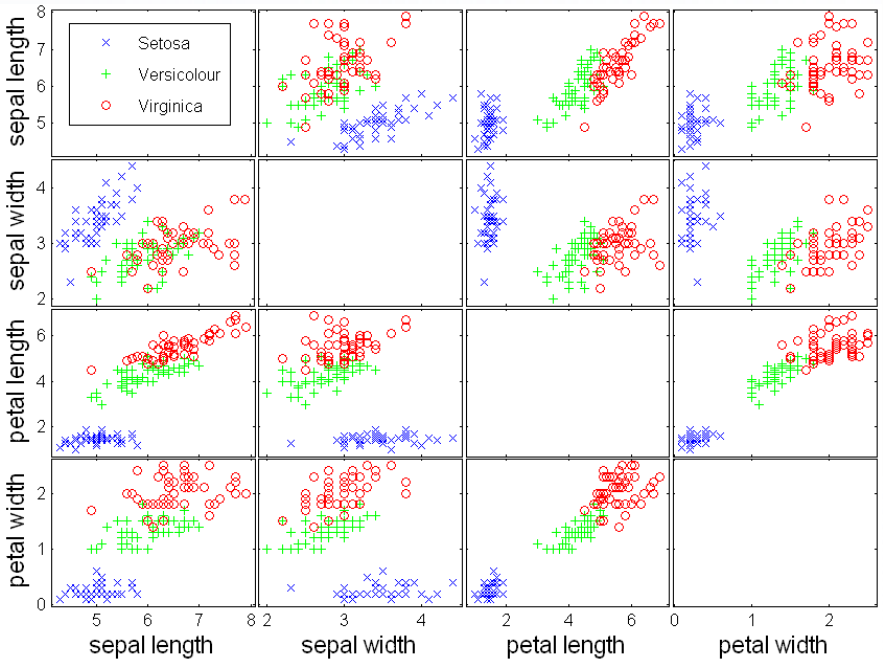
```
from mpl_toolkits import mplot3d
import matplotlib.pyplot as plt

fig = plt.figure()
ax = Axes3D(fig)
ax.scatter(x, y, z)
```

Scatterplot matrix

- A matrix of scatter plots of all pairs of dimensions (variables)
- Inspect many relationships simultaneously.
- Convenient for spotting linear correlation between variables
- Spotting outliers

Scatterplot matrix



Scatterplot matrix

Python

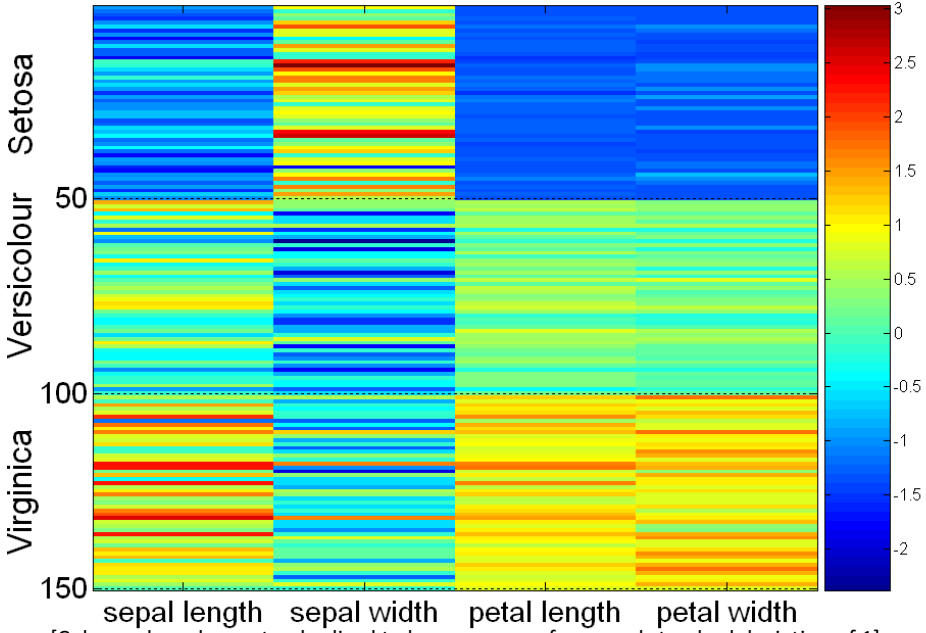
```
import seaborn as sns
sns.pairplot(…)
```

```
from pandas.plotting import scatter_matrix
scatter_matrix(…)
```

