Visualising Multivariate Data

NANYANG THE INNOVATIVE POLYTECHNIC School of Information Technology

Learning Outcomes

By the end of this lesson, you should be able to

- Identify the patterns of multivariate data
- Visualize and compare multiple variables
- Explain and apply multivariate analysis techniques and best practices

Introduction

Multivariate analysis compares multiple instances of several variables at once.

… as compared to multiple instances of a single quantitative variable such as sales revenues per region or compare one variable to another, such as revenue to profit.

The purpose of multivariate analysis is to identify similarities and differences among items, each characterized by common set of variables

Introduction

- Imagine you are an analyst for an automobile company, how to determine which characteristics contribute most to customer satisfaction for a particular types of buyer?
  - Price
  - Gas mileage
  - Number of passengers
  - Cost of insurance
  - Maintenance and repair cost
  - Customer satisfaction rating

- Values of all these variables for each car combine to form its multivariate profile

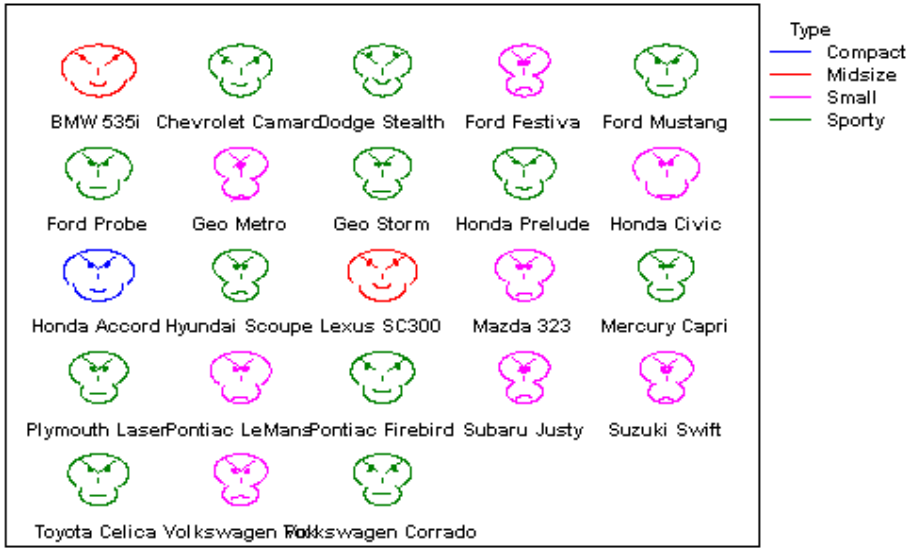
Introduction

Multivariate analysis revolves around the following questions:

- Which items are most alike?
- Which items are most exceptional?
- How can these items be combined into logical groups based on similarity?
- What multivariate profile corresponds best to a particular outcome?

Multivariate Patterns

- Ways to represent several variables worth of information about something as a single composite pattern.
- Displayed in a way that makes it easy to spot similarities and differences even if they are in hundreds.



Multivariate Displays

- Three quite different displays but one does the job poorly (so please avoid it)
  - Glyphs
  - Multivariate heat maps
  - Parallel coordinates plots

- The display that works the best is parallel coordinates plot BUT it can seem absurd and complex at first.

Glyphs

- A glyph is a graphical object designed to convey multiple data values.

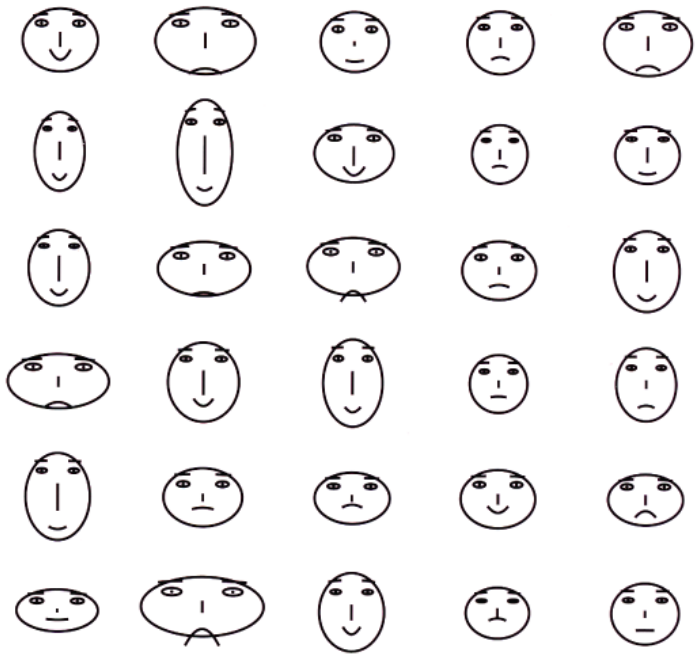
- The above three glyphs represents the health of 3 individual

