Finding Groups: Dimension Reduction and Clustering

Research Questions

- Which group(s) perform better/worse when [intervention]?
- What is the relationship between [suspected cause] and [suspected effect]?

- How can I group [these items] in such a way that I can describe a sufficient amount of the variance?
- How short can [my survey] be to capture approximately the same amount of predictive/explanatory power?
- Is there a way to reduce the number of variables I have in my models without losing too much information?

$$Y = b_1 x_1 + b_2 x_2 + \cdots + b_k x_k + b_0$$

Supervised Learning

- Example: Split-sample cross-validation
  - Split data set
  - Build a model on half
  - Apply solution on other half
  - Calculate the fit

- Providing the model with feedback to determine its usefulness

Unsupervised learning

- Organizes a set of variables or observations to meet some criteria
  - Minimization/Maximization

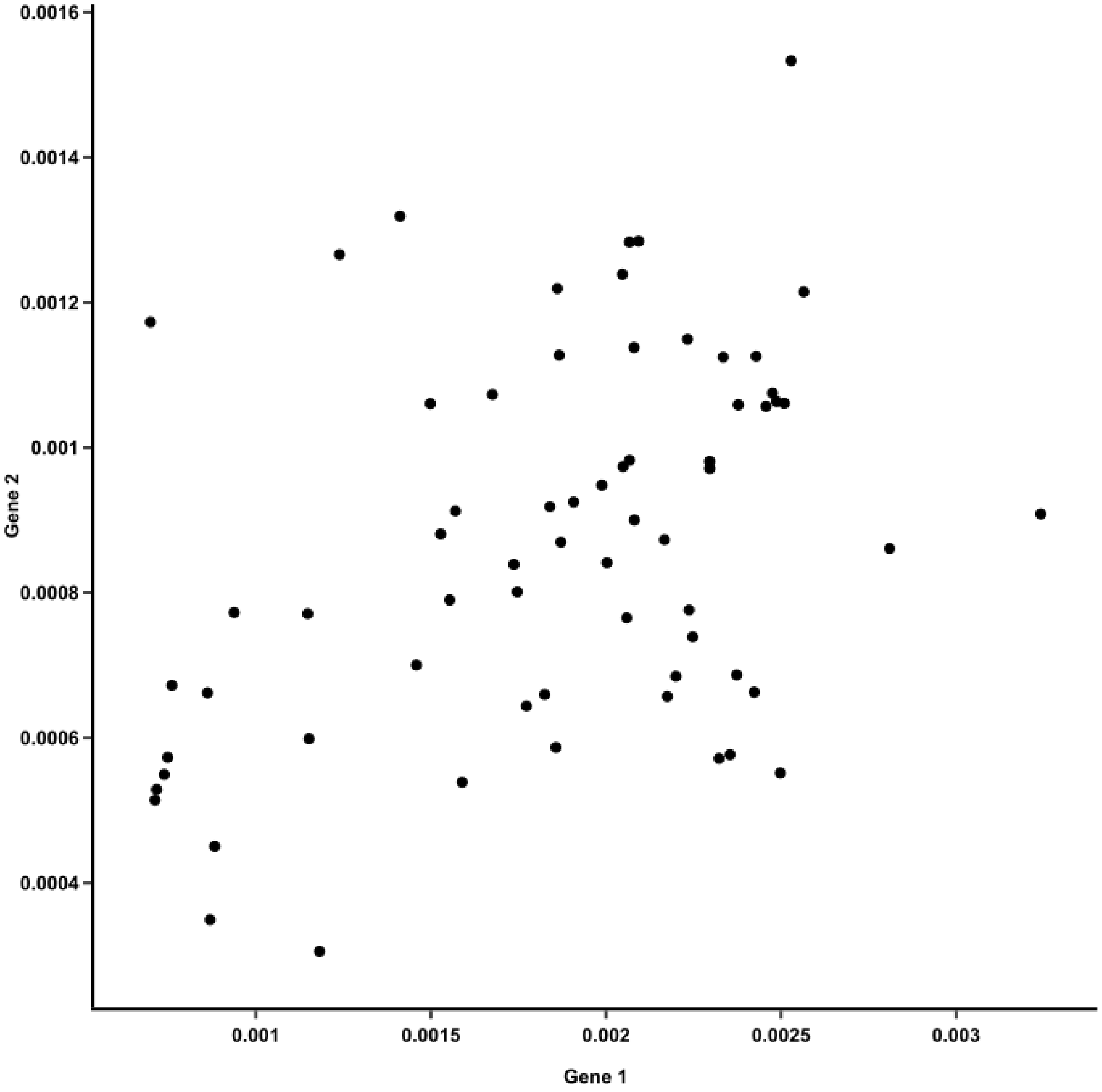
- We let the algorithm just go. No feedback from us.