Heat maps

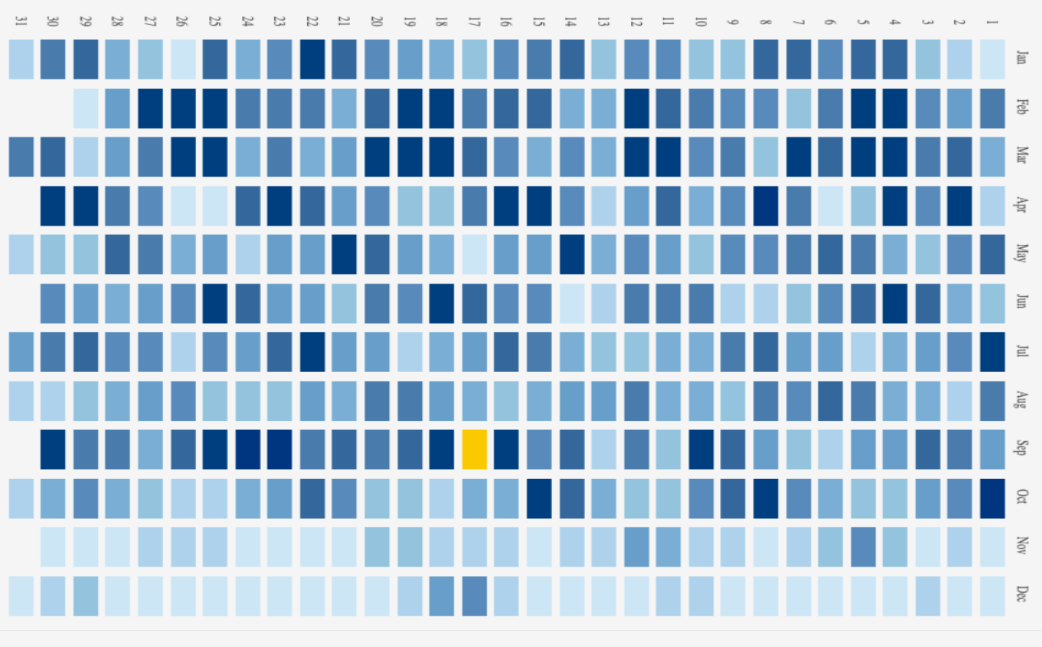
- Plot the data matrix
- Individual values contained in a matrix are represented as colours
- This can be useful when objects are sorted according to class/type
- Typically, features are normalised to prevent one attribute from dominating the plot

Heat map – (normalised) Iris data



[Columns have been standardized to have a mean of zero and standard deviation of 1]

How common is your birthday?



https://www.abc.net.au/news/2017-12-13/australias-most-and-least-popular-birthdays-revealed/9241978

