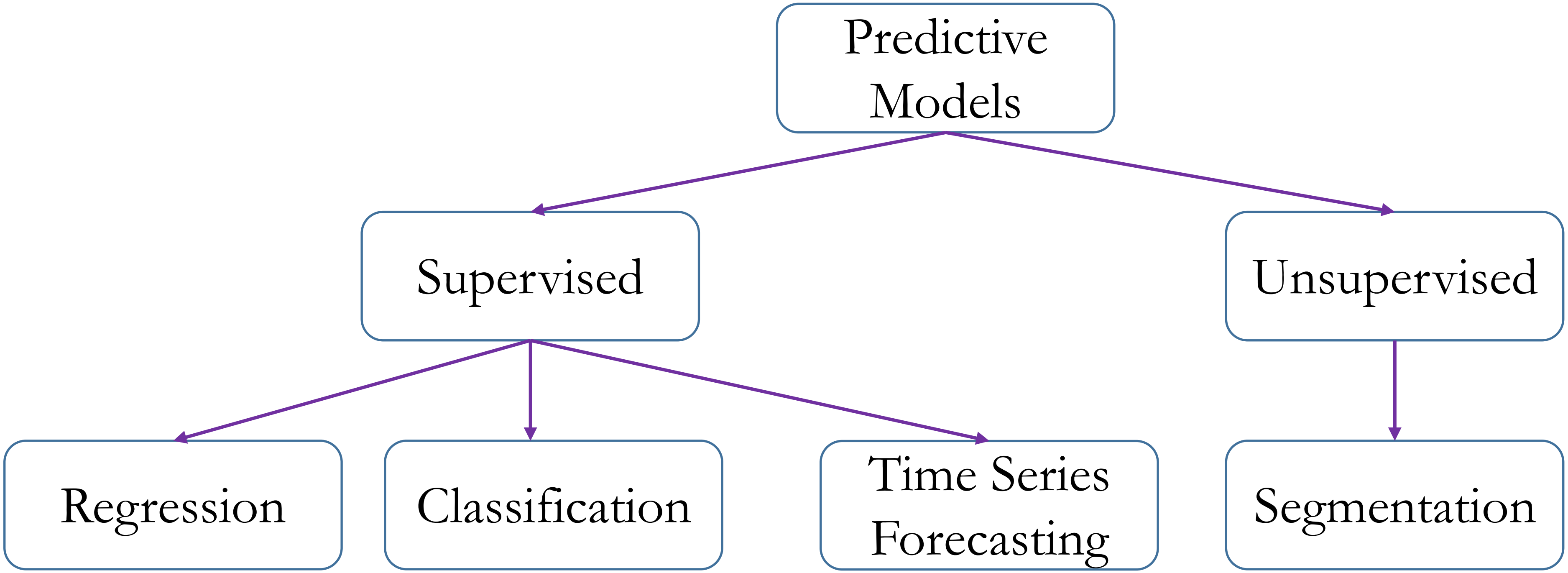
Cluster Analysis



Predictive Models

Supervised

Unsupervised

Regression

Classification

Time Series Forecasting

Segmentation

- ***k*-Nearest Neighbor**
- **Linear Regression**
- **Regression Trees**
- Neural Networks
- Ensembles
- ……

- ***k*-Nearest Neighbor**
- Naïve Bayes
- **Logistic Regression**
- **Classification Trees**
- Neural Networks
- Discriminant Analysis
- Ensembles
- ……

- **Regression-based**
- Smoothing methods
- ……

- **Clustering**
- ……

Introduction

- Popular unsupervised learning method
- Goal
  - Segment the data into set of homogeneous cluster of records
- Based on several measurements made on the records
- Helps improve other supervised learning methods performance
- How?
  - Model each cluster separately than the entire heterogeneous dataset
- Popular clustering methods
  - Hierarchical clustering
  - *k*-means clustering (widely used)

Applications

- Astronomy, Archaeology, medicine, chemistry, education, psychology, linguistics, sociology etc.
- Biologists : Group and organize species
- Chemistry : Mendeleev's periodic table
- Business : Market segmentation (segment customers based on demographics
- Politics : cluster neighborhoods by lifestyles
- Finance : creating balanced portfolios, industry analysis
- Internet : cluster queries that users submit (helps improve search algorithms)