Dimension reduction & clustering

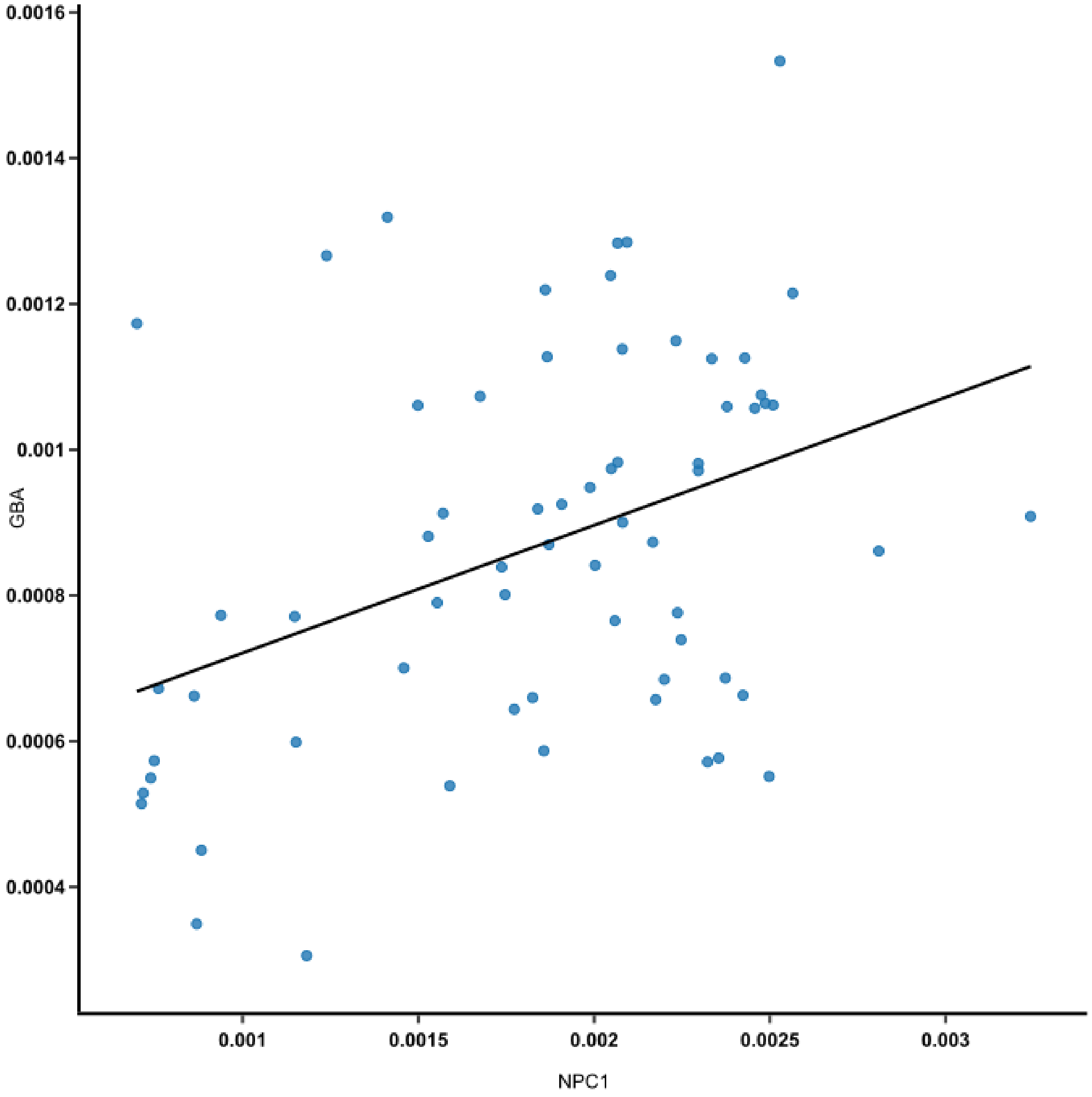
- Unsupervised learning techniques
    - Principal component analysis (PCA)
    - k-means clustering
    - And many more…
        - A few examples