The variables are encoded as follows:

Visual Attribute	Variable
Color	Body temperature
Shape of the head	Blood type
Thickness of the torso	Body mass index
Position of the arms	Heart rate
Position of the legs	Blood sugar level

Glyphs

- Best known example by Herman Chernoff in 1972.
- Facial features are used to represent different variables
  - Size of the eyes
  - Curvature of the mouth
  - Shape of the head



Glyphs

- Whiskers
  - Each line represents s different variable and its length encodes its value

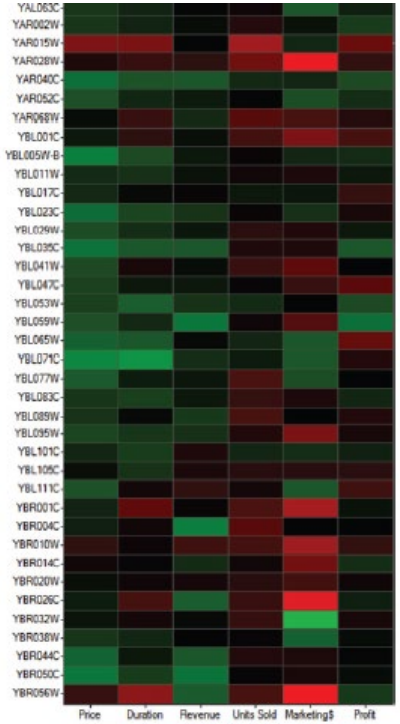
- Stars

Glyphs

- Theoretically glyphs can be use to examine and compare multivariate profiles but it can be hardly used practically.
- It is impractical to remember the variable represented by the features in the graph and compare at the same time.
- Do not let the novelty of display entice you – unless it is used on real data to solve real problem.

Multivariate Heatmaps

- Heatmaps are visual displays that encode quantitative values as variation in color.
- For example, products in row and the following variables in columns:
  - Price
  - Duration
  - Units sold
  - Revenue
  - Marketing Expenses
  - Profit

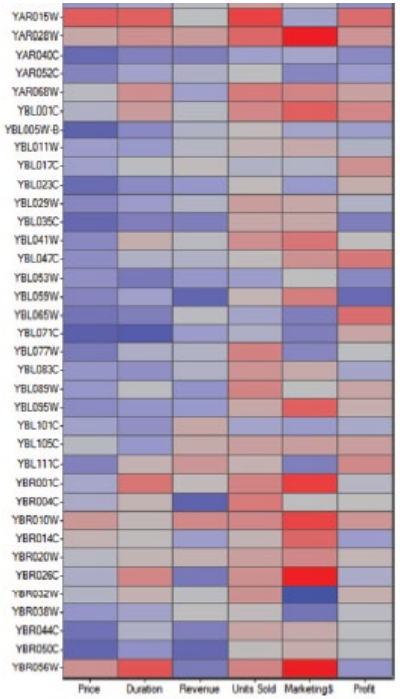


Multivariate Heatmaps

- Dark colour such as black should not be used to encode average value as it is visually prominent.
- Better colors can be used to improve the visualisation.
- It is still difficult to see the combination of colors for particular products as a pattern.

