London Business School

Cluster Analysis

AM04 MAM - Data Science for Business

Dr Kanishka Bhattacharya

Welcome back to Data Science

- Thank you for being early.
- Next 5 sessions I intend to teach from LT including workshops

Zoom classroom etiquette

- Please turn on your cameras and mute your microphone.
- Use chat when instructed, otherwise please raise your virtual hand if you have questions.

Session plan

- Part 1: Unsupervised learning and clustering
- Part 2: K-means clustering
  - Mini Workshop
- Part 3: Interpreting the results of clustering methods

Course contents

Session 6:	• Unsupervised learning: clustering
Session 7:	• Workshop - Clustering • Assignment 3 (group), due by Tuesday 16th Nov
Session 8:	• Contemporary data mining algorithms • Nearest neighbors • Tree based learning
Session 9:	• Ensemble methods
Session 10:	• Workshop: Case competition • Assignment 4 (individual), due in 2 weeks after Session 10: Ensemble Methods

Lectures and assignments

Policies

Late submissions will not be accepted

You should submit your report, and rmd and html files

If your rmd does not knit it means there's something wrong with it.

We will not grade code that does not run.
No html file=no grade
Answer questions and present your argument clearly

Office hours

Will be announced

Class slides (with notes) and rmd files will be posted on canvas

Plagiarism declaration

How to get the most from this class

Read learning outcomes

- Make sure you can answer all questions at the end of each session
- Ask at the end of session if not

Read cases before coming to class

Embedded in html files

Run the code on your own

Try to answer questions in HTML files

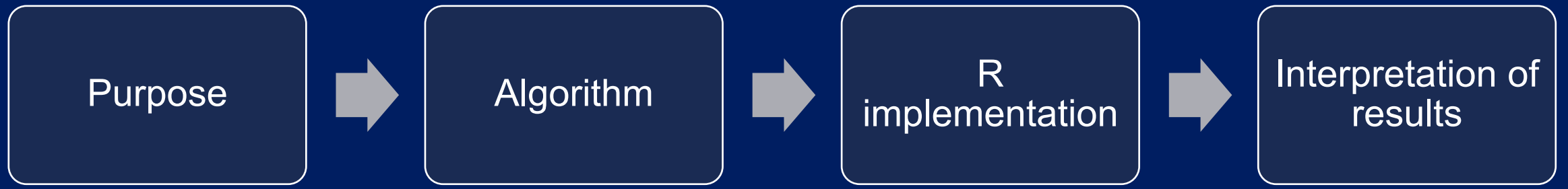
Office hours

Make sure to take full advantage

Assignments

Check the rubric to ascertain you understand what is expected

How to get the most from this class



Assignments

Technical reports

Check the technical writing guide

- How to write technical reports
- Simple word template

Guidance

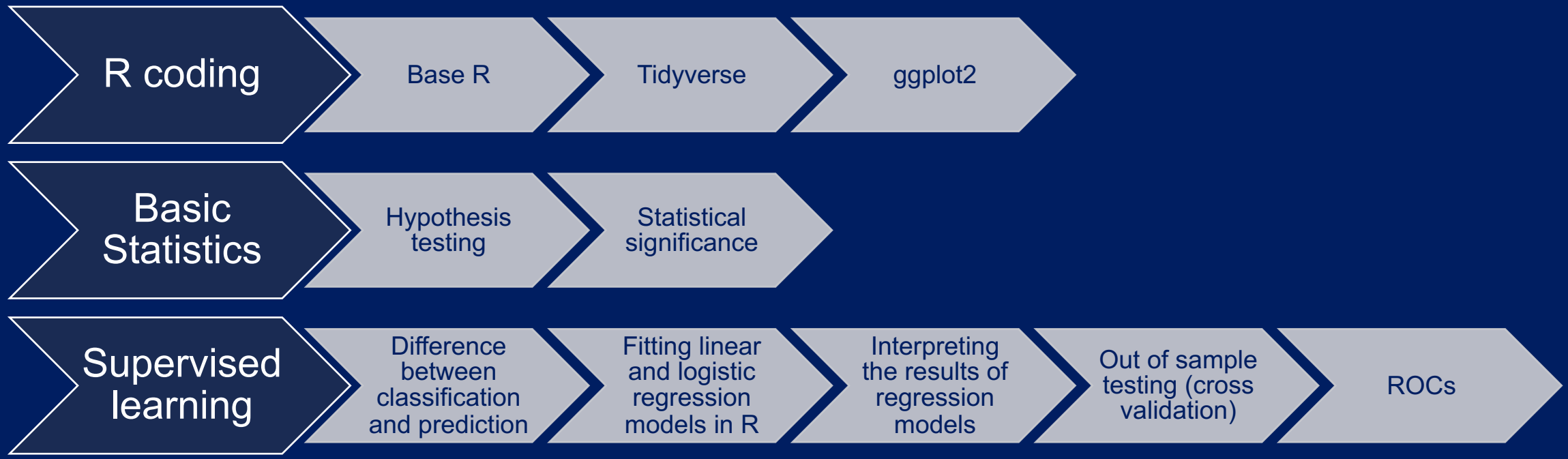
Check learning outcomes for each assignment

Follow instructions where available

Check the deadlines

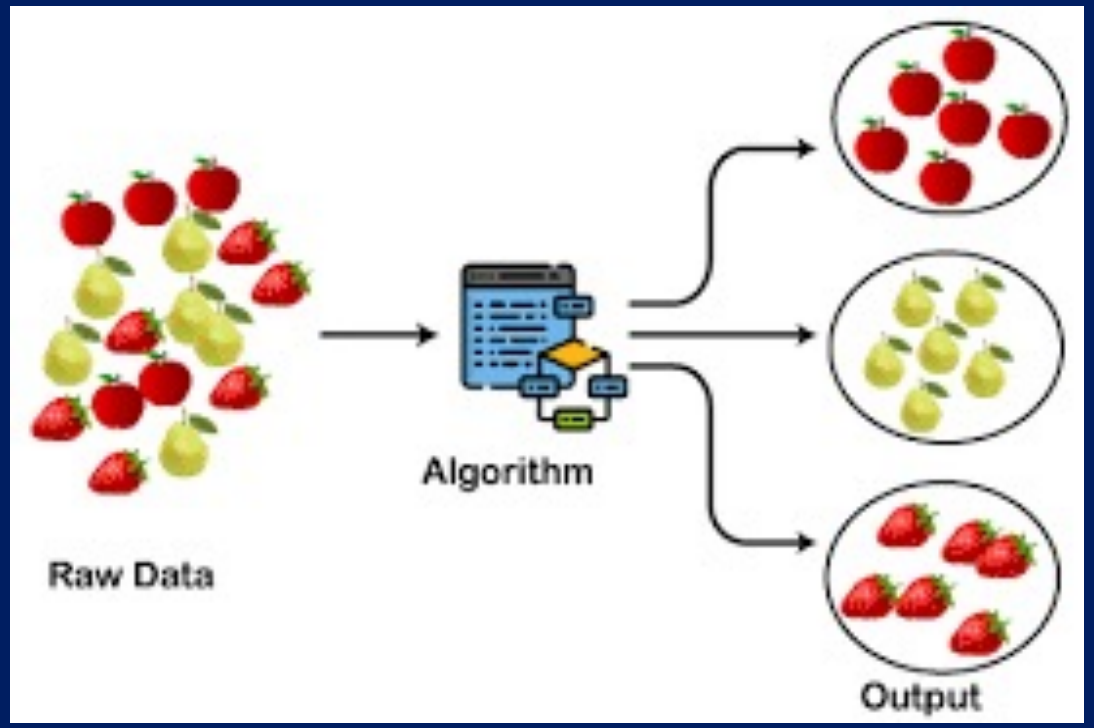
Apply for EC if you cannot complete them on time

Assumed prior knowledge



Today's agenda

- What is clustering (or cluster analysis)?
  - Task of grouping a set of objects in such a way that similar objects are in the same cluster
- Unsupervised learning
  - Inferring knowledge from "unlabeled" data
- Clustering algorithms
  - Distances between observations
  - K-Means clustering
  - Evaluation of clustering results
  - K-Medoids clustering (next week)
  - Hierarchical clustering (next week)
- Summary (revisit learning outcomes)




MAIN TYPES OF CLUSTERING METHODS
01
Partitioning Methods
02
Hierarchical Clustering
03
Fuzzy Clustering
04
Density-Based Clustering
05
Model-Based Clustering

Unsupervised Learning

- Cluster analysis is a type of unsupervised learning
- Unsupervised learning: Draw inferences from datasets consisting of input data without labeled responses.
- It is more subjective than supervised learning
  - There is no simple goal for the analysis, such as prediction of a response.
  - There are many cognitive issues that can affect the results. Examples
    - Confirmation bias
    - Texas sharpshooter fallacy
    - Bandwagon effect and etc.
  - I will try to make it as objective as possible
  - Evaluation of clustering
    - Determining the number of clusters: We don't even know how many clusters are there in the data
    - Measuring clustering quality: There are different ways to measure clustering quality based on distance measures.
    - Why don't we use out-of-sample testing?

Motivation

- Example: Obama election campaign
  - Identify target groups:
    - Goal: Registration, persuasion, and turnout.
  - Each target group has different characteristics
    - Identify democrat leaning group and convince them to vote
      - People were sent messages on Facebook to click a button to automatically urge those targeted voters to take certain actions, such as registering to vote, voting early or getting to the polls.
      - The campaign found that roughly 1 in 5 people contacted by a Facebook pal acted on the request, in large part because the message came from someone they knew.
    - Identify undecided and convince them to vote democratic
  - Identify subgroups within these groups based on their demographic info and produce targeted marketing campaigns.
    - They were able to put their target voters through some really complicated modelling, to say, if Miami-Dade women under 35 are the targets, [here is] how to reach them."
    - For example they noticed that George Clooney had an almost gravitational tug on West Coast females aged 40 to 49. How do you think they used this information?