- Maximize variance or minimize error

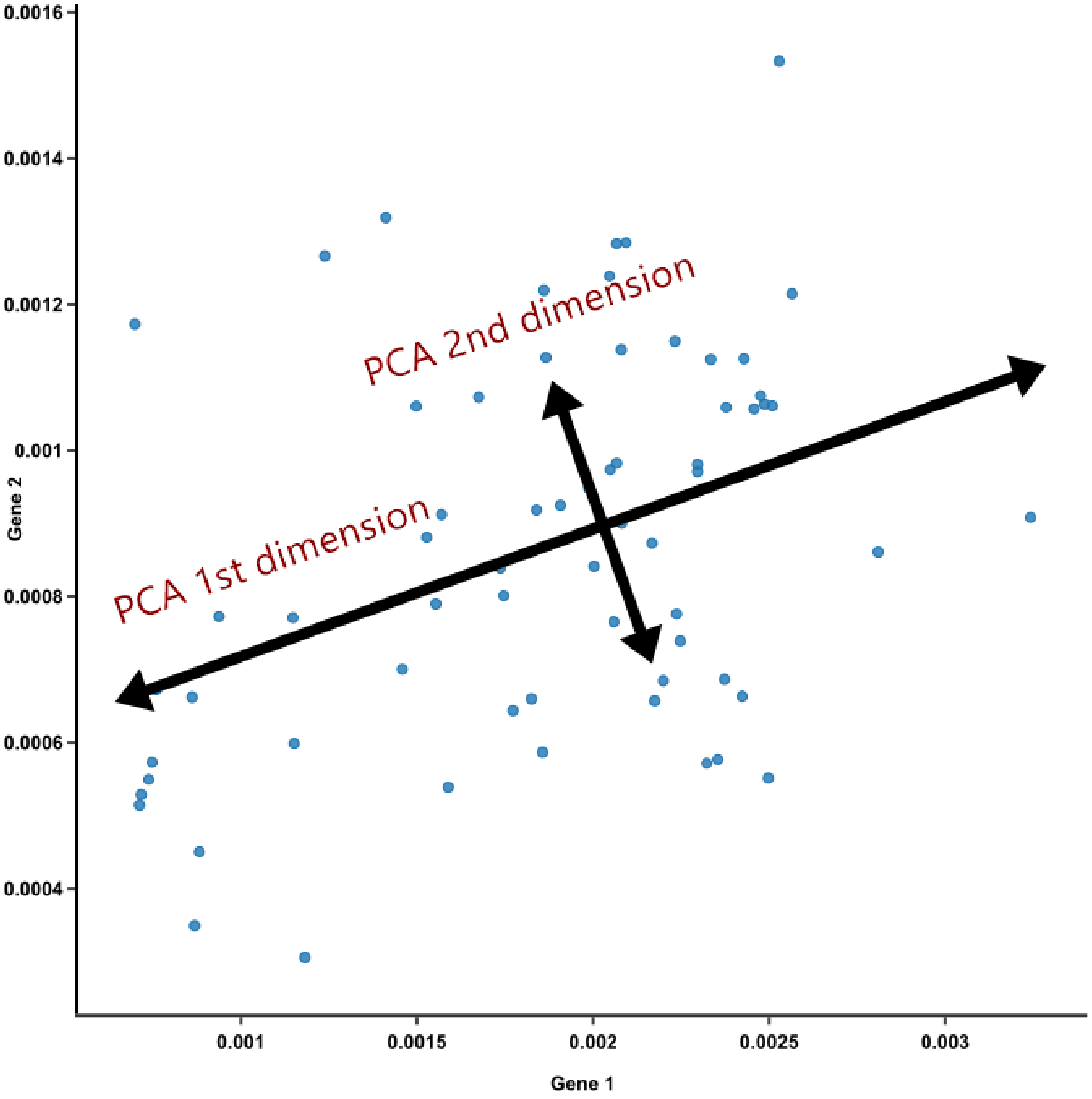
Principal Component Analysis



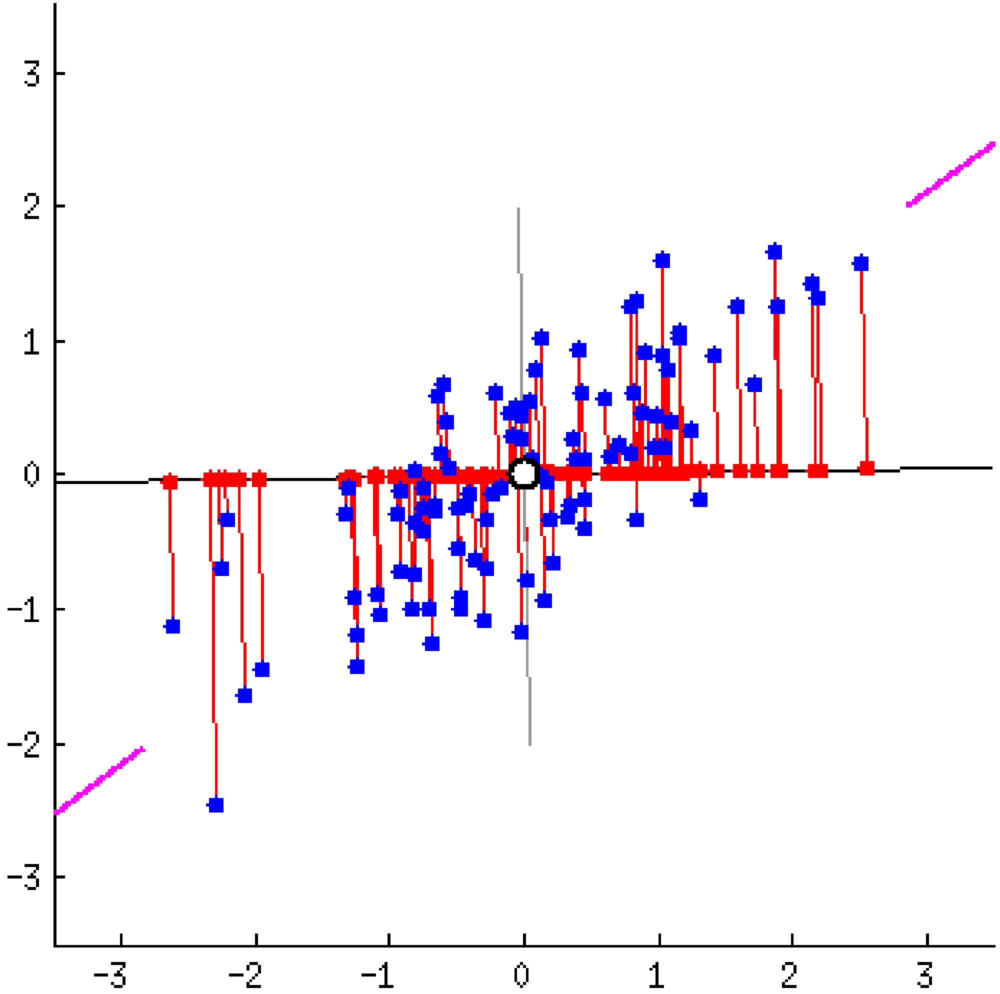
Principal Component Analysis



