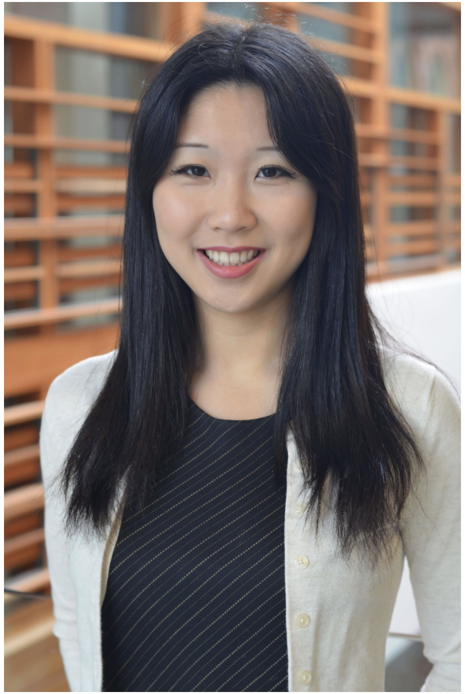
HALE Sports Summer Internship Report

September 23, 2020

Who are we?

Wenjie Gu
CBQG ’21 @ Harvard

Shaoling Han
CBQG ’21 @ Harvard



Jenny Wang
HDS ’21 @ Harvard

What did we do this summer?

- Place the athletes on our platform into distinct groups by seeing if they have similar patterns in their health data, including:
    - Genetic data
    - Clinical lab data
    - Microbiome data
    - Performance data