Heat maps

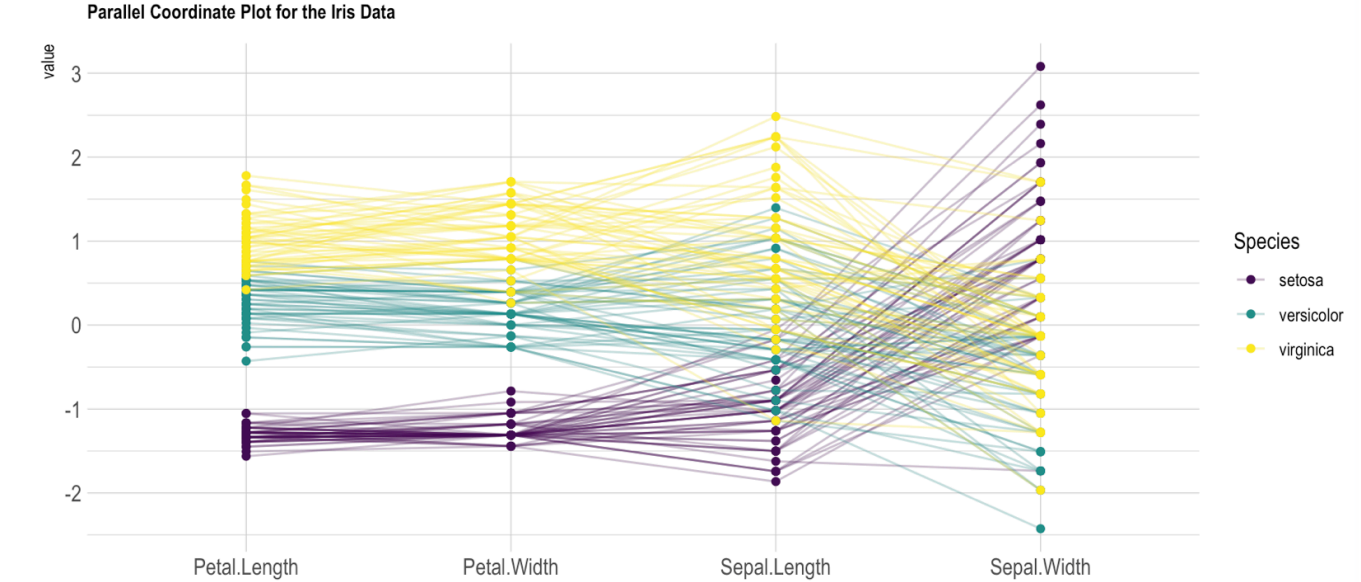
Python

```
import seaborn as sns
sns.heatmap(…)
```

Parallel Coordinates

- A widely used visualisation technique for exploring large, multi-dimensional data sets
- Use a set of parallel axes (coordinate axes)
- The values of each data object are plotted as a point on each corresponding coordinate axis and the points are connected by a line.
- Thus, each data object is represented as a line

Parallel coordinates – Iris data



https://www.data-to-viz.com/graph/parallel.html

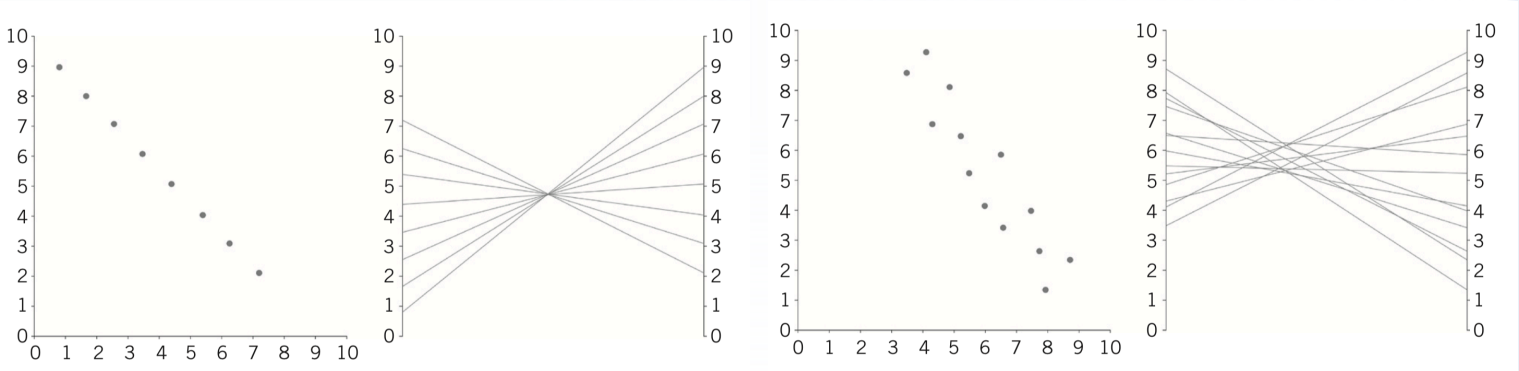
Patterns in parallel coordinates

- Reveal a distinct class of object group
  - Often, the lines representing a distinct class of objects group together, at least for some features
  - Ordering of attributes is important in seeing such groupings.

- Show data characteristics such as different data distributions
  - E.g. petal length has low variance for setosa species

Patterns – cont.

