Rotman

INTRO TO R PROGRAMMING

R Tutorial (RSM456) – Session 3

January 22, 2024 Prepared by Jay Cao / TDMDAL

Website: https://tdmdal.github.io/r-intro-2024-winter/

Rotman School of Management
UNIVERSITY OF TORONTO

K-Means Cluster Analysis

- A method to partition n observations into k clusters
  - such that total **w**ithin-cluster **s**um of **s**quares (WSS, sum of squared distance between observations to cluster centroid) is minimized

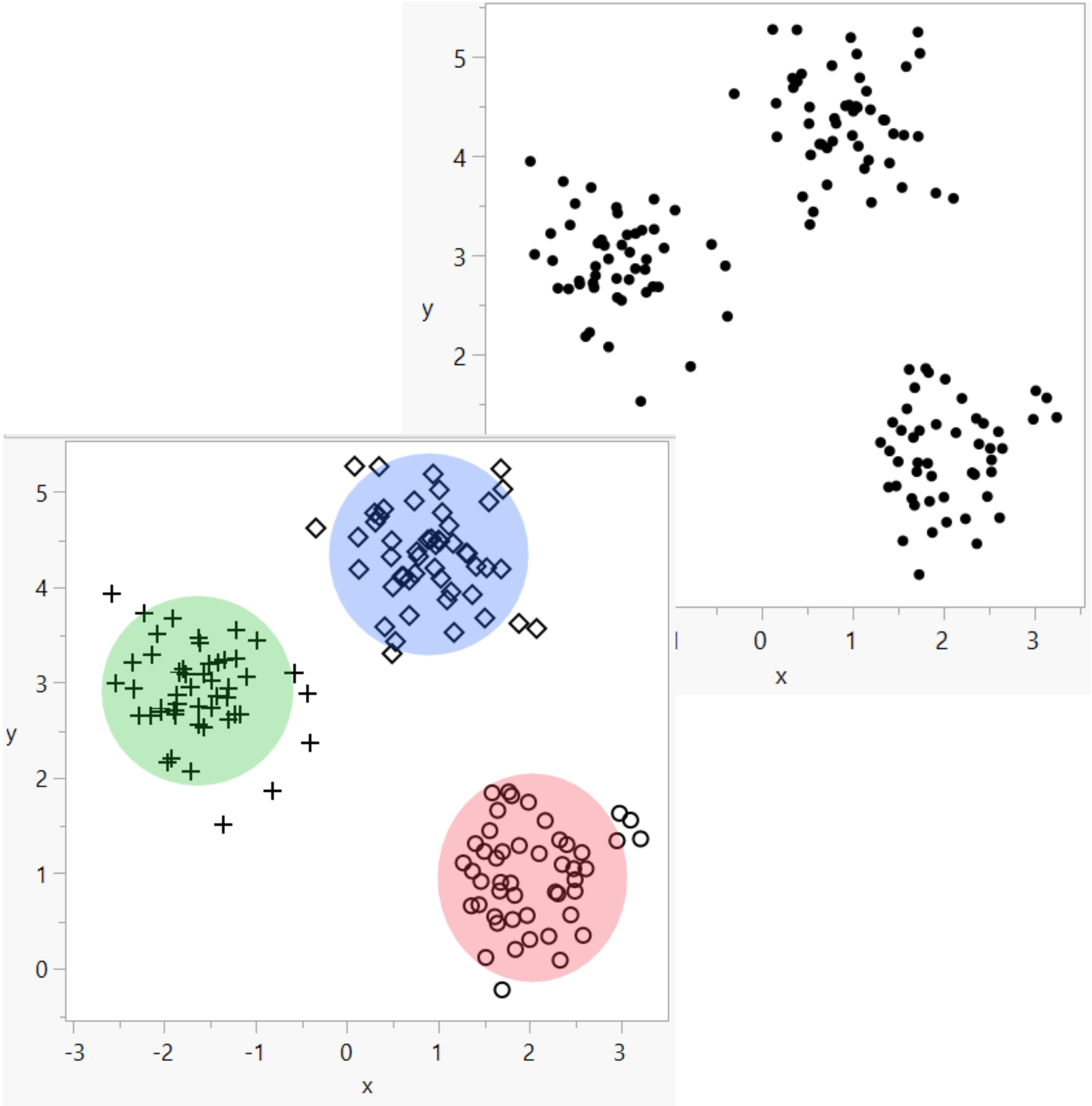
- A **cluster** refers to a collection of data points aggregated together because of certain similarities
  - Similarity based on a distance measure

- Need to set k
  - There are methods to help you decide the value of k

K-Means Cluster Analysis: A Simple Example

- Observations: 150 2-d points
- Set $k = 3$
  - partition each observation to one of the 3 clusters $S = \{S_1, S_2, S_3\}$
- K-means clustering algorithm finds 3 clusters such that

$$\underset{S}{\operatorname{argmin}} \sum_{i=1}^{3} \underbrace{\sum_{x \in S_i} \|x - \mu_i\|^2}_{\text{Within-cluster sum of squares}}$$



K-means in R – Country Risk Exercise

- Import the `country_risk.xlsx` data

- Prepare the data for k-means clustering
  - Perform correlation analysis and choose features
  - Standardize the features

- Perform a K-means cluster analysis
  - Determine *k* using the “elbow” method
  - Run k-mean clustering algorithm for a chosen *k* (i.e., fit/learn/estimate the model)
  - Interpret/name the clusters