Interesting application

- Design of new set of sizes for army uniforms for women in US army
- Study came up with a new clothing size system with only 20 sizes, where different sizes fit different body types
- 20 sizes are combinations of five measurements :
  - Chest
  - Neck
  - Shoulder circumference
  - Sleeve outseam
  - Neck-to-buttock length

Data on Public Utilities

- 22 Public utilities
- Interested in forming groups of similar utilities
- Attributes
  - company : name of the utility
  - Fixed_Charge : fixed-charge covering ratio (income/debt)
  - ROR : rate of return on capital
  - Cost : cost per kilowatt capacity in place
  - Load_Factor : annual load factor
  - Demand_Growth : peak kilowatt demand growth from 1974 to 1975
  - Sales : sales (kilowatthour use per year)
  - Nuclear : percentage nuclear
  - Fuel_Cost : total fuel costs (cents per kilowatthour)

Today’s class mandatory steps

- Create a folder name “**p. cluster**” within the folder

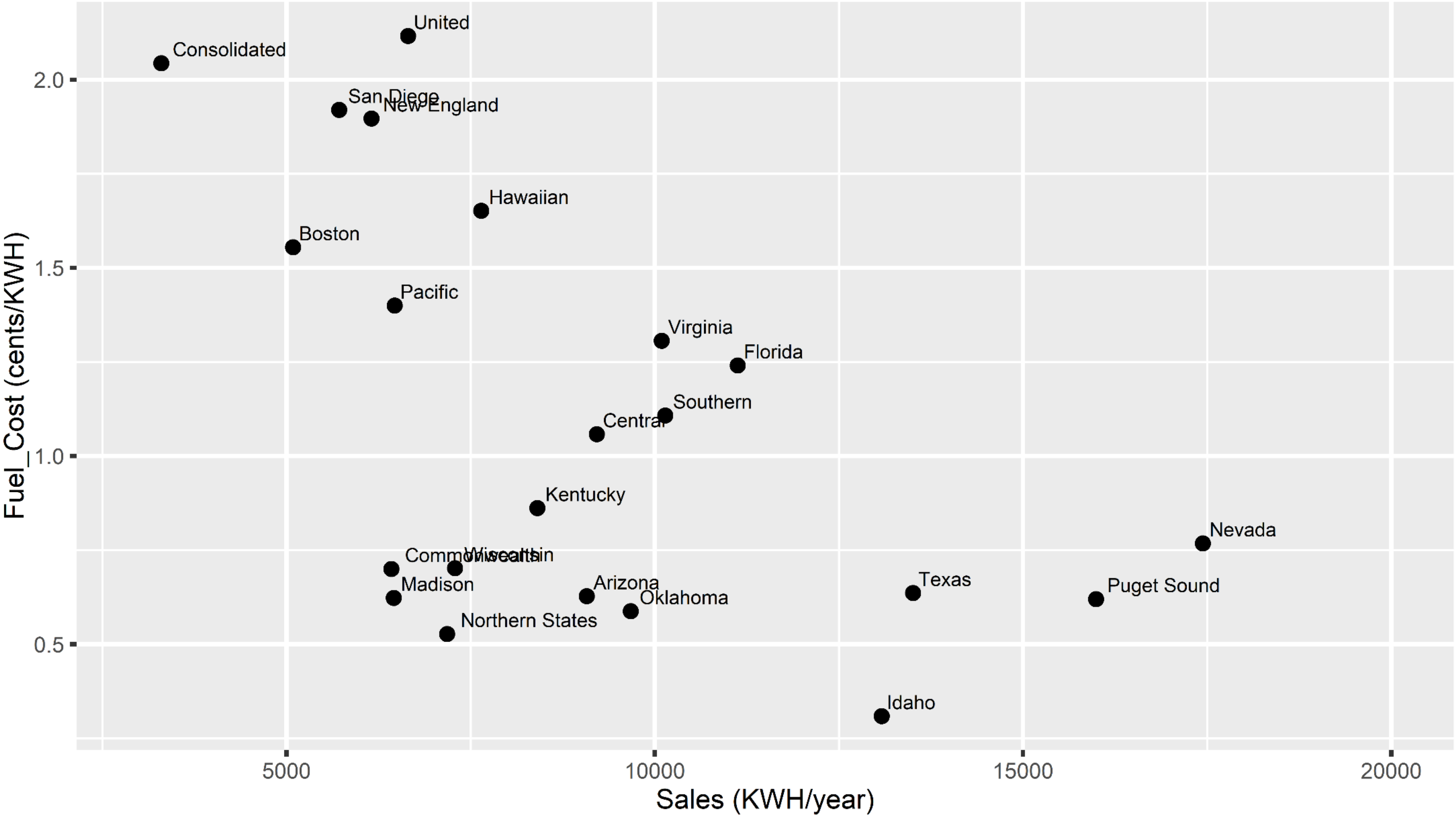
  “**oba_455_555_ddpm_r/rproject**”