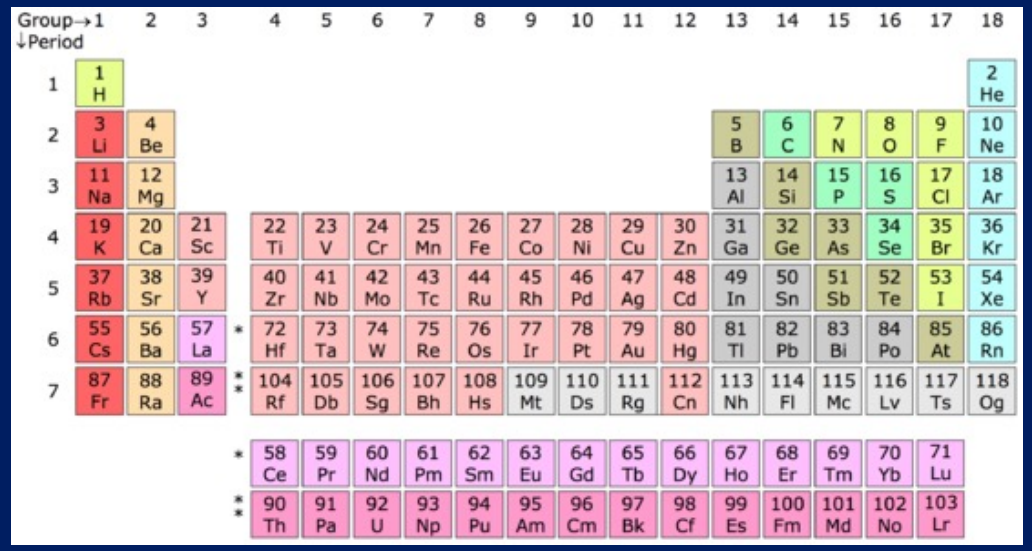
How Obama's Team Used Big Data to Rally Voters

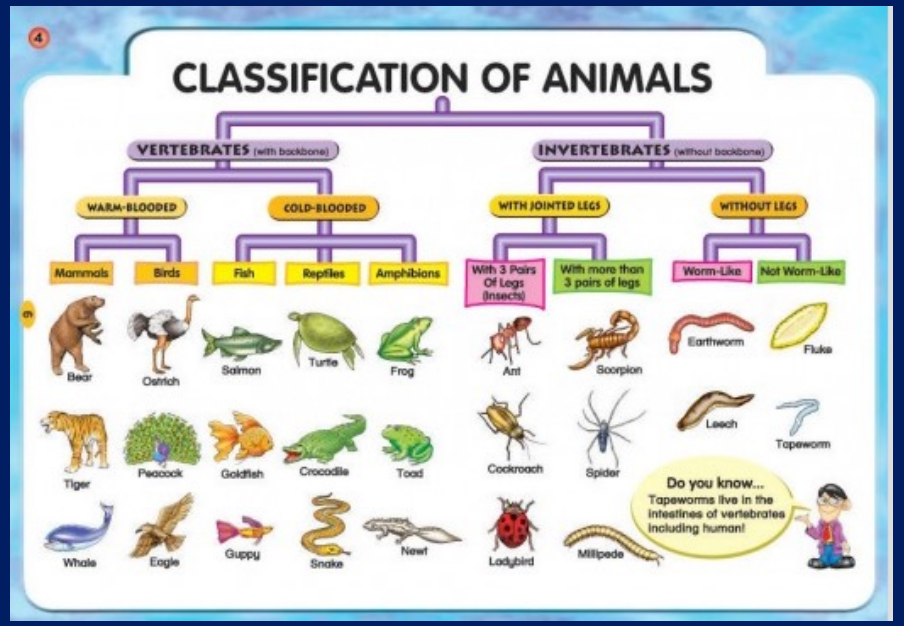
How President Obama's campaign used big data to rally individual voters.

by **Sasha Issenberg** December 19, 2012

Clustering

- Clustering is the process of grouping a set of data objects into multiple groups or clusters so that objects within a cluster have high similarity but are very dissimilar to objects in other clusters.
- Dissimilarities and similarities are assessed based on the attribute values describing the objects and often involve distance measures.
- Other applications:
  - Chemistry: Periodic table
  - Biology: Taxonomy



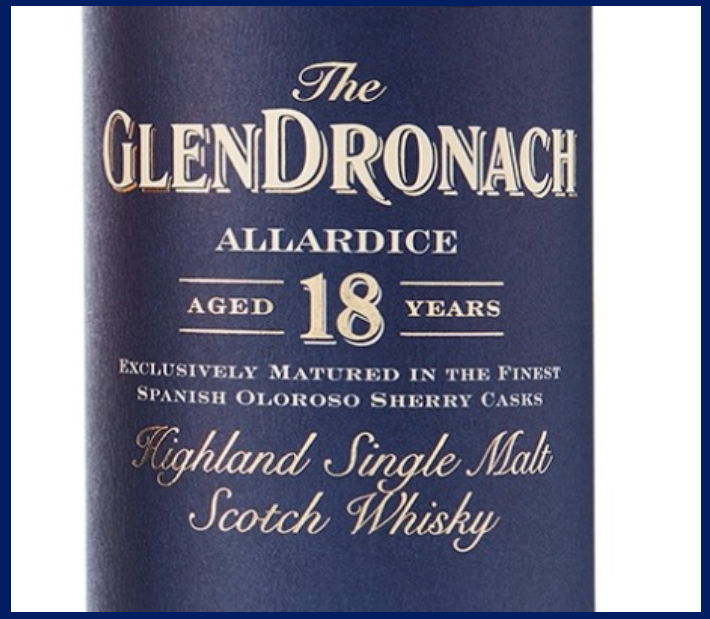


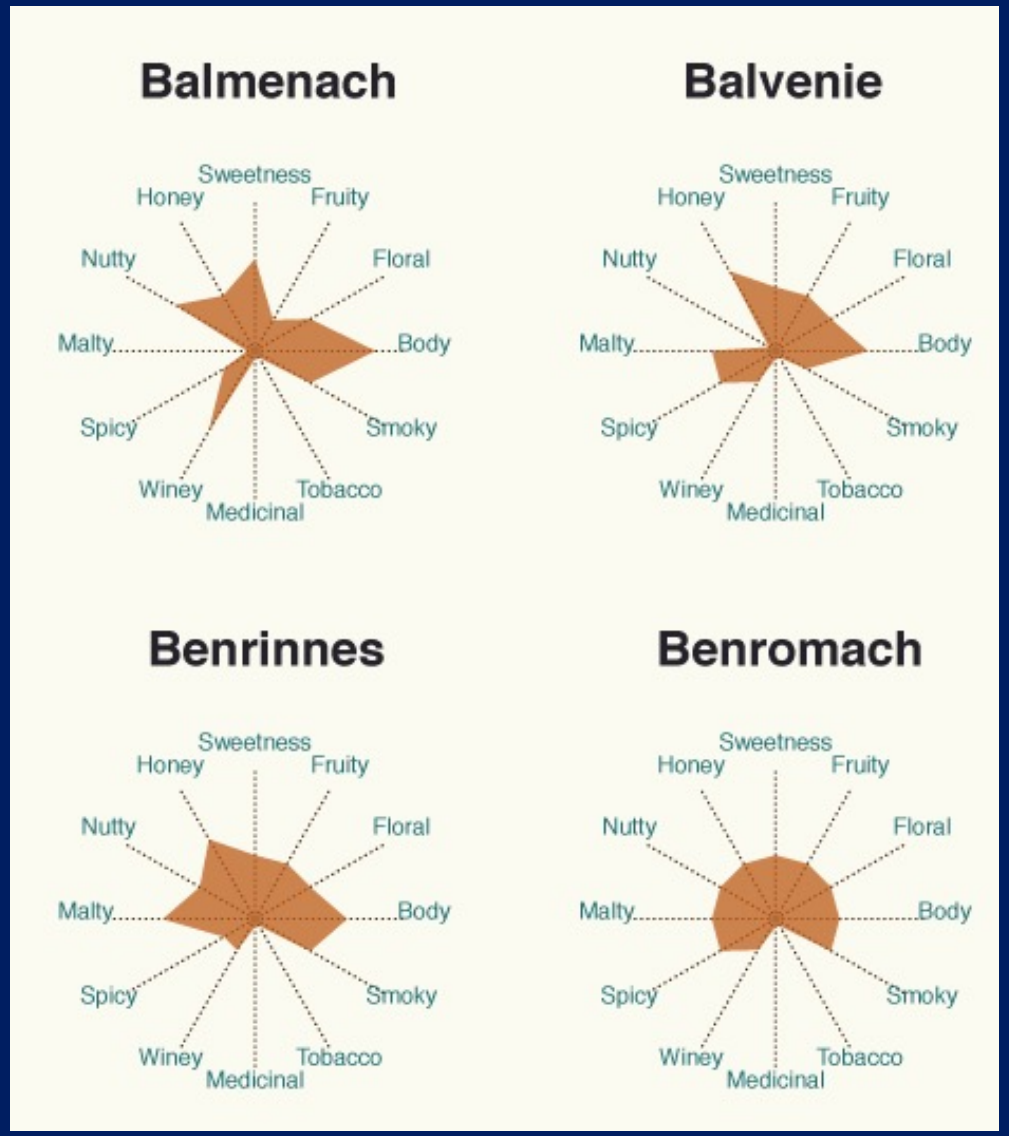
Motivation

- Marketing
  - A large set of data is available for income, occupation and so on for a large number of people
  - Market segmentation: Identify subgroups of people who might be more receptive to a particular form of advertising or more likely to purchase a particular product.
  - The task of performing market segmentation amounts to clustering the people in the data set