- Trial and error process: end product came together in several stages

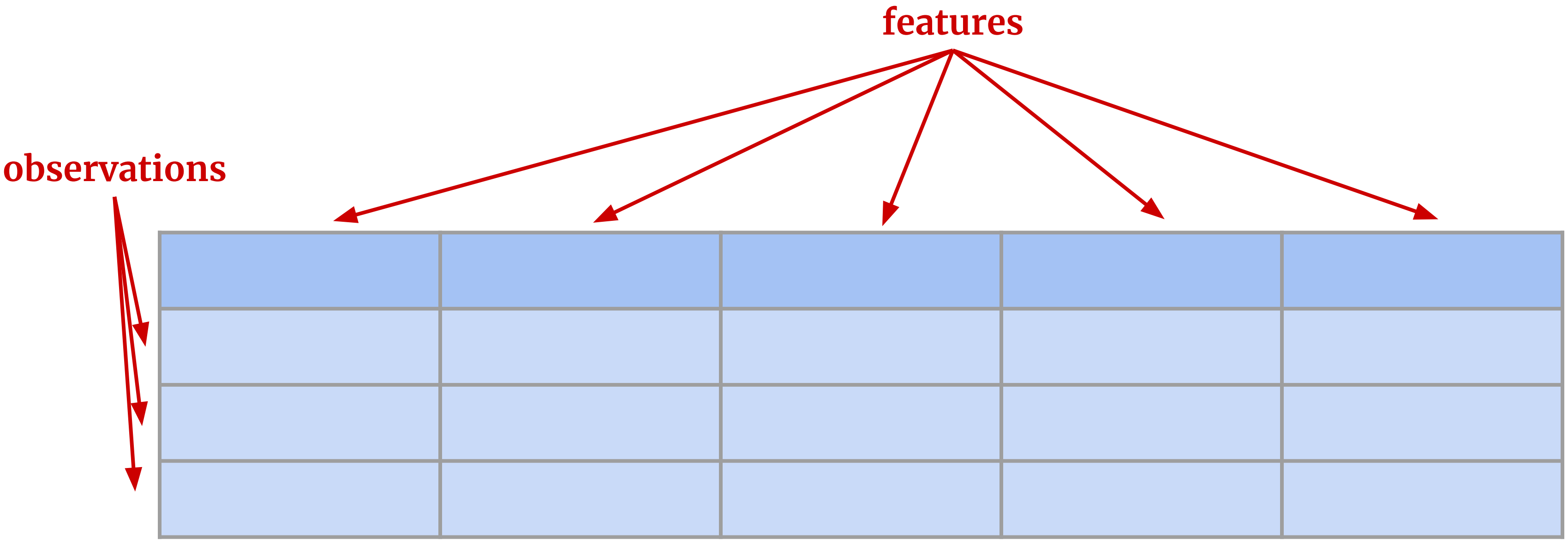
Outline

1. Data table basics
2. Terminology
3. Cluster analysis
4. Subject matter expert report

Outline

1. Data table basics
2. Terminology
3. Cluster analysis
4. Subject matter expert report

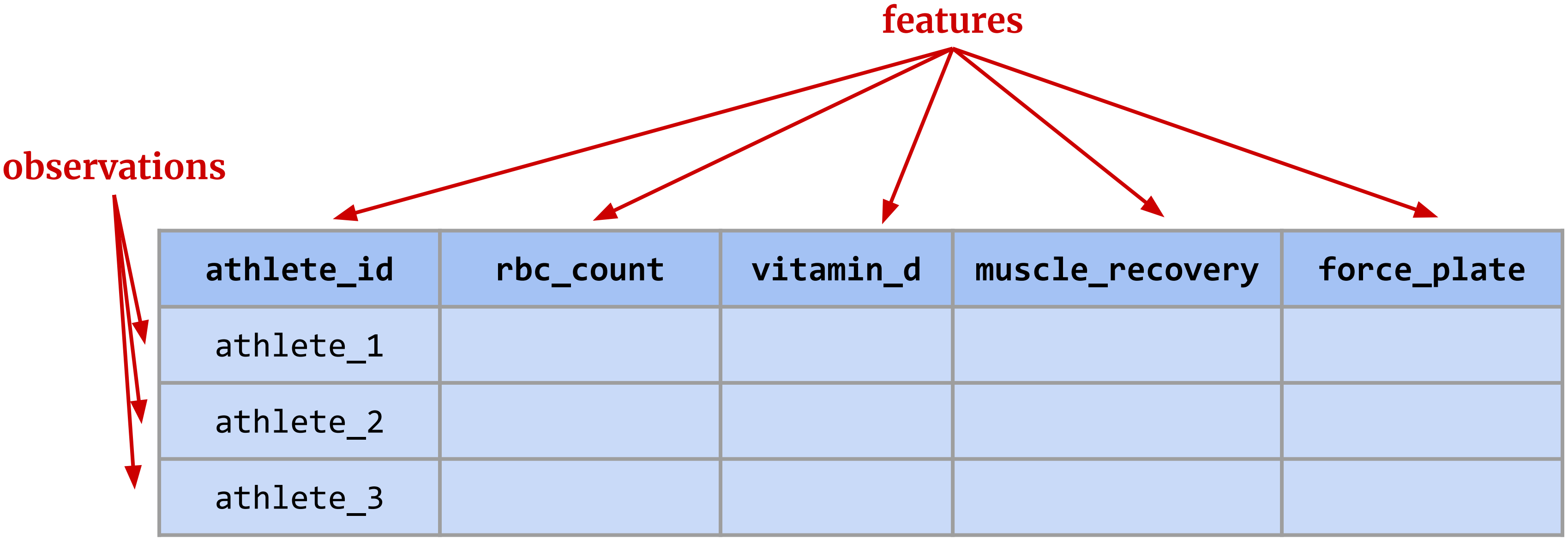
Data table basics



features

observations

Data table basics



features

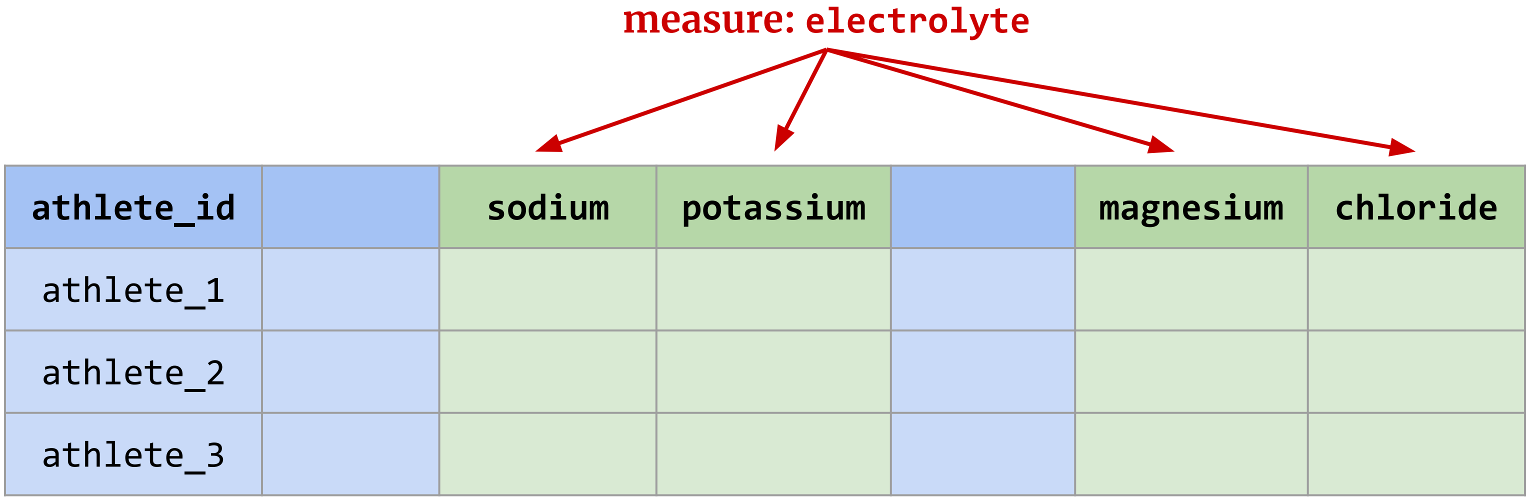
observations

athlete_id	rbc_count	vitamin_d	muscle_recovery	force_plate
athlete_1				
athlete_2				
athlete_3				

Outline

1. Data table basics
2. Terminology
3. Cluster analysis
4. Subject matter expert report

Measure



measure: electrolyte

athlete_id		sodium	potassium		magnesium	chloride
athlete_1						
athlete_2						
athlete_3						

Attribute

- Aspect of sports performance
  - Endurance
  - Strength
  - Power

- Each measure can influence one or more attributes