Principal Component Analysis

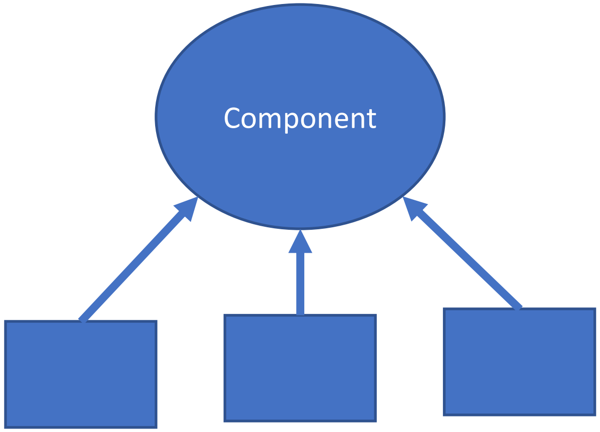


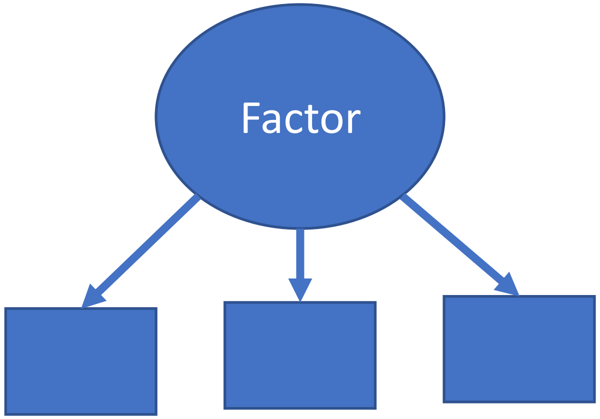
Principal Component Analysis



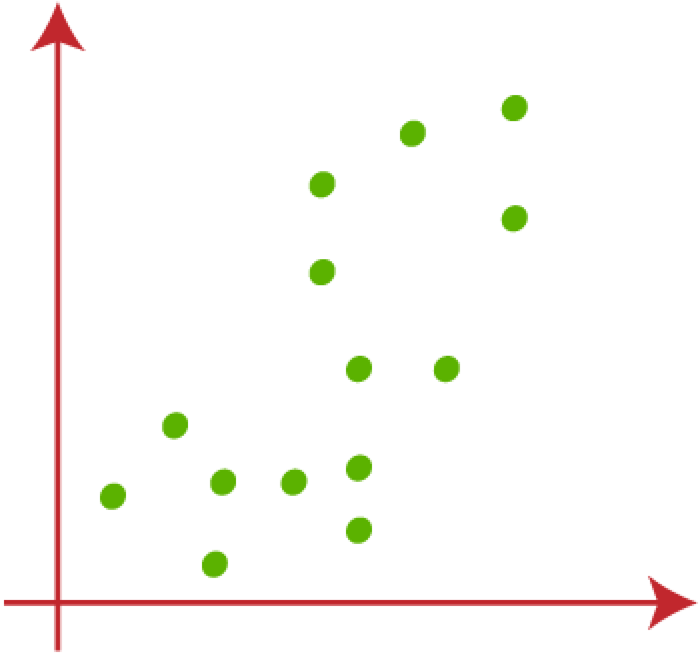
PCA and EFA

- PCA is similar to Exploratory factor analysis
- With a large number of features and observations, essentially the same answers
- Maximizing variance → PCA
- Estimating shared/common variance → EFA