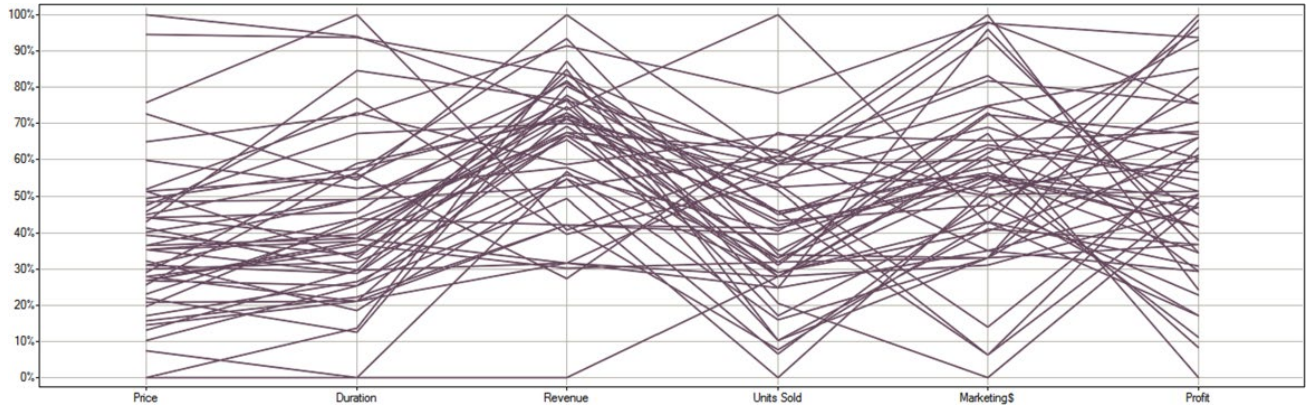
Multivariate Heatmaps

- This is a better display
- Positive value are blue, negative values are red and values near zero (average) fade from blue to red to light gray.
- The light gray used to represent numbers close to zero intuitively represents low values.



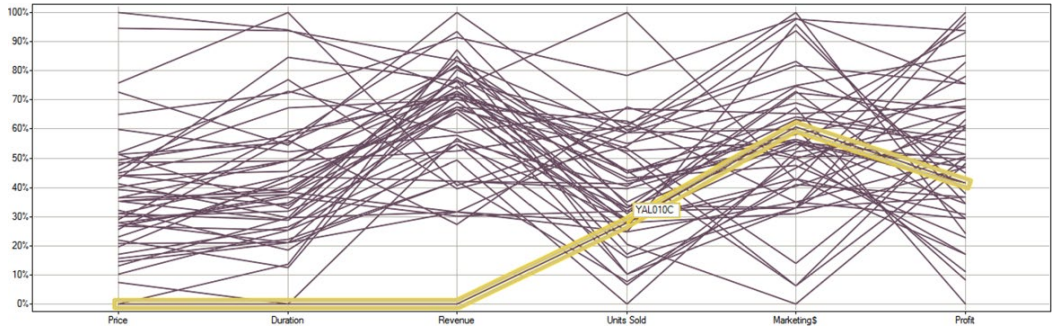
Parallel Coordinates Plots

- It includes six variables (price, duration on the market, revenue, units sold, marketing expenses, and profit) for 49 products, one product represented by each of the lines that extends from left to right across the graph.