Outline

1. Data table basics
2. Terminology
3. Cluster analysis
4. Subject matter expert report

Overview

- We do cluster analysis to group together athletes who have similar test results

- Cluster analysis for every measure

Toy example

Red blood cells

athlete_id	hemoglobin	hematocrits
athlete_1	5	2
athlete_2	3	4
athlete_3	3	4
athlete_4	5	2

Toy example

Red blood cells

athlete_id	hemoglobin	hematocrits
athlete_1	5	7
athlete_2	3	4
athlete_3	3	4
athlete_4	5	7

k-modes clustering

- Clustering on categories assigned to numerical values
  - Examples: `in range`, `below range`, `above range`

athlete_id	hemoglobin	hematocrits
athlete_1	above range	above range
athlete_2	in range	in range
athlete_3	in range	in range
athlete_4	above range	above range

k-means clustering

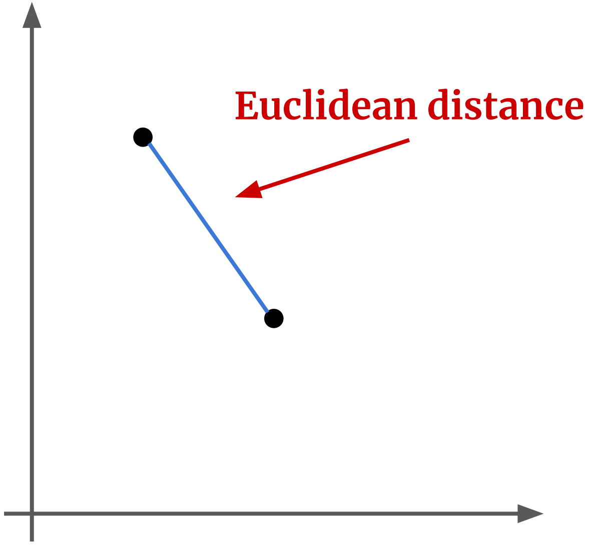
- Categorizing numerical values may cause us to lose important information contained in the numbers themselves