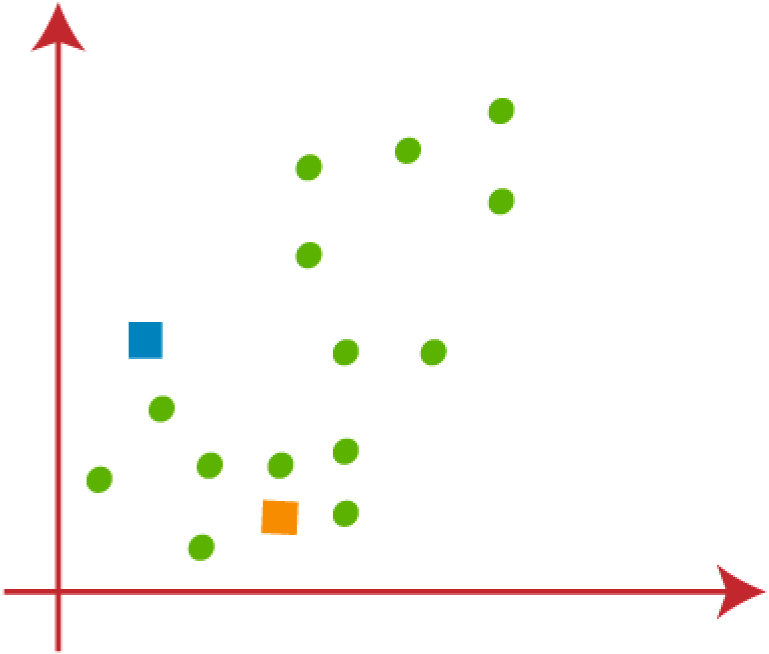




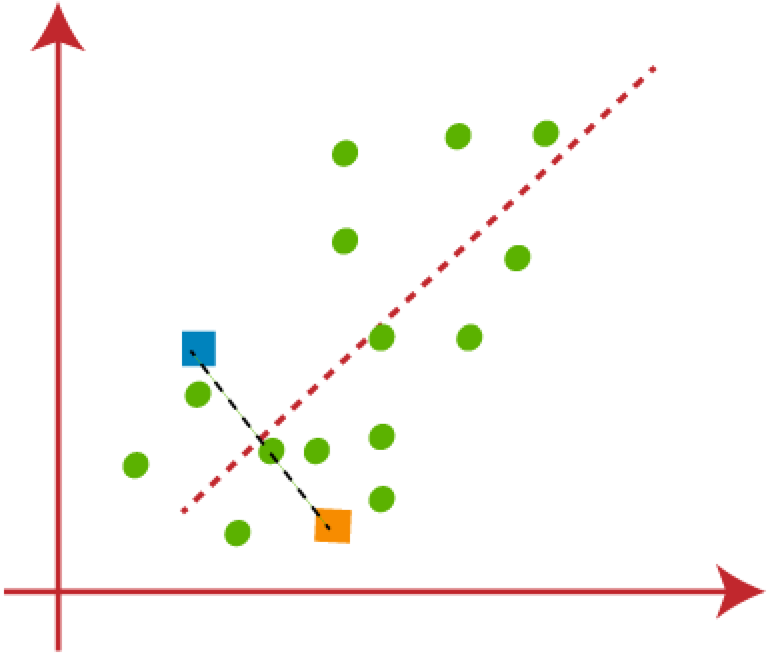
k-means clustering



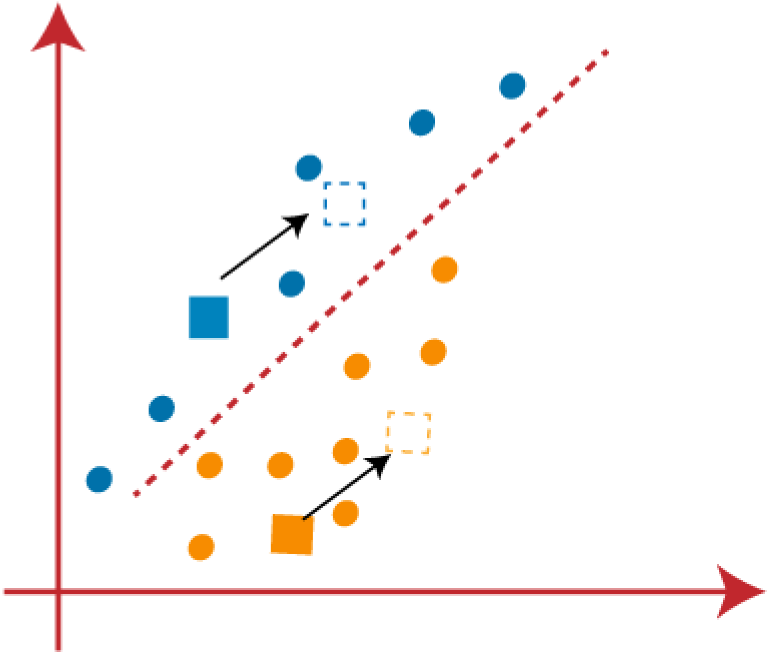
k-means clustering



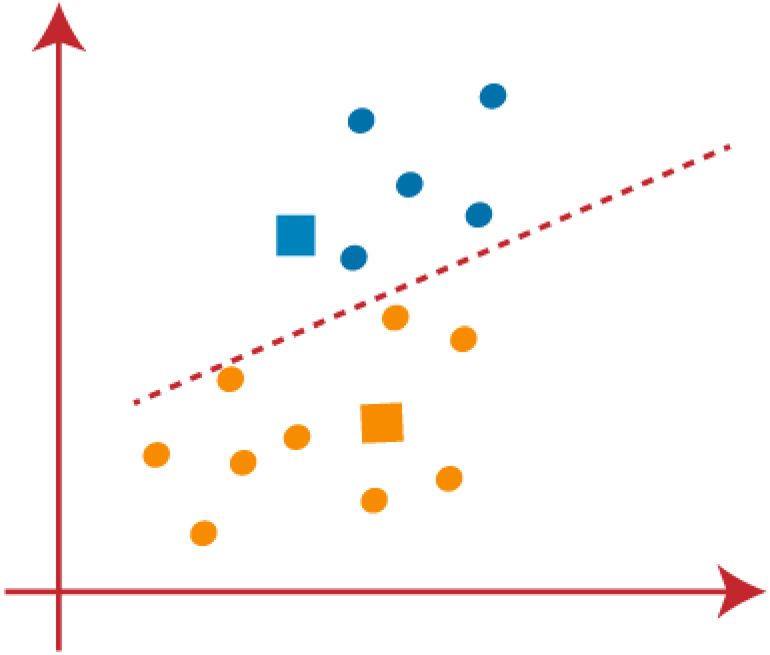
k-means clustering



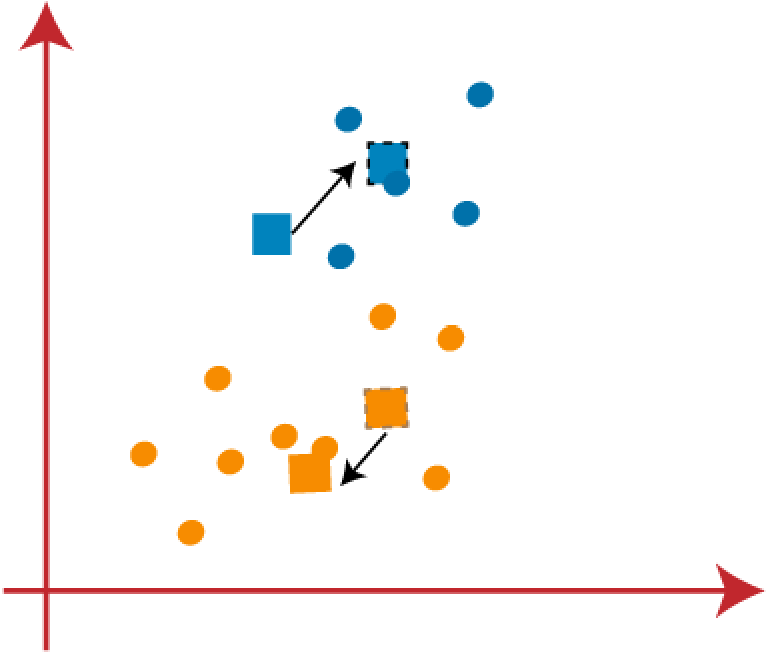