- Download “**cluster_code.R**”, and all **csv** files from canvas
- Place all downloaded files in

  “**oba_455_555_ddpm_r /rproject/ p. cluster**”

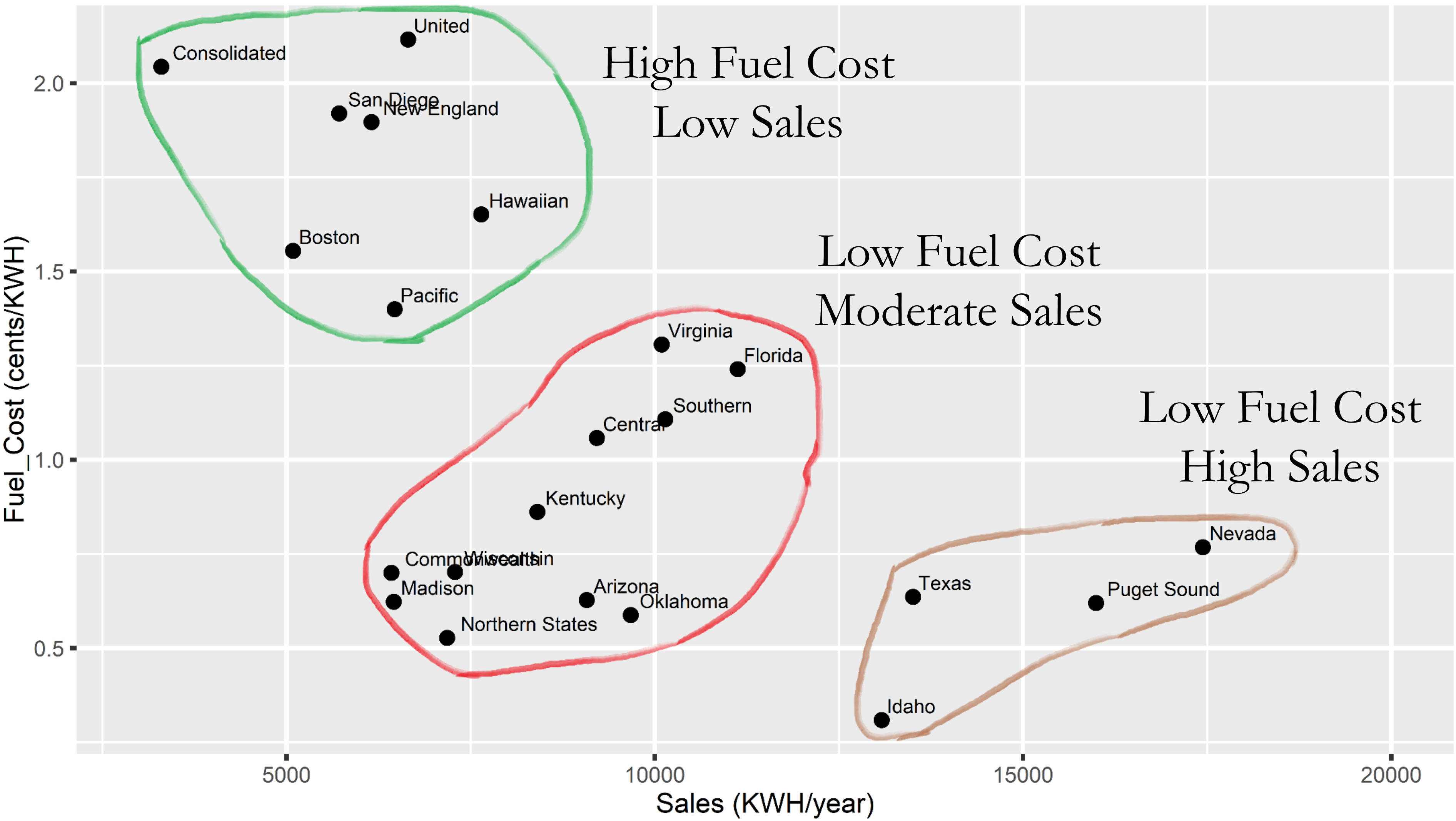
- Open RStudio project
- Open “**cluster_code.R**” file within RStudio