- Clustering on numbers, as opposed to categories

- Put athletes in the same cluster if they are similar

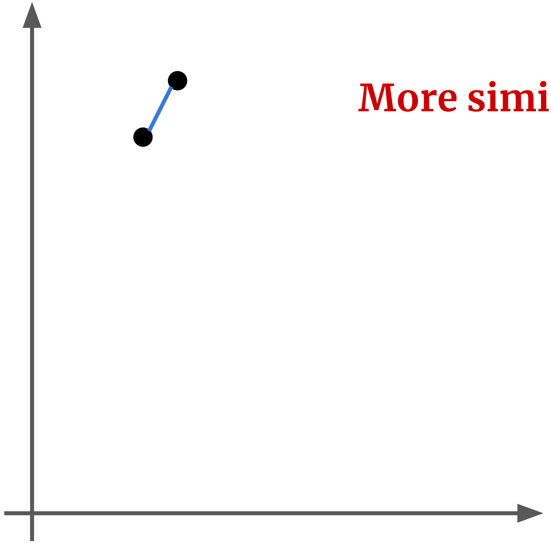
How to decide if numeric observations are similar?

- By looking at the distance between points

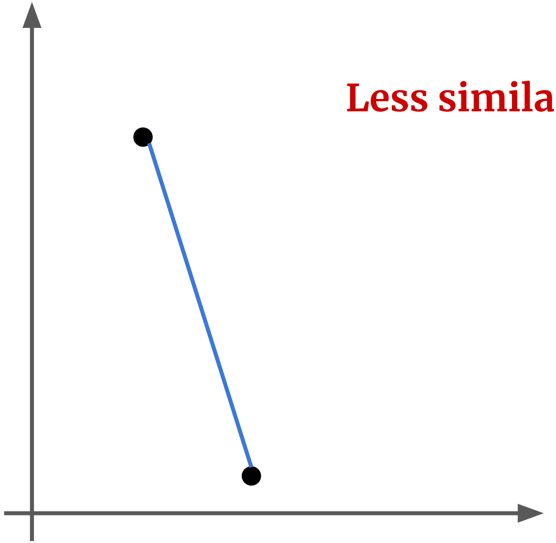


How to decide if numeric observations are similar?

- By looking at the distance between points



More similar



Less similar

Back to our data table...

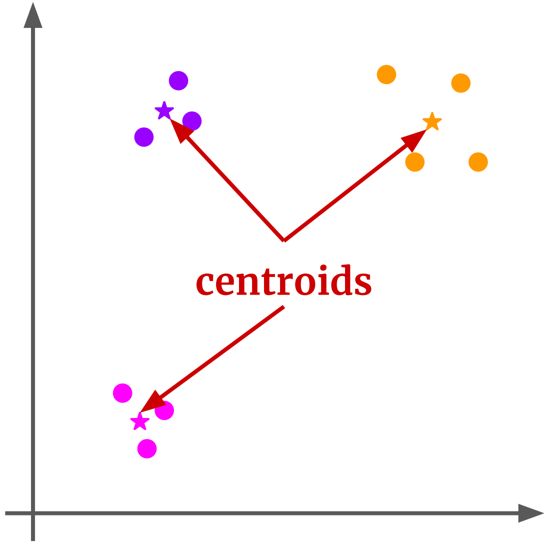
- Each feature would be an “axis”, and each athlete would be a “point”

- Some measures may have 4 or more features, which means 4 or more “axes”

- How do we measure Euclidean distance with so many axes?