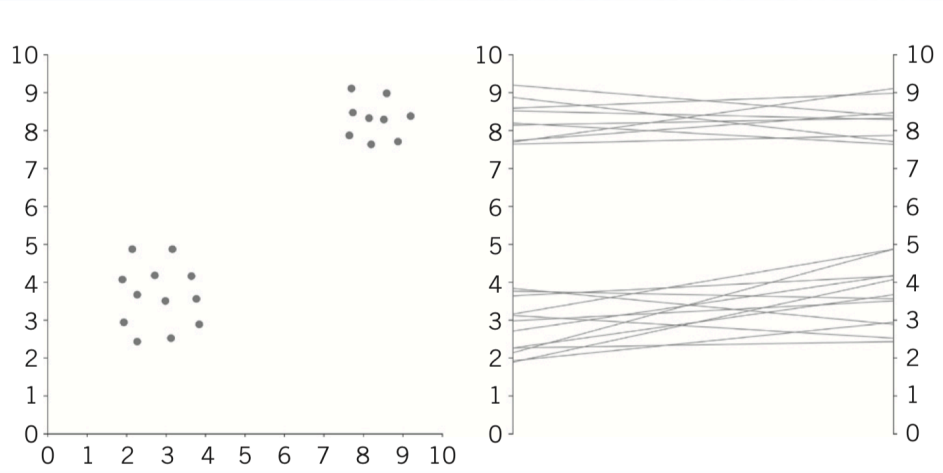
- Highlight specific patterns of association on different features



http://joules.de/files/heinrich_parallel_2015.pdf

Patterns – cont.

- Highlight specific patterns of association on different features

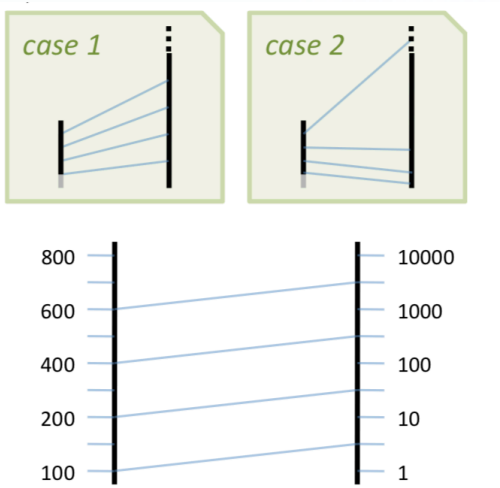


http://joules.de/files/heinrich_parallel_2015.pdf

Axes scaling with parallel coordinates

Scaling of Axes

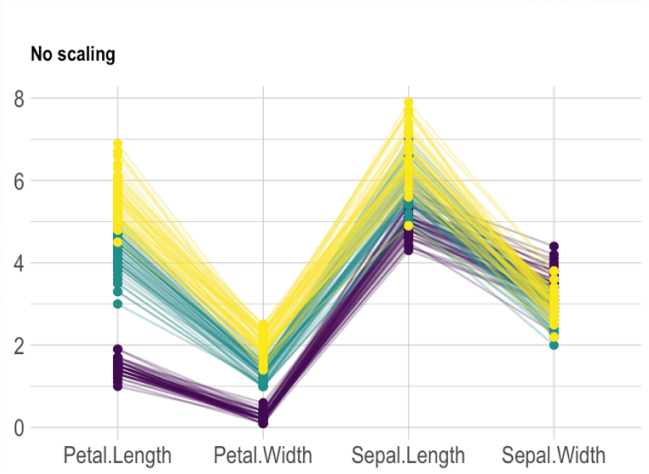
- Inconsistent scaling can lead to mis-interpretation

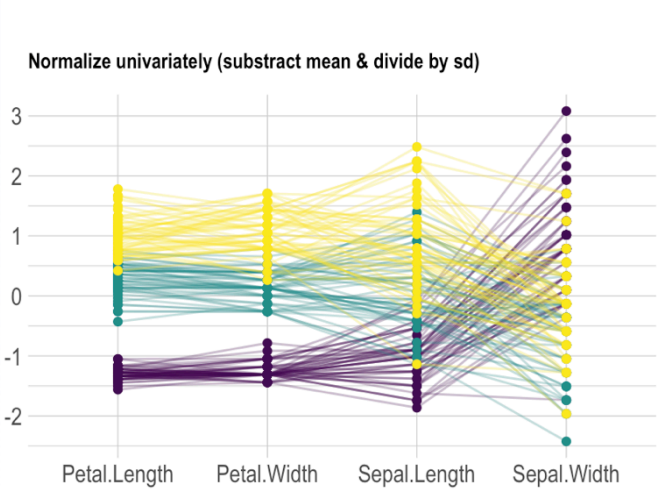


https://aedeegee.github.io/cgf12.pdf

Axes scaling – cont.