Whisky analytics

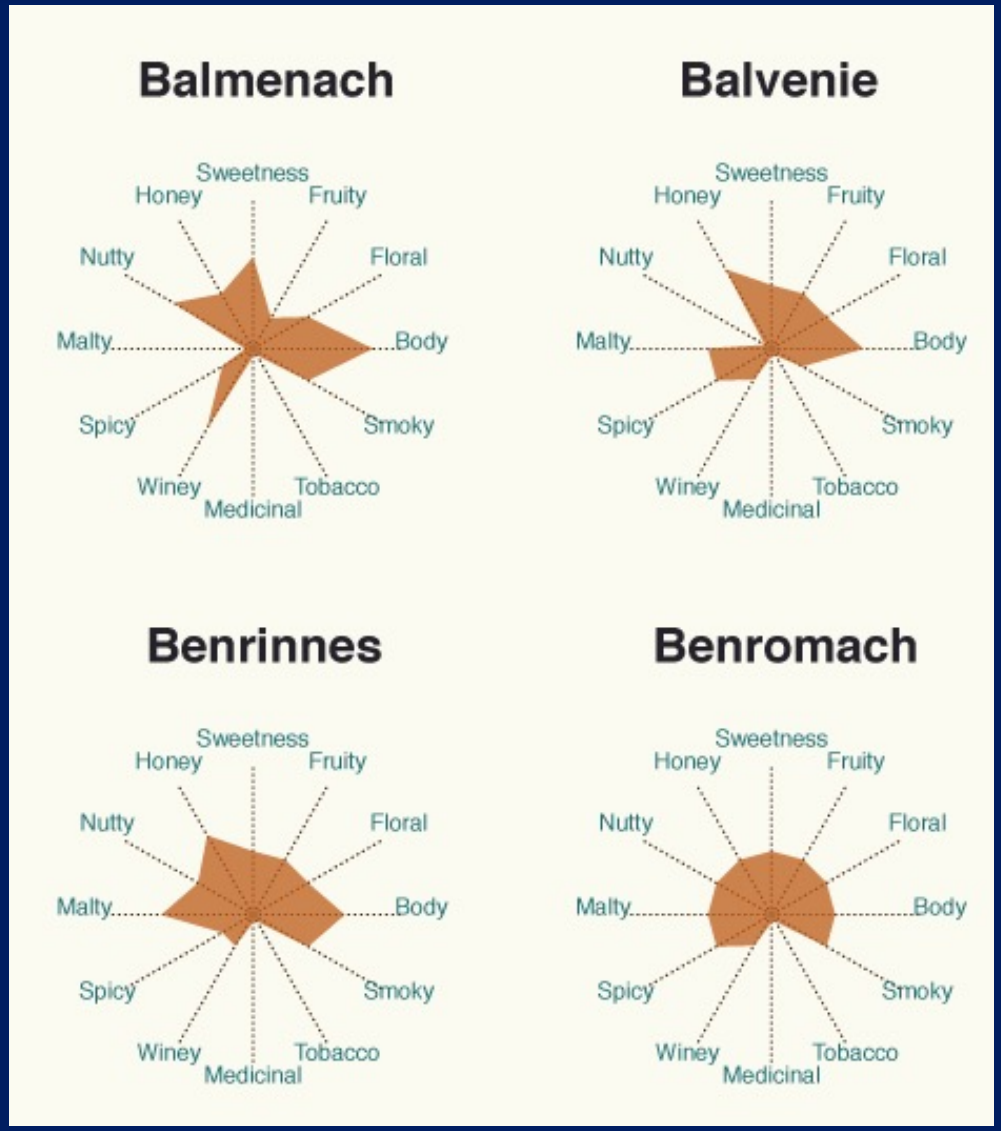
- What are the major types of single-malt whiskies that can be recognized?
  - What are their chief characteristics and the best representatives?
- What is the geographic component in that classification?





Whisky analytics

- Goal: Identify clusters in whiskies from 86 distilleries based on tasting notes
- Techniques: Use K-means, Partitioning around mediods (PAM), and Hierarchical clustering methods
- Applications: Recommendation engines and hypothesis “testing”.



Clustering algorithms

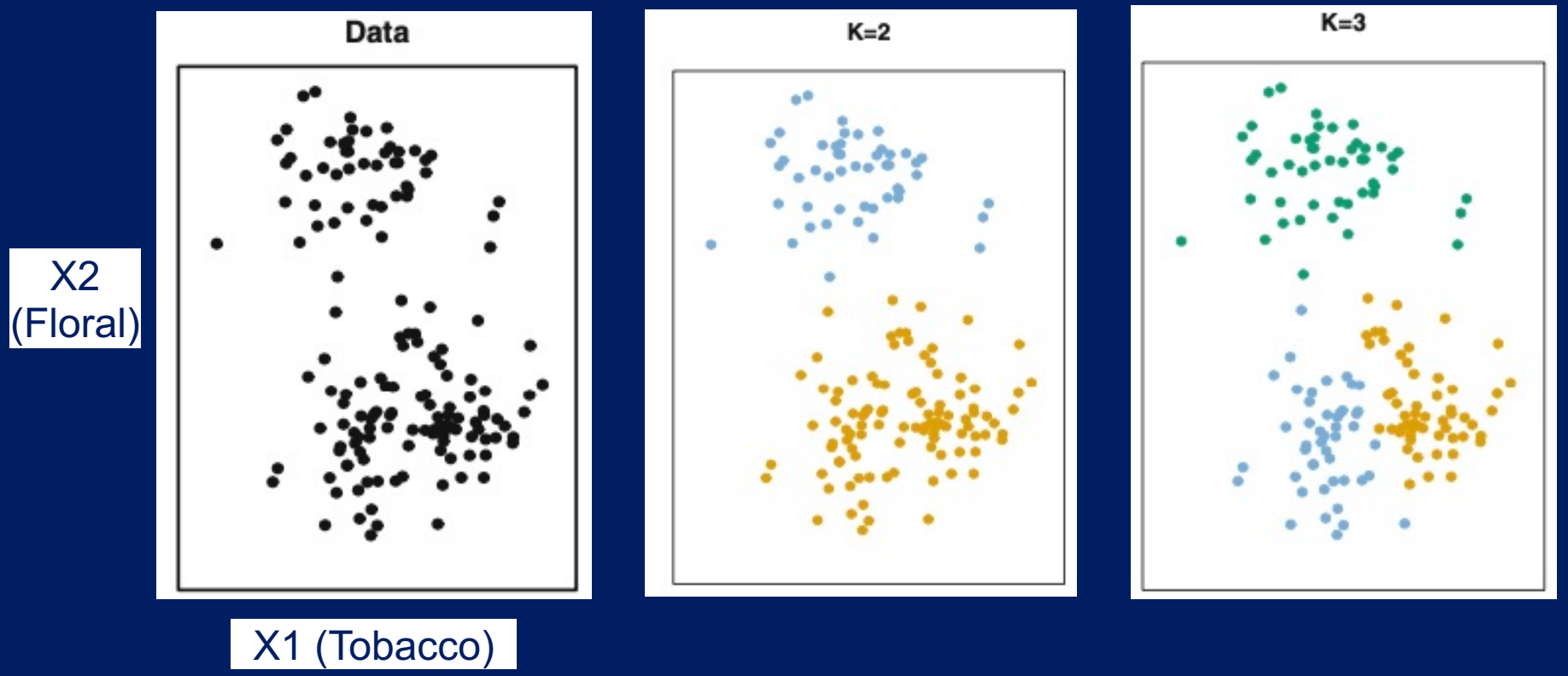
- Partitioning methods: K-means clustering
  - Partition a set of objects into k clusters in the best way possible.
  - We seek to partition the observations into a pre-specified number of clusters
  - Iterative approach to determine the “best way” possible.
- Partitioning Around Medoids (PAM)
- Hierarchical clustering
  - Start with all observations in separate clusters and successively merge the objects or groups close to one another
  - Evaluate the clusters by looking at the groups at different stages of the process(graphical approach)

K-Means Clustering

- In K-means clustering, we seek to partition the observations into a pre-specified number of clusters
- For example consider the following synthetic data points. (see next slide)
  - Each data point has two attributes x1 and x2
  - If I want to have two clusters the figure in the middle is what I end up with
  - If I want three it is the one on the right
- Let's focus on K=2. How do you think we should divide the points between two clusters?
  - We want disjoint sets; a point can belong to only one set
  - We want all points covered; each point has to belong to a point
  - We want the points to in the same cluster as similar as possible
  - How do I measure similarity? We use within cluster variation.

K-Means Clustering

- We can formulate this as an optimization problem for fixed K.
- Minimize within cluster variation
- Subject to: All points should be assigned to one and only one cluster



Optimization formulation

K-Means Clustering

- How do we define within cluster variance?
- We use the sum of squared error for each observation.
  - How much error would I make if I estimate each variable of an observation by the average of these variables among all the observations in the same cluster?
  - Assume that the following observations are assigned to the same cluster

Whisky	Tobac	Floral
a	3	1
b	5	3
Average	4	2

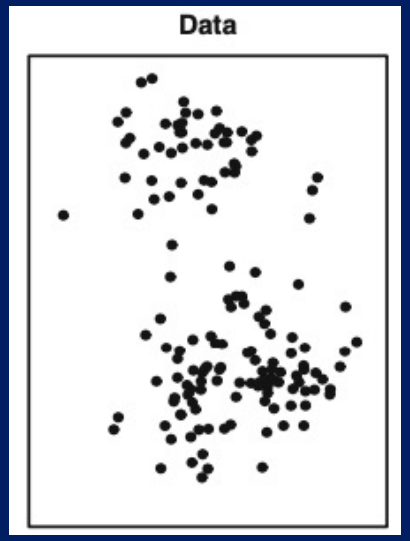
  - Total variation in this cluster = $(3-4)^2 + (5-4)^2 + (1-2)^2 + (3-2)^2$
    =4
  - What happens when we assign another observation to this cluster?
    - The mean changes hence the variance