k-means clustering



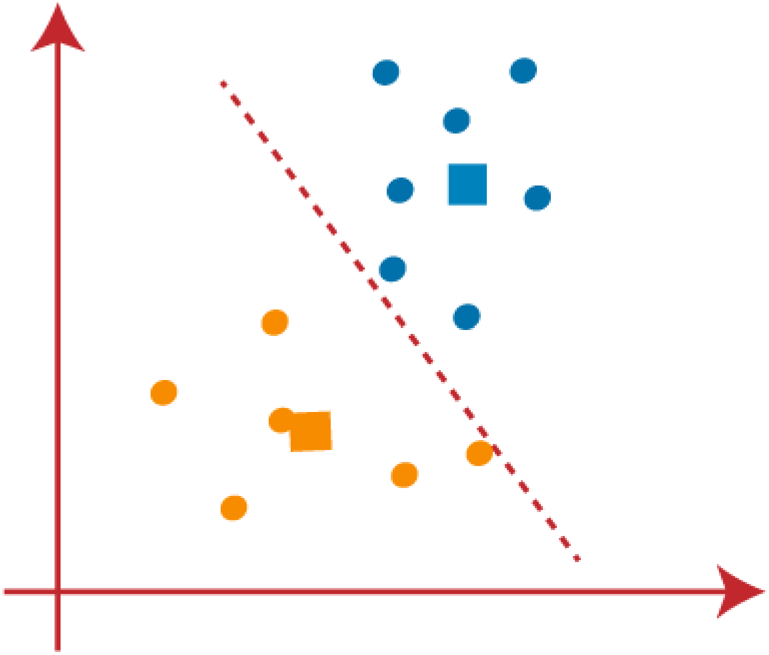
k-means clustering



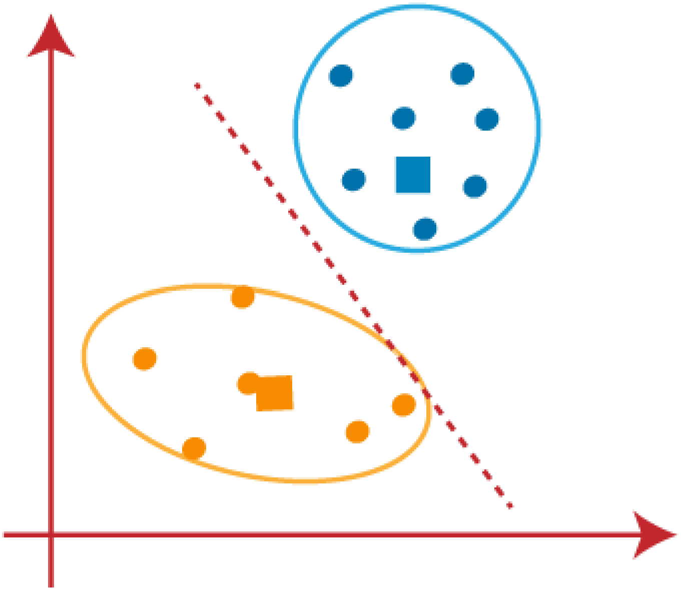
k-means clustering



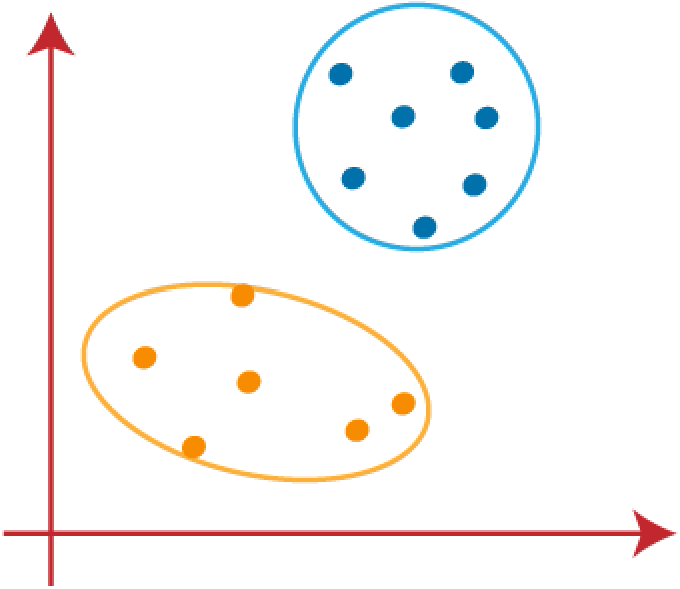
k-means clustering



k-means clustering



k-means clustering



Why do dimension reduction?

- Rhetorical simplification
  - Distill many variables to just a handful
    - Cleaner stories
- Pattern recognition
  - Find relationships you might not have thought of
    - Quantitatively bundle variables that “go together”