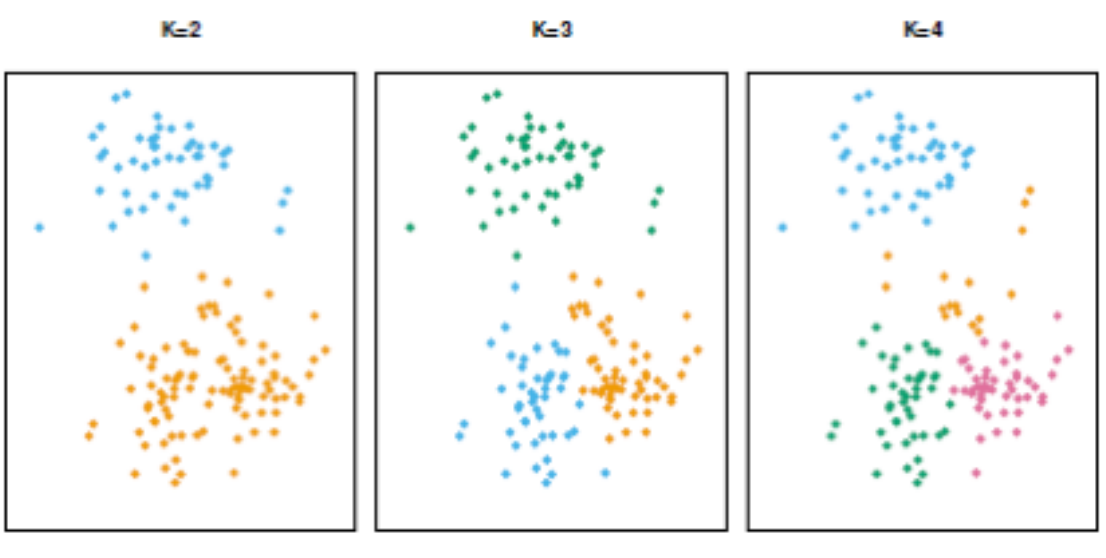

H2O.ai

K-Means Clustering

- K-Means clustering groups observations based on numeric features
  - Assumes clusters are roughly the same sized hyperspheres
  - Minimize Euclidean distance between observations and cluster centers
- Number of methods for choosing the number of clusters, k
  - Choose several and evaluate performance
  - Use business rules


K=2
K=3
K=4

Pros and Cons of K-Means Clustering

Pros

- Fast, Scalable Algorithm

Cons

- Choice of k can be tricky
- Euclidean distance not robust
  - Hyperspheres not common
  - Sensitive to correlated measures
  - Sensitive to scaling
  - Sensitive to skewed measures
  - Sensitive to outliers
- Categorical data requires preprocessing
  - Multiple Correspondence Analysis
  - Multi-Dimensional Scaling

K-Means Clustering

- K-Means clustering groups observations based on numeric features
  - Assumes clusters are roughly the same sized hyperspheres
  - Minimize Euclidean distance between observations and cluster centers
- Number of methods for choosing the number of clusters, k
  - Choose several and evaluate performance
  - Use business rules