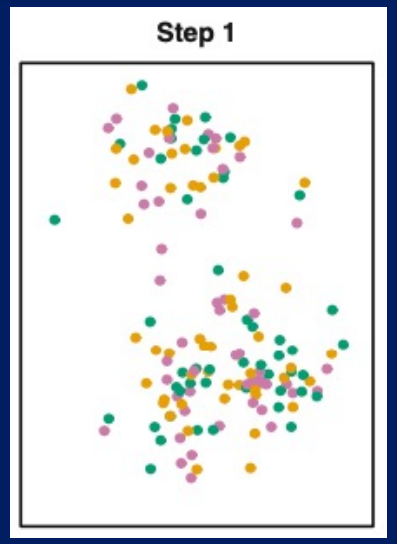
K-Means Clustering Algorithm

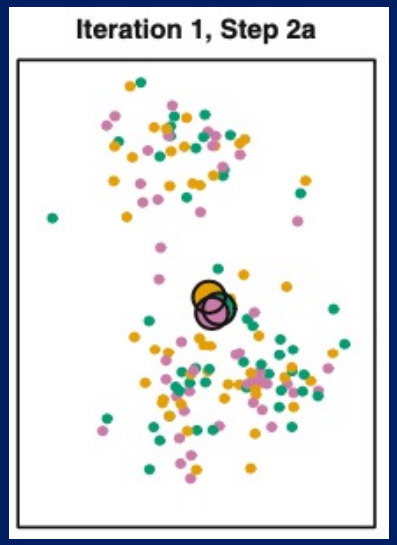
- The clustering problem becomes an optimization problem
  - Minimize total within cluster variation (K is fixed)
  - Subject to: All points should be assigned to one and only one cluster
- It turns out this optimization problem is too difficult to solve in a reasonable time, hence we have an algorithm which tries to find good solutions.
- K-Means clustering algorithm
  1. Choose # of clusters desired, *k*
  2. Start with a partition into k clusters (Often based on random selection of k centroids)
     - 2a. Re-compute centroids (average for each variable within a cluster)
     - 2b. At each step, move each record to cluster with closest centroid
  3. Stop when moving records increases within-cluster dispersion

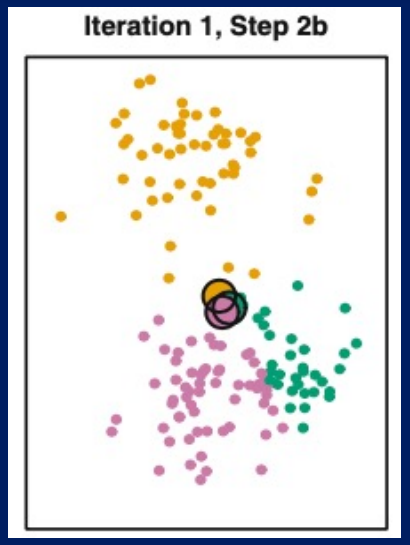
     See next slide for a demonstration of this algorithm.

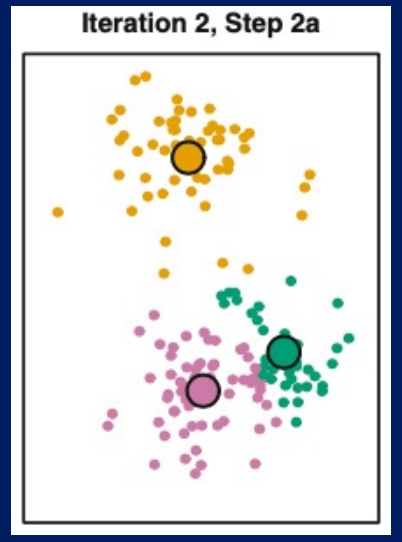
K-Means Clustering Algorithm



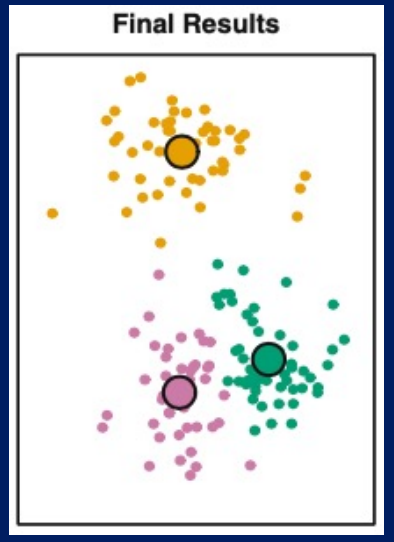








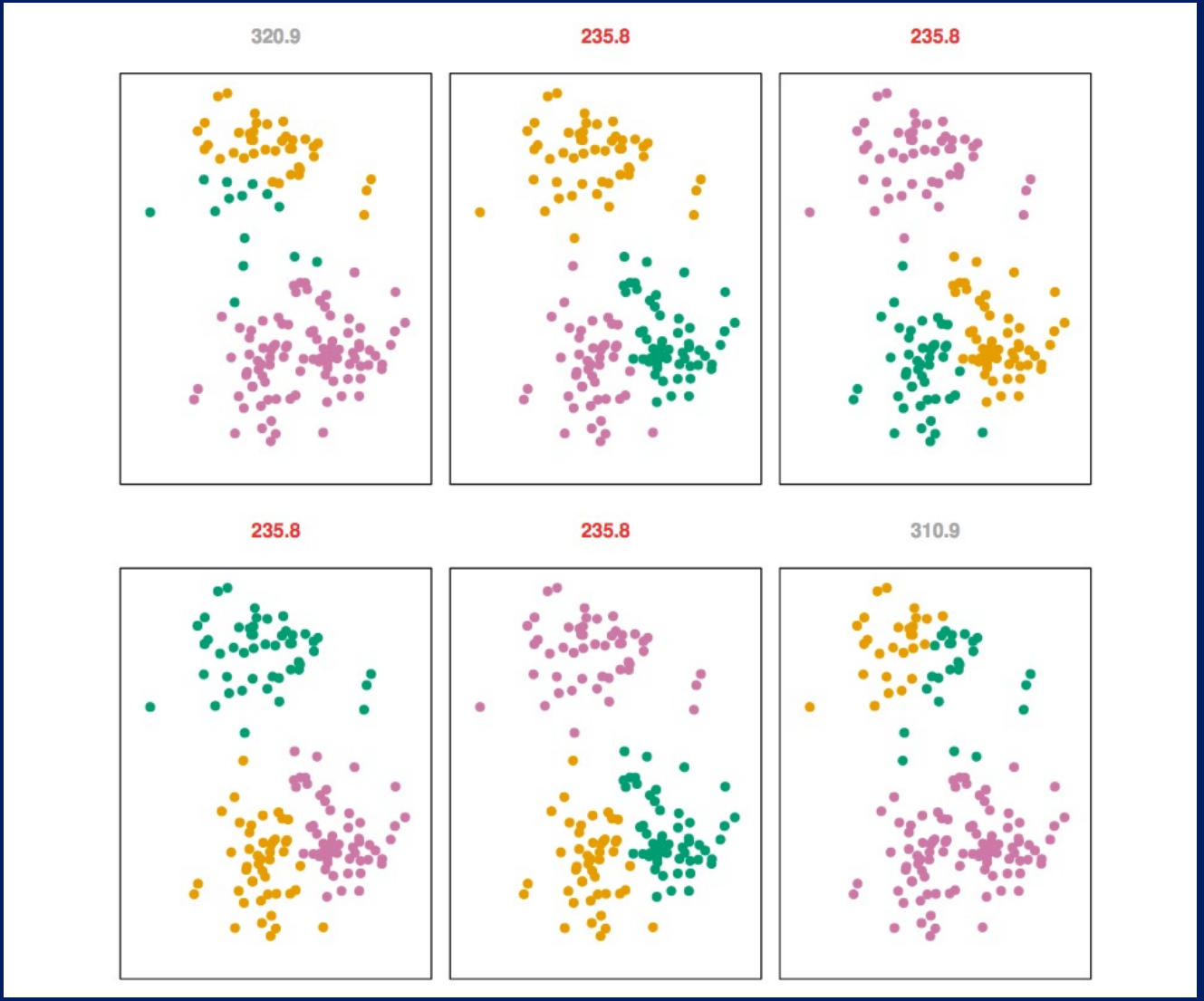
.....



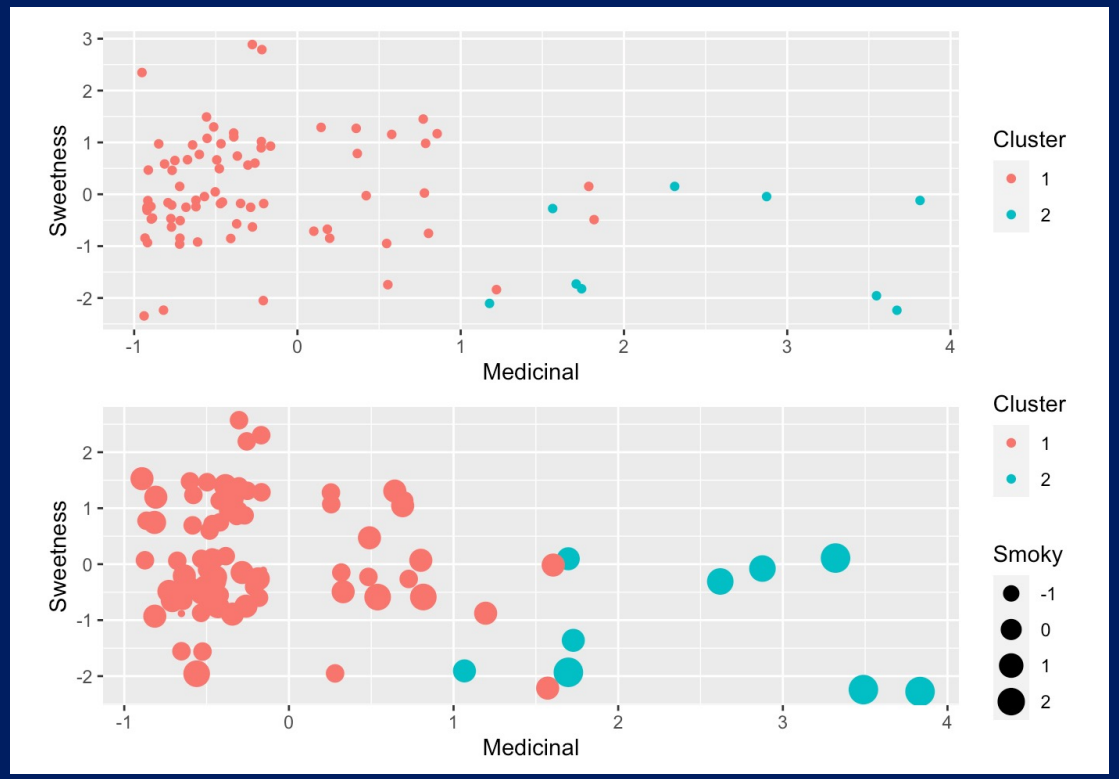

GREAT SUCCESS!