Scatter plot of only two variables



Difficult to visualize 8 variables at a time

Reveals ~ 3 clusters by looking at only 2 var



High Fuel Cost
Low Sales

Low Fuel Cost
Moderate Sales

Low Fuel Cost
High Sales

Based on distance between observations

Distances

- Types
  - **Euclidean**
  - Mahalonobis
  - Manhattan
  - Maximum coordinate
  - ……
- Distance between observations is highly influenced by scale
- Distance have to be unit free
- Solution ?
  - Standardization/Normalization

Transformation of data

- Standardization/Normalization
- Subtract mean from each observation
- Divide the result by standard deviation

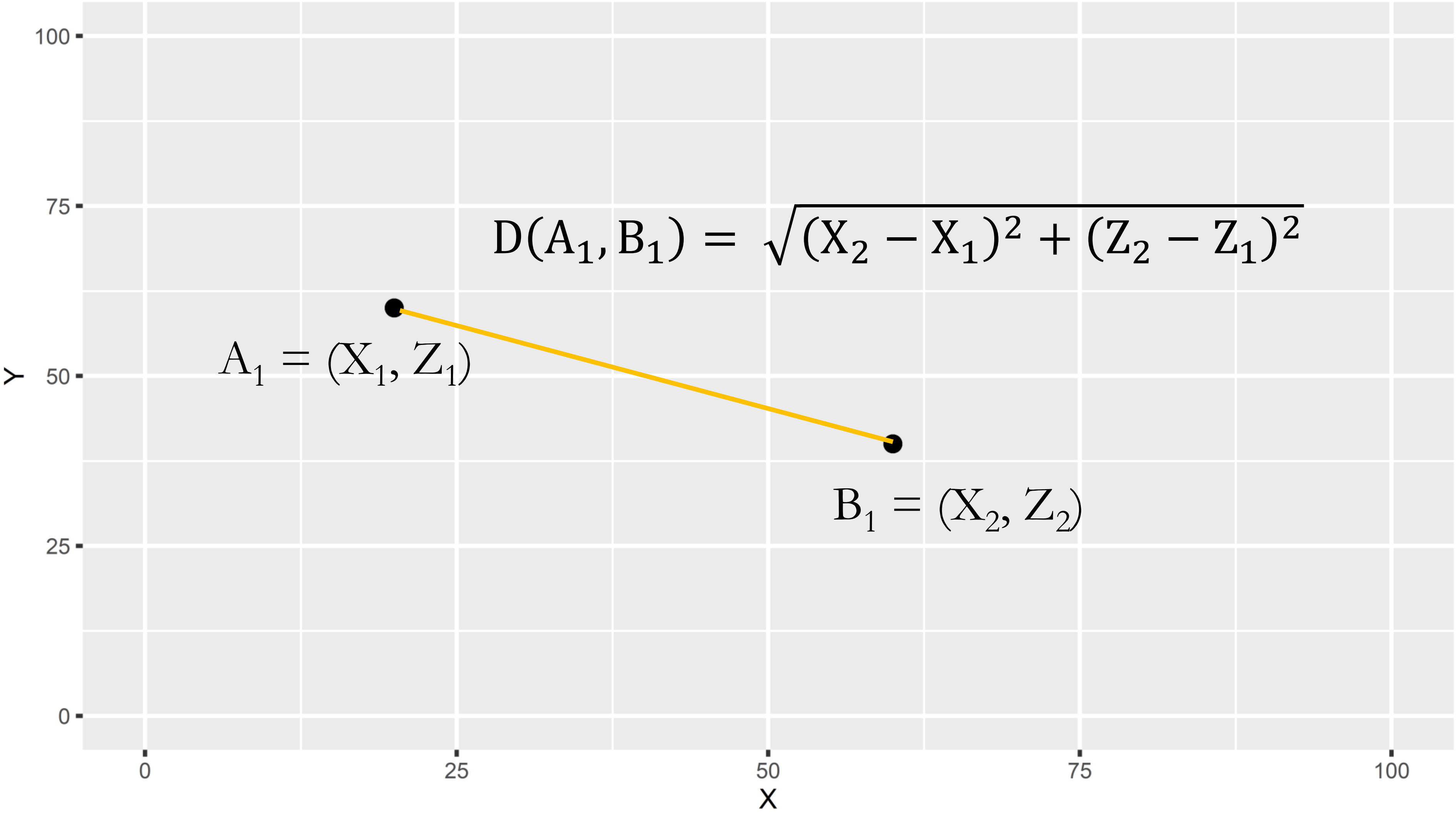
X
64
18
24
46
72

- **m = mean(c(64, 18, 24, 46, 72))**
- **s = sd(c(64, 18, 24, 46, 72))**
- X_norm = (X-m)/s

X_norm
0.8076
-1.1273
-0.8749
0.0505
1.1441

- Mean of normalized data is 0
- Standard deviation of normalized data is 1

Euclidean Distance between points



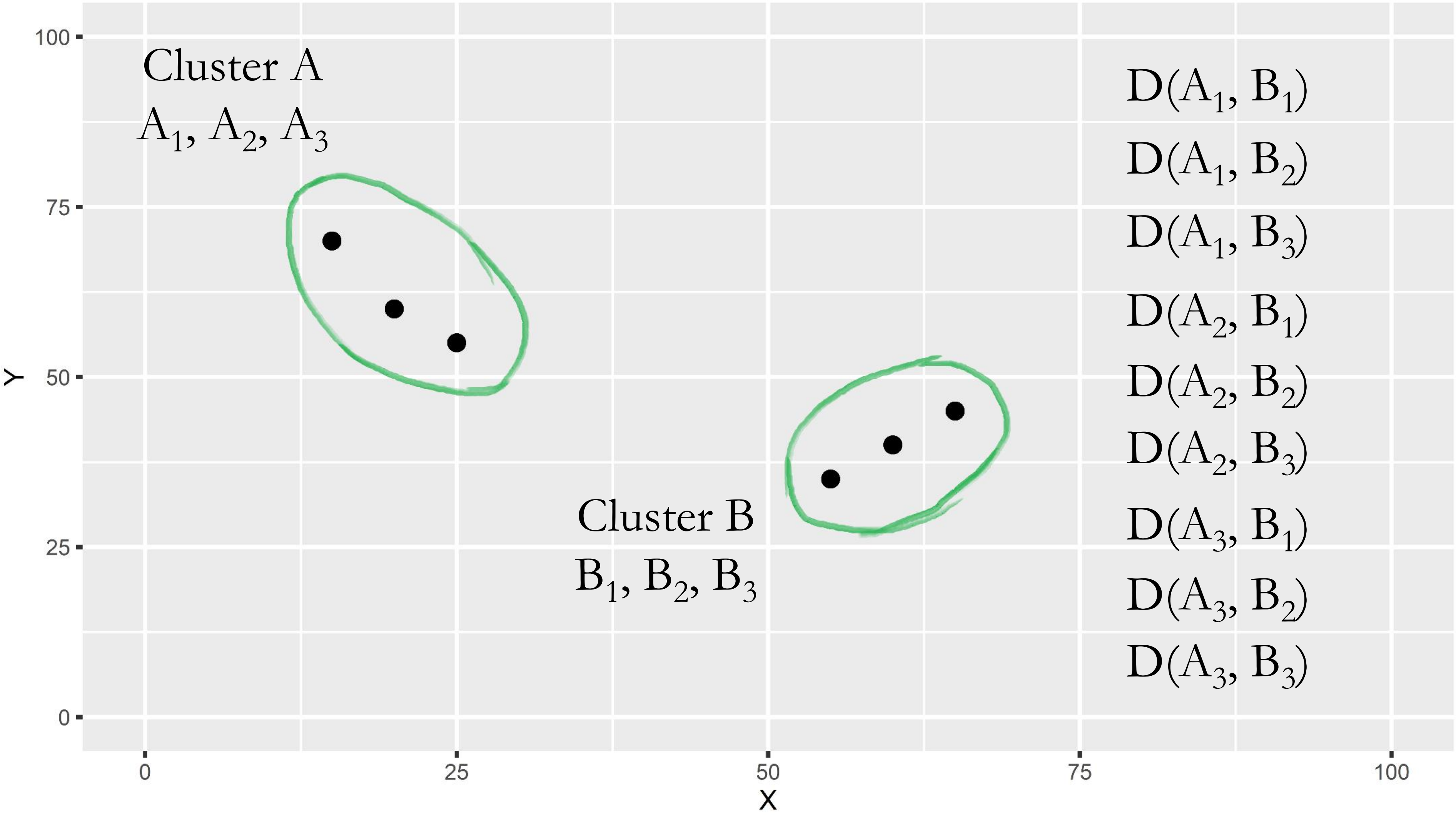
$$D(A_1, B_1) = \sqrt{(X_2 - X_1)^2 + (Z_2 - Z_1)^2}$$