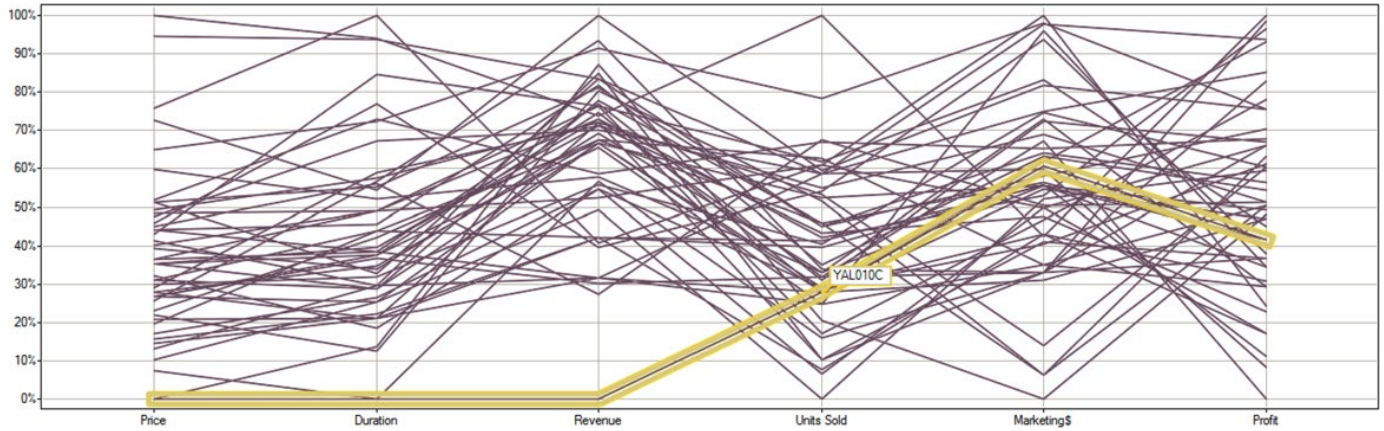
Parallel Coordinates Plots

- Regular line graphs – connect values along an interval scale such as time
- Parallel coordinates plots – connect values associated with entirely different variables.
- The individual patterns formed by different lines can be compared for similarities or differences



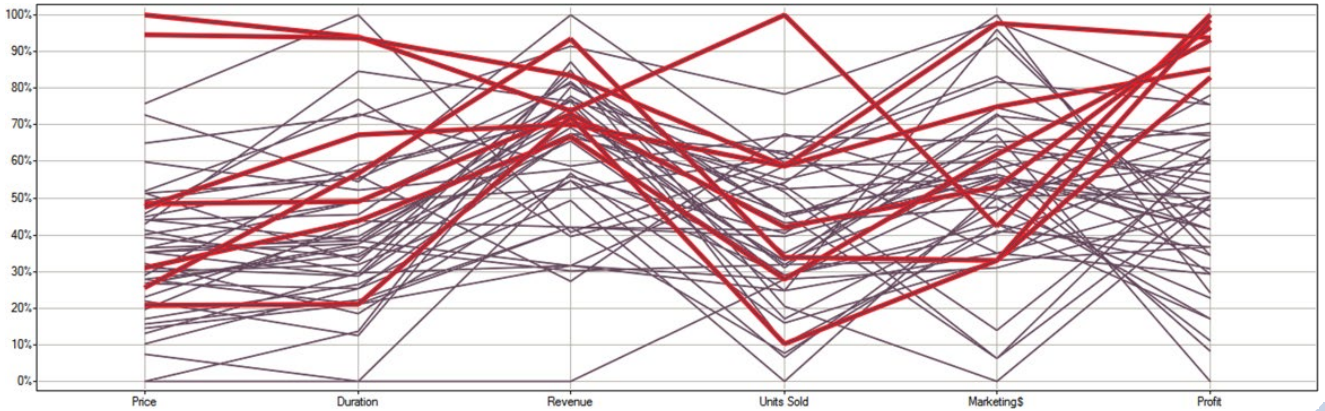
Parallel Coordinates Plots

- What else can you observe?



Parallel Coordinates Plots

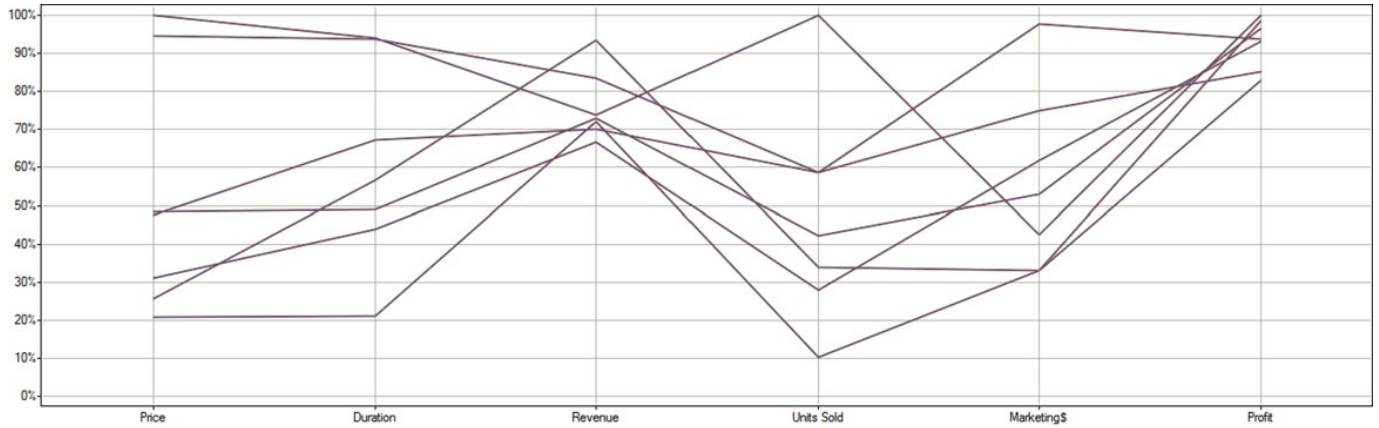
- Usually we want to look for multivariate profiles with particular condition. E.g. high profit
- Important to have interaction with the data to cut through the clutter