Example

Results of K-Means with six different starting points



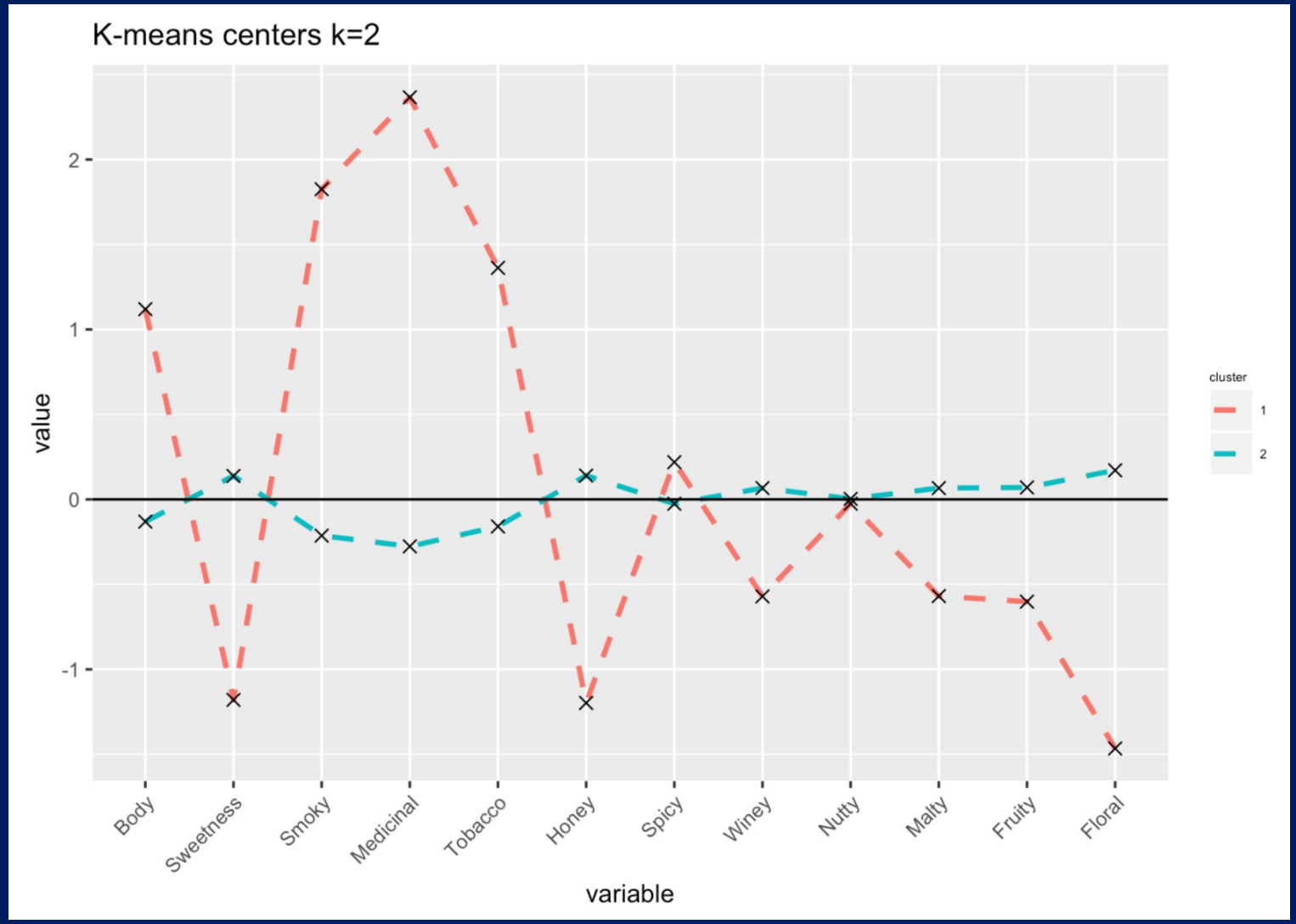
Example: Whisky Data



- Visualizing based on features
- Note, the small sample size for one group will not give us reliable clusters
- There will always be border cases – in some clustering approaches you can have mutually non-exclusive groups