$A_1 = (X_1, Z_1)$

$B_1 = (X_2, Z_2)$

X_1, X_2, Z_1, Z_2 are all standardized values

Types of distances between clusters



Cluster A
A_1, A_2, A_3

Cluster B
B_1, B_2, B_3

$D(A_1, B_1)$
$D(A_1, B_2)$
$D(A_1, B_3)$
$D(A_2, B_1)$
$D(A_2, B_2)$
$D(A_2, B_3)$
$D(A_3, B_1)$
$D(A_3, B_2)$
$D(A_3, B_3)$

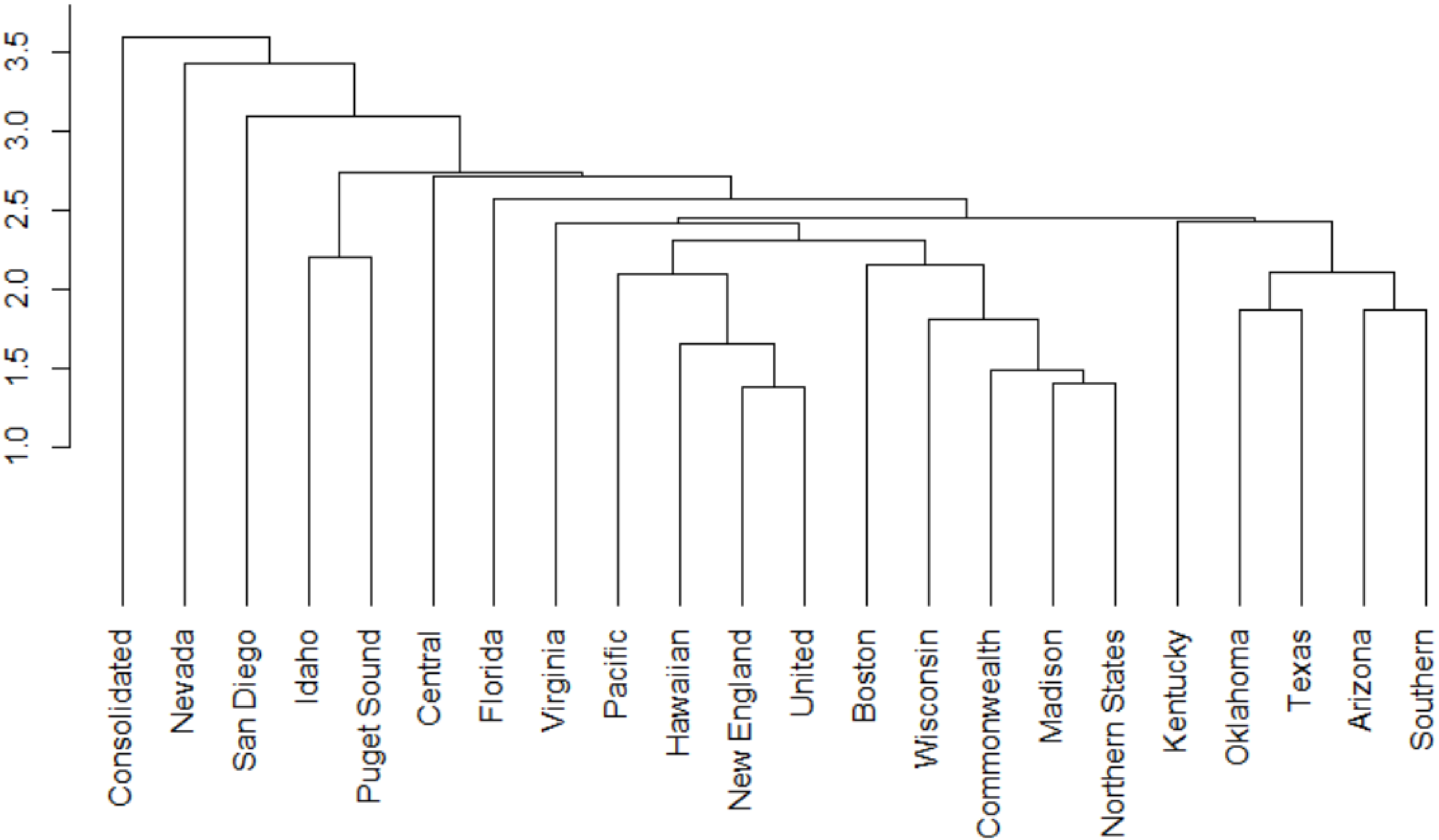
Types of distances between clusters

- Minimum cluster distance
  - Minimum of all possible pairs D(A_i, B_j)
- Maximum cluster distance
  - Maximum of all possible pairs D(A_i, B_j)
- Average cluster distance
  - Average of all possible pairs D(A_i, B_j)
- Centroid cluster distance
  - Distance between center of cluster A and B

Hierarchical clustering

- Two types
  - Agglomerative
  - Divisive
- Agglomerative
  - Step 1 : Start with "n" clusters (each record = cluster)
  - Step 2 : The two closest records are merged into one cluster
  - Step 3 : The two clusters with the smallest distance are merged i.e. either single records are added to existing clusters or two existing clusters are combined

Dendogram



© Pradeep Pendem

(Dis)Advantages of Hierarchal Clustering

- Does not require to specify the number of clusters
- Pictorial representation by Dendogram
- Easier to interpret and purely data-driven
- For large datasets, the algorithm is expensive and slow
- One pass : once records are allocated to a group cannot be re-allocated
- Re-ordering or dropping the data can lead to different solution
- Sensitive to outliers