- Axes scaling affects the visualization
- May choose to scale all features via a pre-processing step





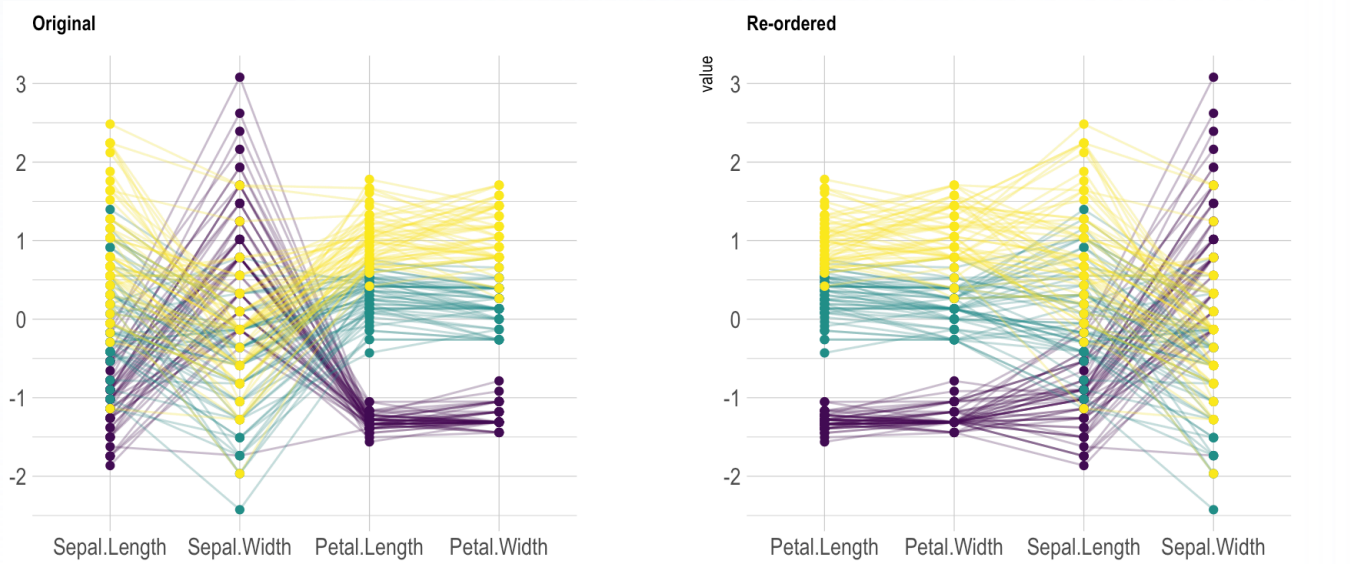
https://www.data-to-viz.com/graph/parallel.html

Axes ordering in parallel coordinates

Ordering of axes

- Influences the relationships that can be seen. Correlations between pairs of features may only be visible in certain orderings
- Can decrease the clutter
- Can reveal distinct class more clearly

Parallel coordinates – ordering of axes



https://www.data-to-viz.com/graph/parallel.html

Parallel coordinates code

Python code

- *parallel_coordinates* in *pandas.tools.plotting*
- Will practice in workshop

Very high dimensional data?

- Parallel coordinates leads to clutter and over-plotting with very large dataset and very high dimensions
  - Not enough space to draw all lines
  - Difficult to trace a line for a data object

- Only look at an important subset of attributes
  - Domain experts
  - Feature selection techniques
- Dimensionality reduction techniques

Covered later in the subject

Acknowledgements

- Material partly adapted from
  - “Data Mining Concepts and Techniques”, Han et al, 2nd edition 2006.
  - “Introduction to Data Mining”, Tan et al 2005.
  - “The grammar of graphics”, Leland Wilkinson 2005.