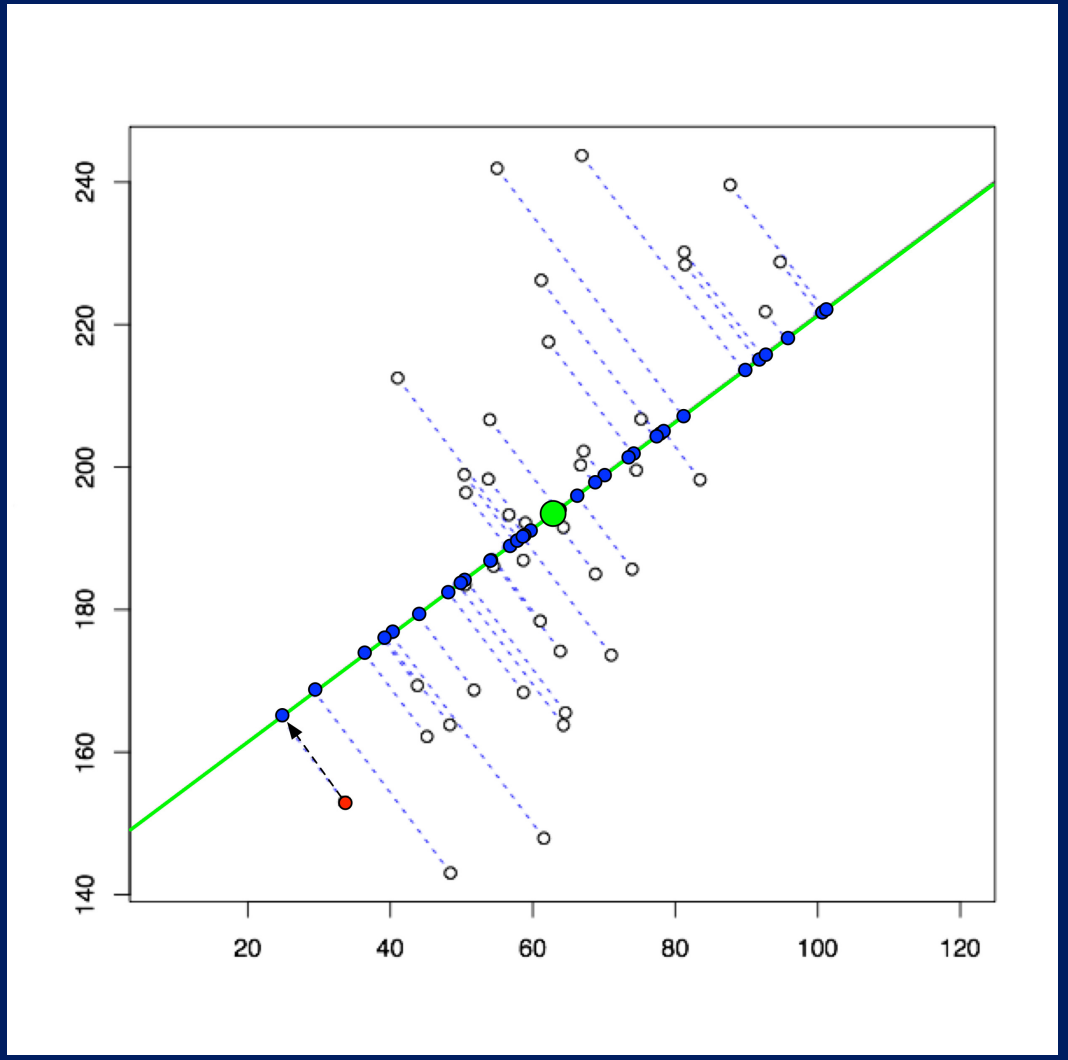
Example: Whisky Data

Visualizing based on centers



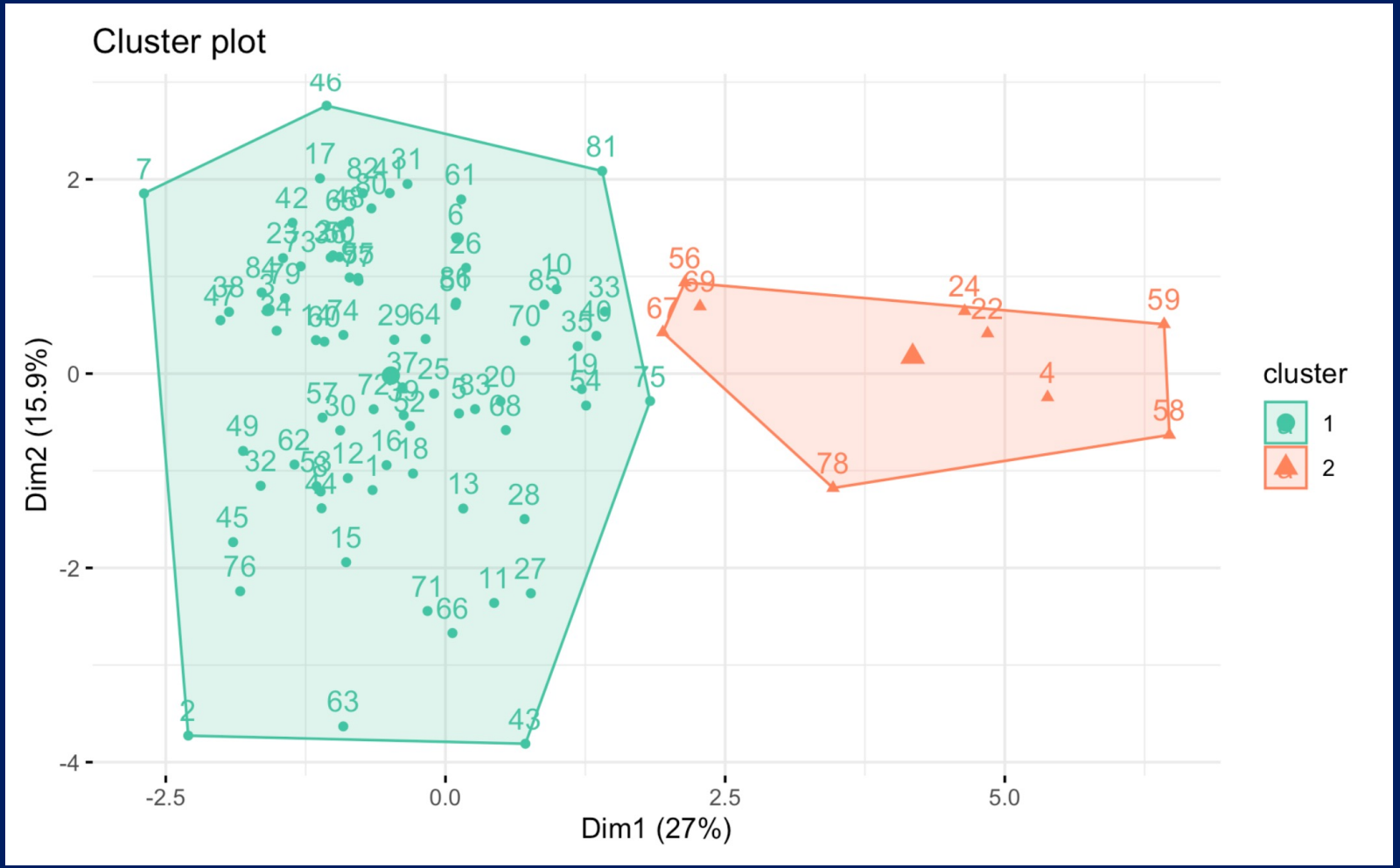
Example: Whisky Data

Visualizing Principal Components



Example: Whisky Data

Visualizing Principal Components



Validating Clusters: Interpretation

- **How can we evaluate whether the clustering results are good?**
- **Goal:** obtain meaningful and useful clusters
  - Obtain summary statistics
- Example: Clustering for marketing
  - Identifiability: refers to the extent that managers can recognize segments in the marketplace
  - Sustainability: criterion is satisfied if the segments represent a large enough portion of the data to ensure profitable customization of the marketing program.
  - Accessibility: The extent to which managers can reach the identified segments through their marketing campaigns is captured
  - Actionability: whether customers in segment and the marketing mix necessary to satisfy their needs are consistent with the goals and core competencies of the firm
- **Caveats:**

  (1) Random chance can often produce apparent clusters

  (2) Different cluster methods produce different results

Visualizing clusters: Summary

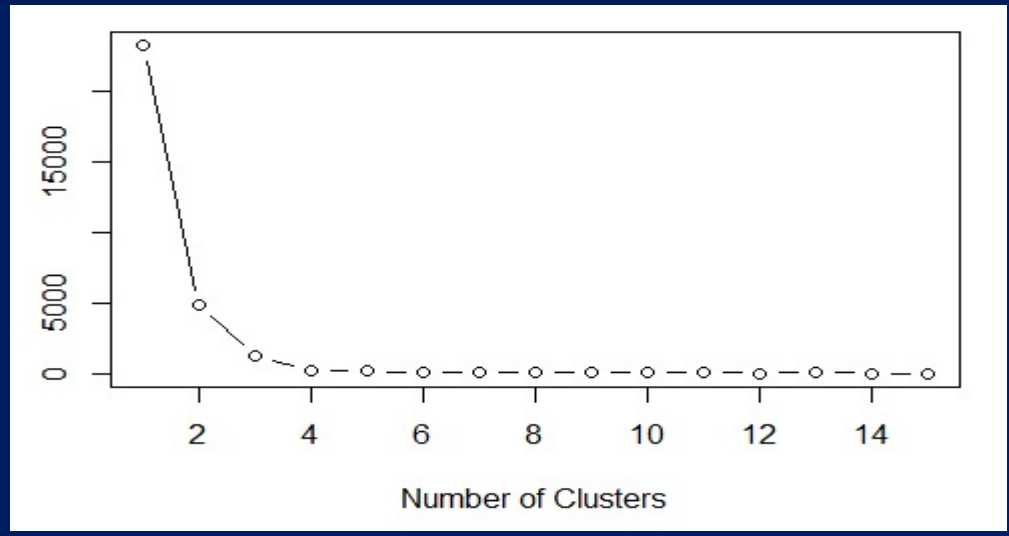
- We can check how points are clustered by plotting locations of each observation vs two of the variables (maybe three)
  - If there are many dimensions, this is not very useful
- We can use dimension reduction and plot the points as explained above. But this lacks interpretability
  - PCA will be covered in your Machine Learning Class
  - I will just show you how it plays out in cluster visualization
- We can visualize the centers of the clusters

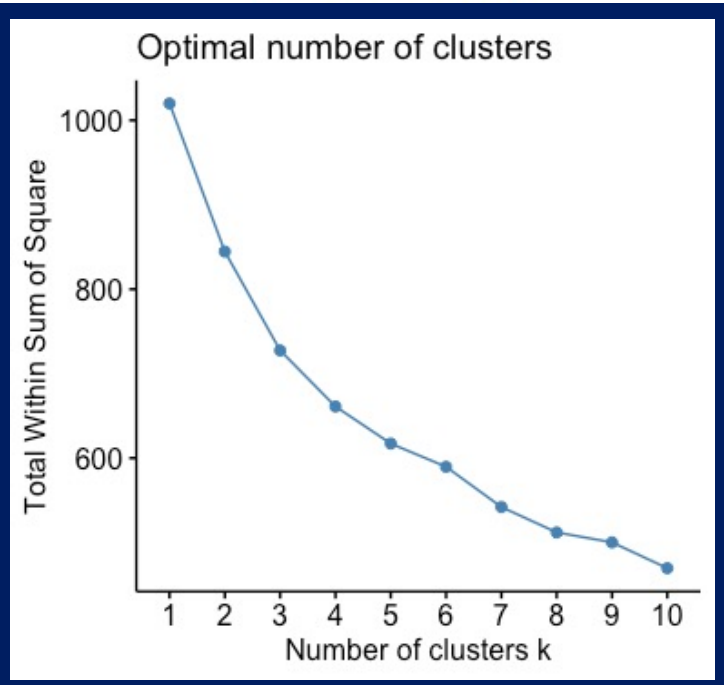
Determining the number of clusters

- Elbow method
- Visualization
- Silhouette method
- NbClust() function: 30 indices for choosing the “optimal” number of clusters

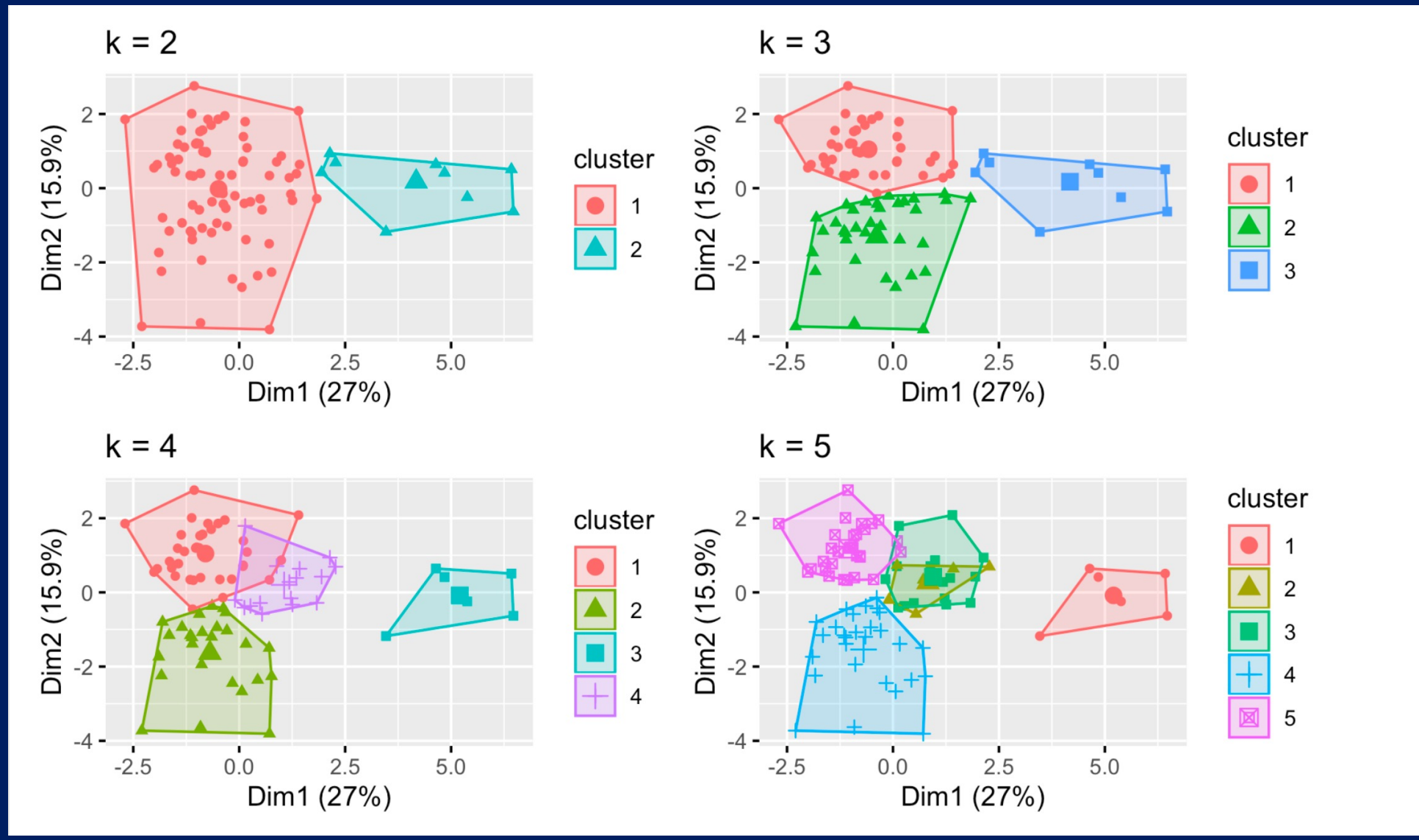
Number of clusters: Elbow plot

- Look for a bend or elbow in the sum of squared error (SSE) scree plot.
- On the example on top plot, 3 or 4 seems to be reasonable
- Can total variation within cluster increase with # of clusters?
- How many clusters do you need in the second case?