k-means Clustering

- Specify the number of clusters, *k*

- Process

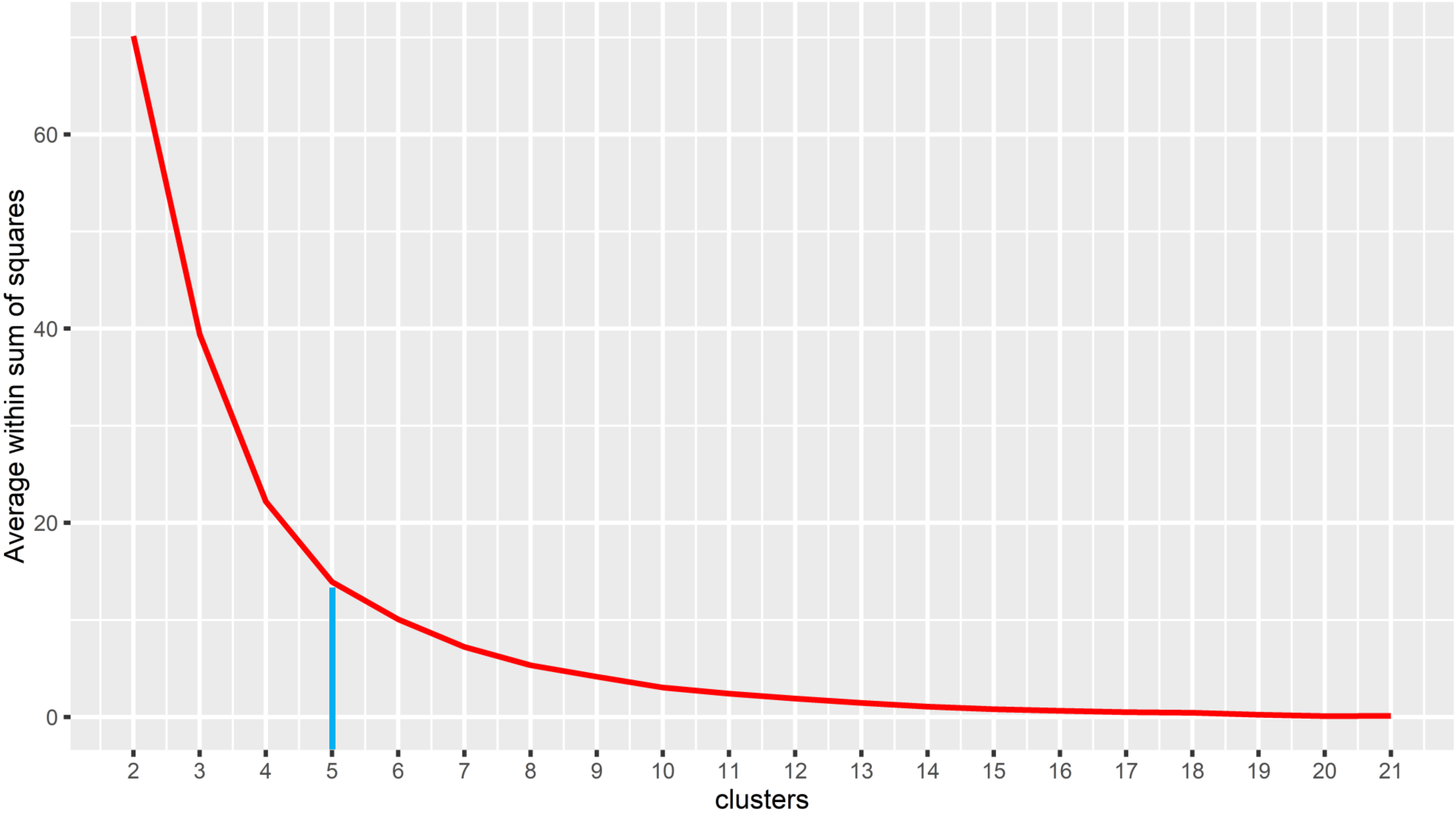
  - Step 1 : Start with "*k*" initial clusters

  - Step 2 : Each record is reassigned to the cluster with the "closest" centroid

  - Step 3 : Recompute the centroid of clusters that lost or gained a record, and repeat Step 2

  - Stop when moving any more records between clusters increases cluster dispersion

Choosing the number of clusters (k)



© Pradeep Pendem

Final Project presentation

- Presentation (10%)
  - 15-minute presentation followed by a 10-minute Q&A
  - **May 31st (Tue) & Jun 02nd (Thu)**
  - Groups are randomly assigned to the 2 days
  - Groups should send the ppt file by 8 am on their presentation date
  - Each member of the group should **mention the contribution** of their work in the last slide of the presentation file
- **Everyone** must be present in the class on the presentation days
  - Zero scores for presentation assessment if absent

May 31st presentations

- ACB
- ATJ
- HJJ
- P

Jun 02nd presentations

- AJA
- DJK
- MRV
- TAP

Final Report

- Formal report
  - Introduction, Problem description, Approach (Regression / Classification)
  - Data Analysis, Results, Inference
  - Conclusions, recommendations
- Regression : *k*-NN as Regression, Linear Regression & Regression Tree
- Classification : *k*-NN as classification, Logistic Regression & Classification Tree
- Assess the performance & recommend best predictive model
- 8-10 pages including any tables and graphs (excluding code)
- Two or Three key insights from the entire analysis
- Submit the code with comments at end of the report
  - 10 of 30 points penalty on not submitting the code

Final Project

- Final Report (30%)
  - Due by **Jun 08th, 8AM** (Exam day) for all groups
  - Each member of the group should **mention the contribution** of their work in the report

Thank You