Products with profits above 80% have been identified.

Parallel Coordinates Plots

- Multivariate pattern associated with highly profitable products:



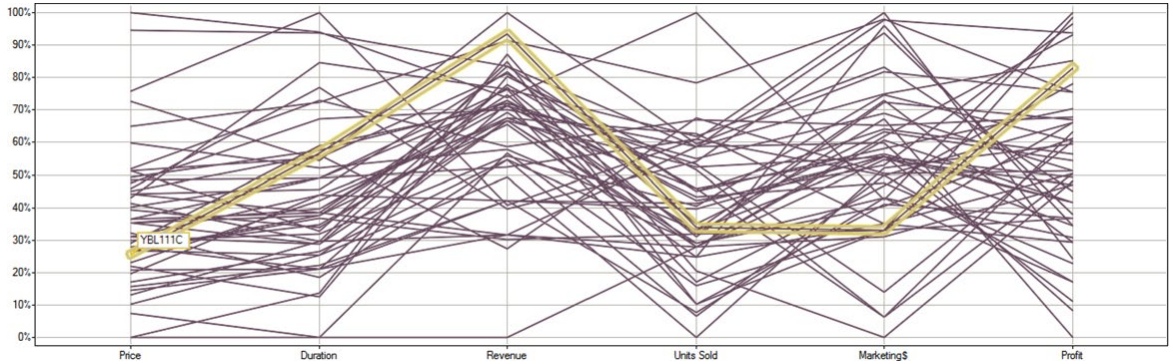
- What interesting insights can be found?

Multivariate Analysis Techniques and Best Practices

- Ranking items by similarity
- Clustering items by similarity

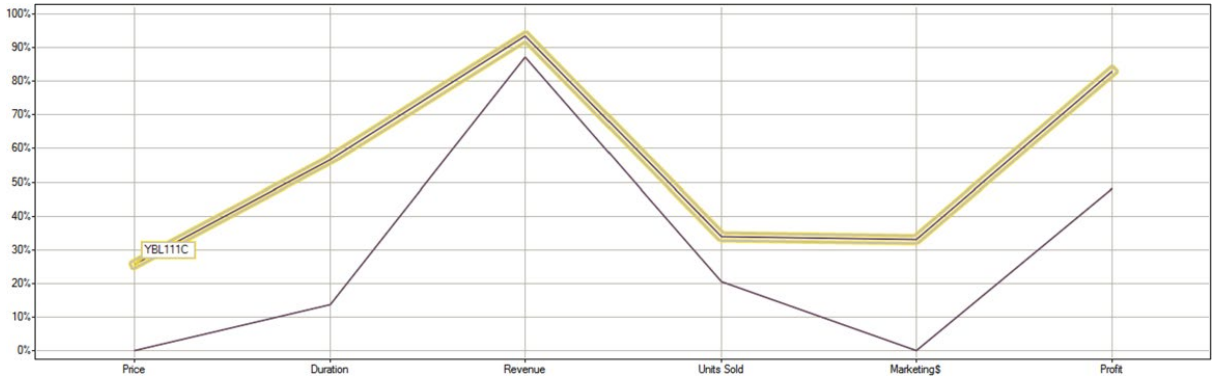
Ranking Items by Similarity

- Find items that come close to a certain pattern or matching a particular multivariate profile.