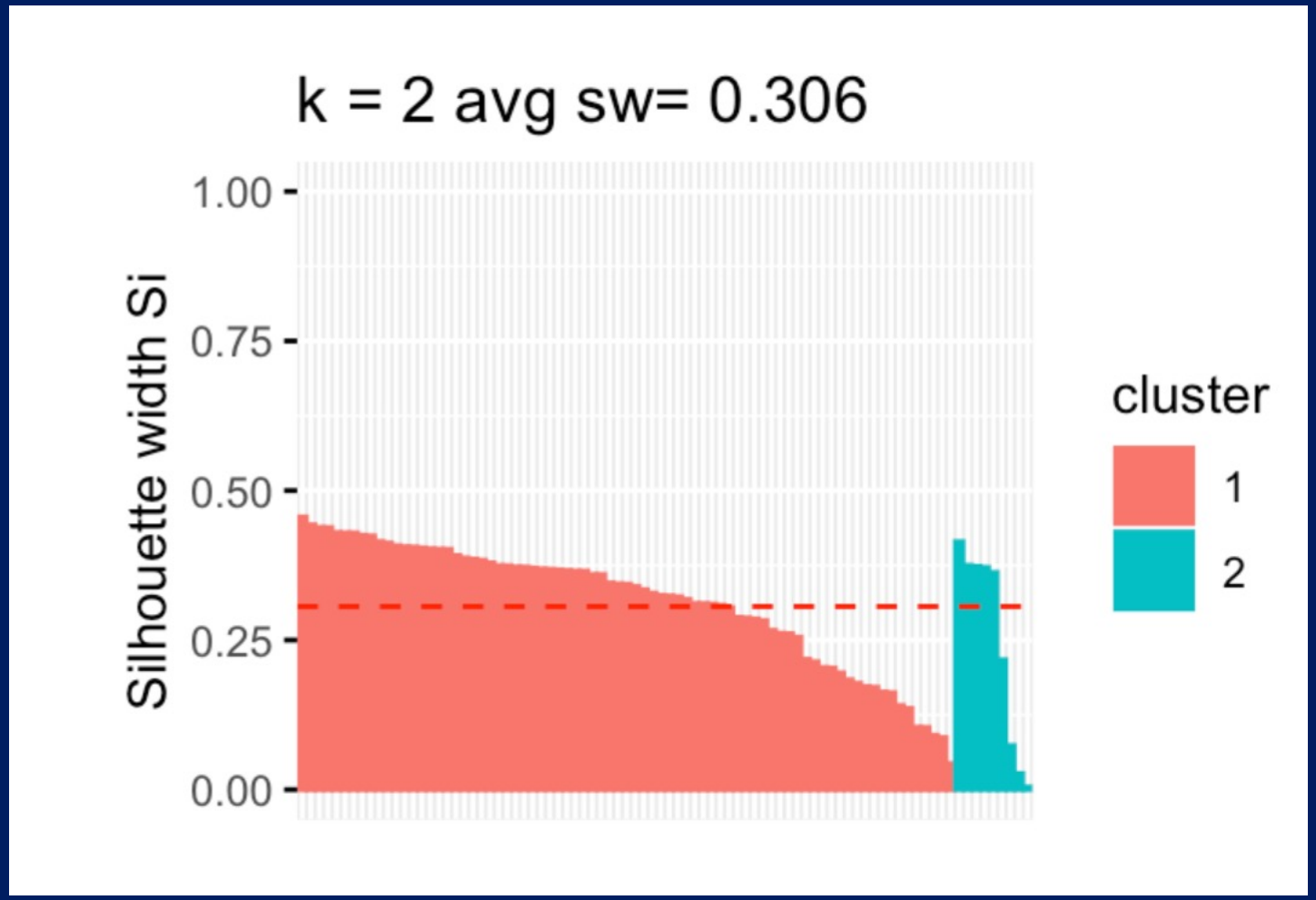




Visualizing clusters



Silhouette Analysis



Silhouette Analysis

- SA refers to a method of interpretation and validation of consistency within clusters of data.
  - Its a neat way to find out the optimum value for k during k-means clustering.
  - The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation).

$$s(i) = \frac{N(i) - C(i)}{\max(C(i), N(i))}$$

    - See next slide for definitions
  - The silhouette ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters.
    - If most objects have a high value, then the clustering configuration is appropriate.
    - If many points have a low or negative value, then the clustering configuration may have too many or too few clusters.
- Mean Silhouette score: Mean score can be simply calculated by taking the mean of silhouette score of all the examples in the data set. This gives us one value representing the Silhouette score of the entire cluster.

Silhouette Analysis

- For each data point $i \in C_i$ data point i in the cluster C_i, define in cluster distance as

$$C(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i,j)$$

- For each data point $i \in C_i$, define distance to the nearest neighbor as

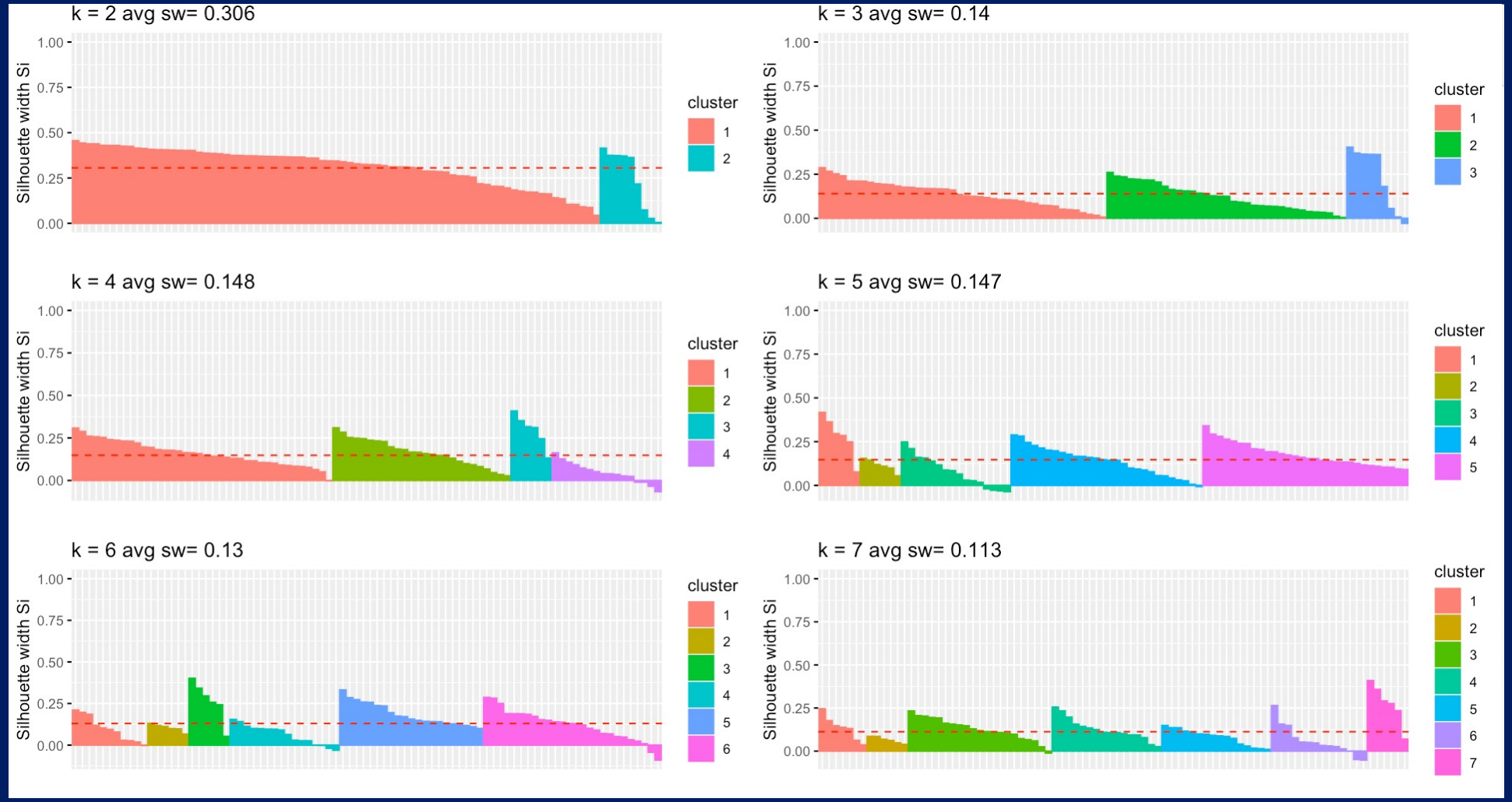
$$N(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i,j)$$

- The silhouette s(i) for point i is

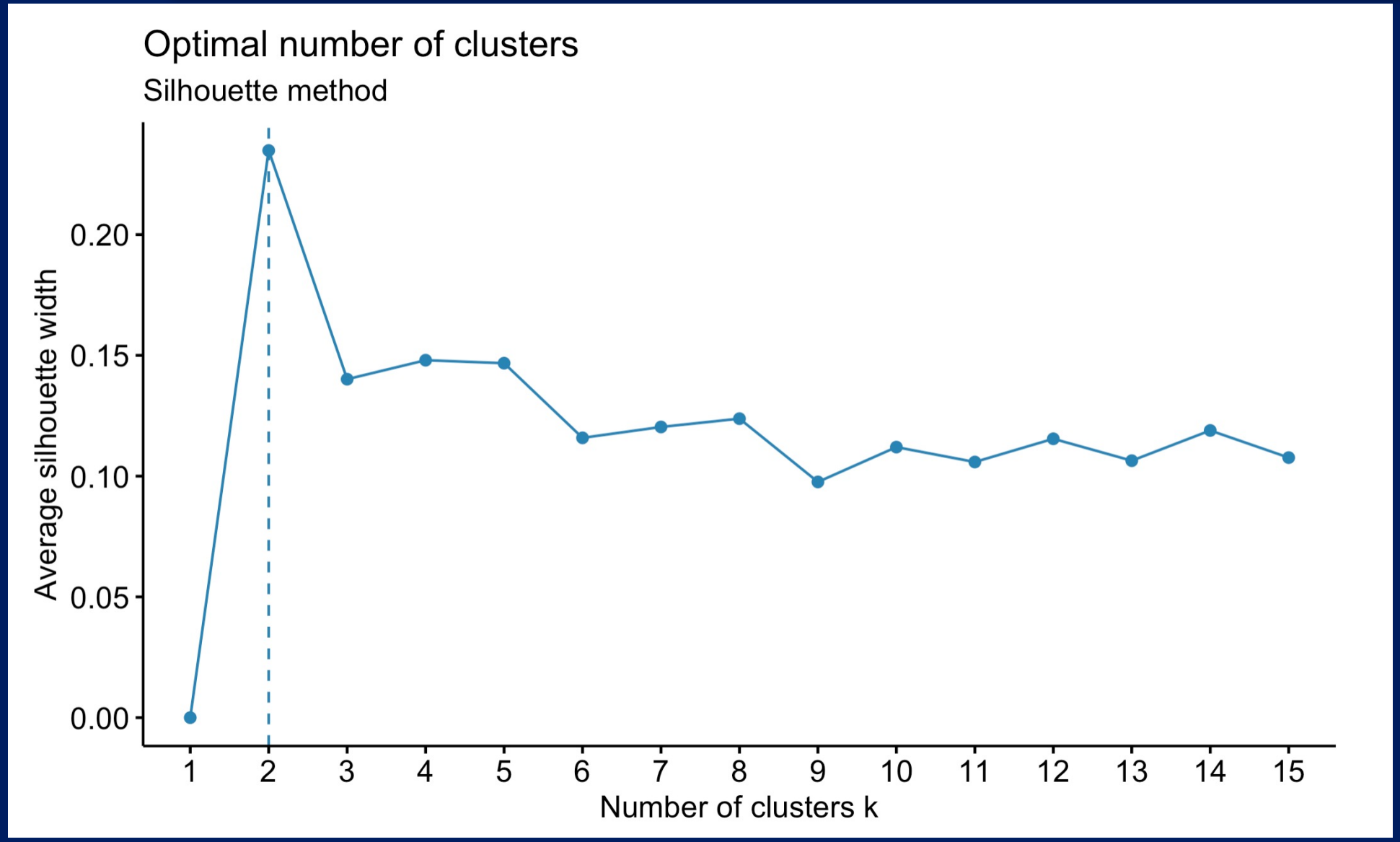
$$s(i) = \frac{N(i) - C(i)}{\max(C(i), N(i))}$$

- C(i) is the average distance of point 'i' with all other points within the same same cluster
- N(i) is the average distance of point 'i' with all points in the closest cluster
- S(i) is the SA coefficient

Silhouette Analysis

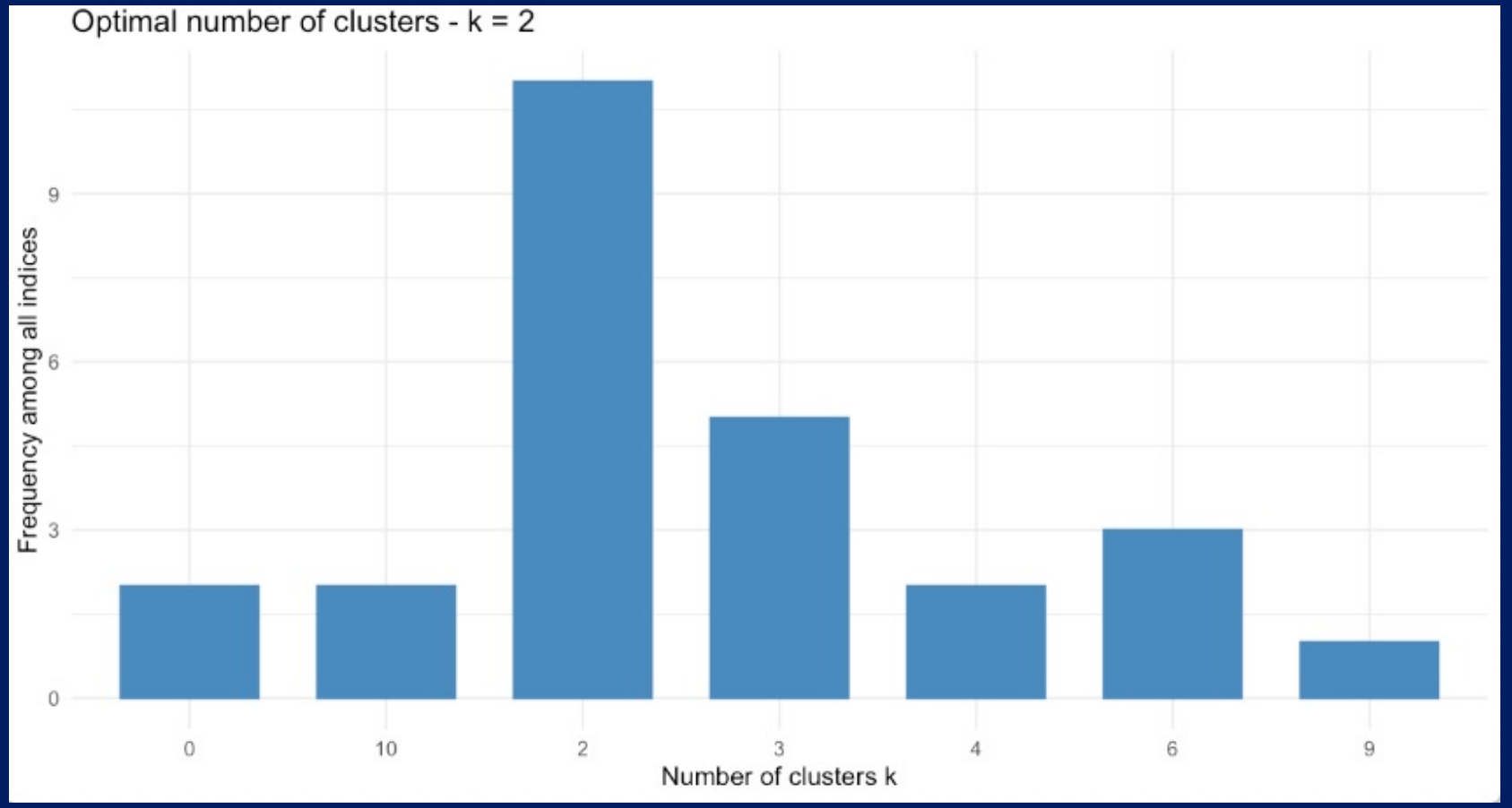


Silhouette Analysis



NBClust

- There are many more
- NbClust package provides 30 indices for determining the number of clusters and proposes to user the best clustering scheme from the different results obtained by varying all combinations of number of clusters, distance measures, and clustering methods.



Clustering steps

1. Once you clean the data start by clustering data into a few clusters, 2 or 3 are good starting points
2. Check how many points are in each cluster
   - Small clusters are usually not very useful
   - Visualize PCA to verify separation
3. Plot the means of clusters to detect differences between clusters
4. Think about variables that may not be very critical for clustering purposes or think about eliminating those that are highly correlated.
5. Increase the number of clusters and check how that changes clustering results
6. Use elbow chart and silhouette analysis (there are also other methods) to verify the number of clusters you are using are not way off
   - Don’t use these methods blindly. Their results only provide guidance but not definitive answers.

Summary

- Cluster analysis is an exploratory tool. Useful only when it produces **meaningful** clusters
- We usually use a few methods to verify the accuracy of our conclusions.
- Be wary of chance results; data may not have definitive “real” clusters- Texas sharpshooter fallacy

Learning objectives

Conclusions

1. What is the difference between supervised and unsupervised learning?
2. What is the main objective of clustering?
3. How does K-Means algorithm work?
   - Objective
   - Inputs
   - Outputs
4. How do we determine how many clusters we have in the data?
   - Elbow chart
   - PCA visualization
   - Comparing clustering results with different clusters
   - Silhouette analysis

K-means Algorithm: Issues

- Results are sensitive to outliers (why?)
  - How to resolve this?
    - Use different distance measures such as absolute error
    - Use the median of observations instead of average.
    - Use K-Medoids (next)
- K is fixed in advance
  - Use elbow charts and see when there is no improvement
- The algorithm may end up in a local optimal
  - Re-start with different initial clusters and see if there is a big difference
- The clusters are expected to be of similar size
  - Use other methods to verify the results (next)