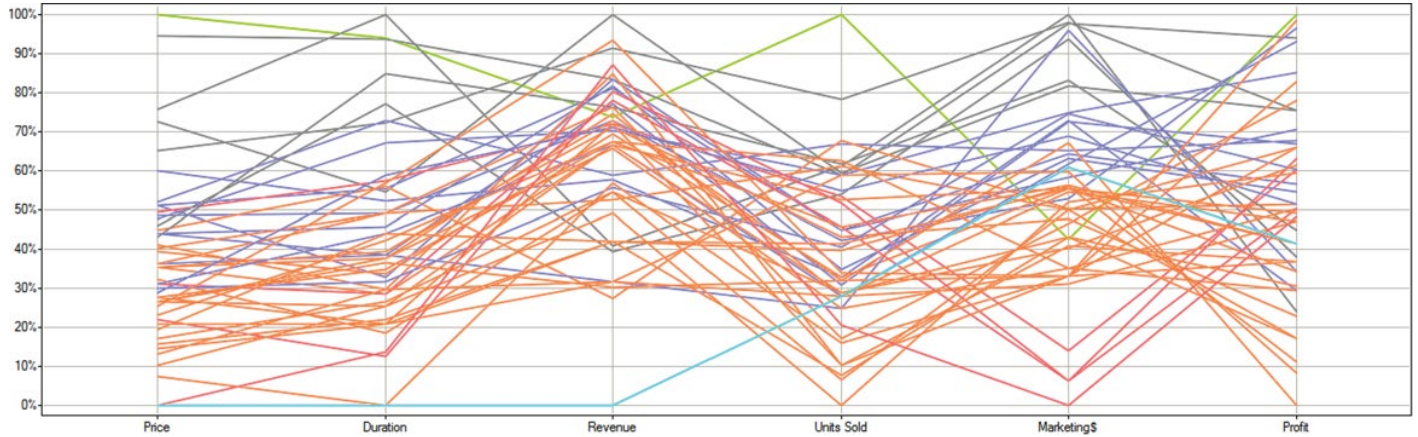
Ranking Items by Similarity

- With *pattern detection algorithm*, product(s) with similar profile can be identified and ranked for further analysis.
- Correlation study can be done to identify the one that is most similar to it.



Clustering Items by Similarity

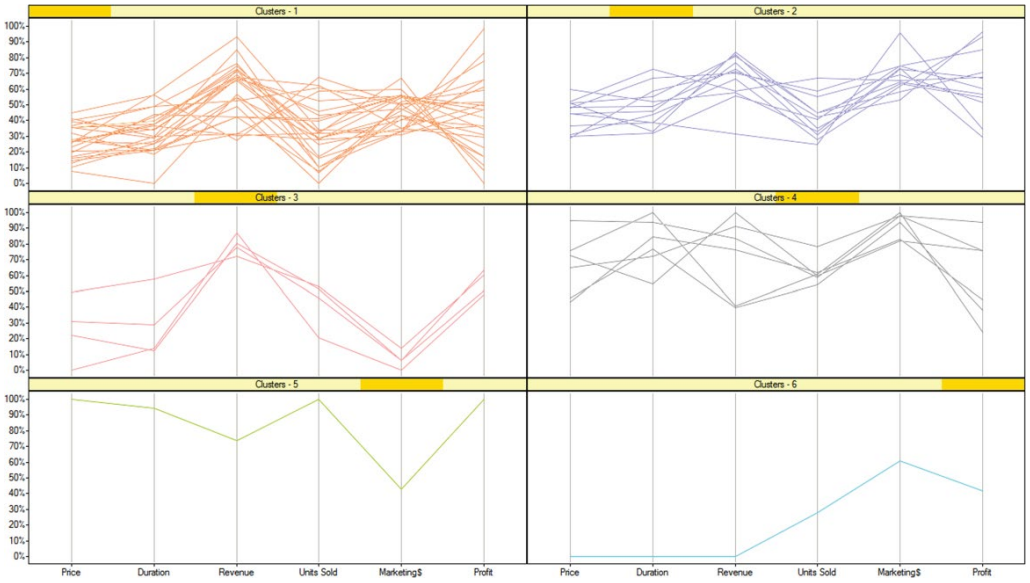
- Segmenting data into groups whose items share a similar feature or features using *statistical clustering algorithm*.



Even though the groups are distinguished by color, it's hard to see their similarities in the midst of this visual clutter.

Clustering Items by Similarity

- Trellis display – one group per graph



Summary

- Characteristics of multivariate data
- How to display multivariate data
- Techniques and best practices to consider for visualising multivariate data