Centroids in *k*-means clustering

- Each cluster is “centered” around its respective centroid

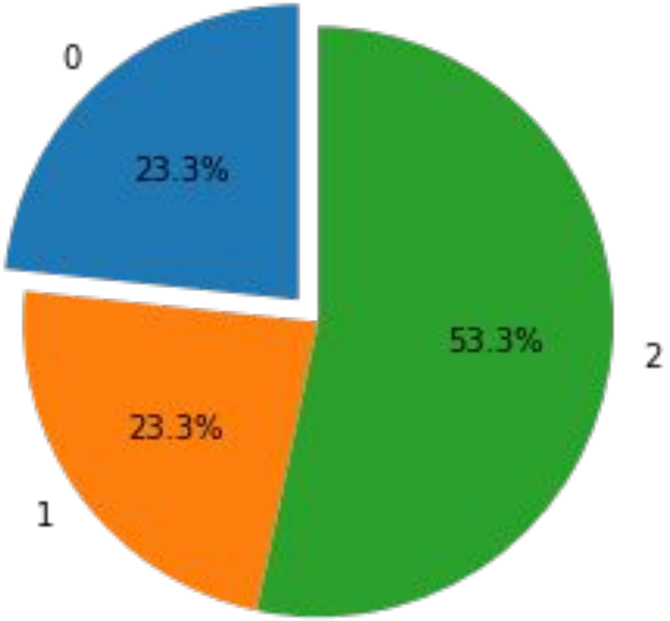


Reporting *k*-means clustering results

Measure: `electrolyte`

cluster	sodium	potassium	magnesium	chloride
cluster_0	140	4.0	2.2	100
cluster_1	110	2.1	0.4	77
cluster_2	180	8.9	5.5	136

Reporting *k*-means clustering results

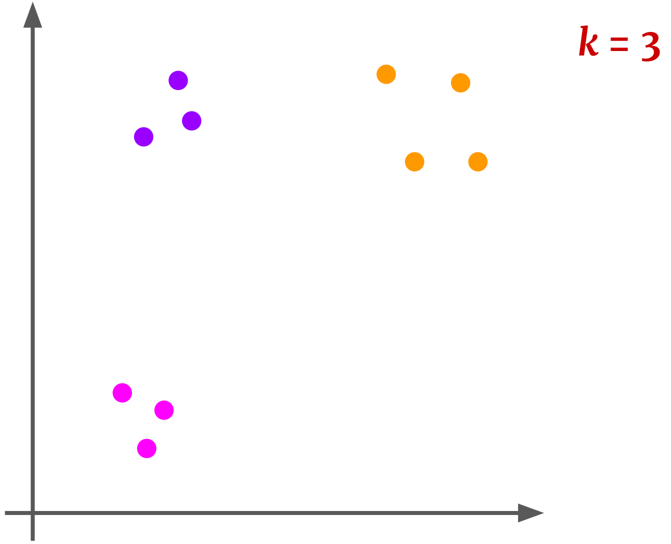


Data of an individual athlete:

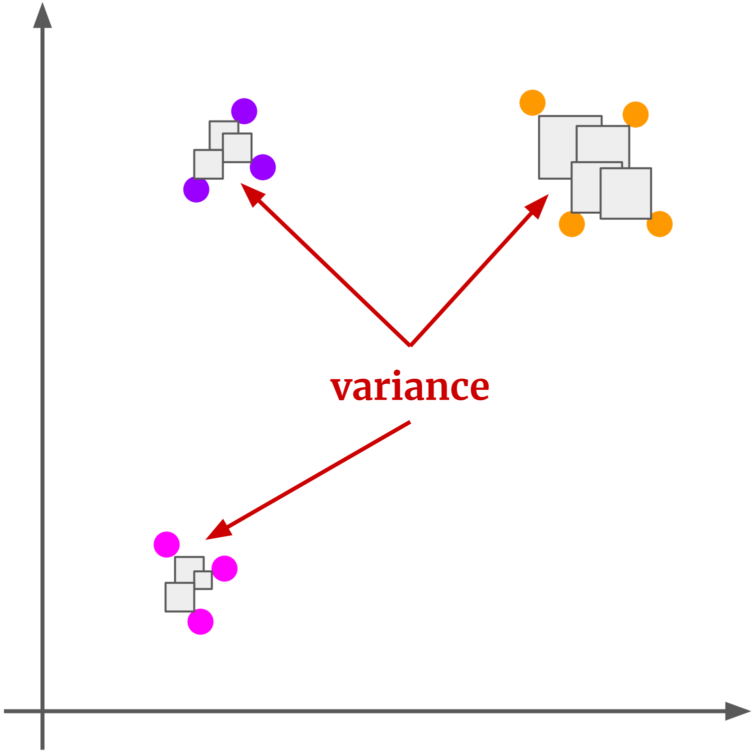
- Sodium: 135
- Potassium: 3.9
- Magnesium: 2.5
- Chloride: 98

What is *k* and why do we care about it?

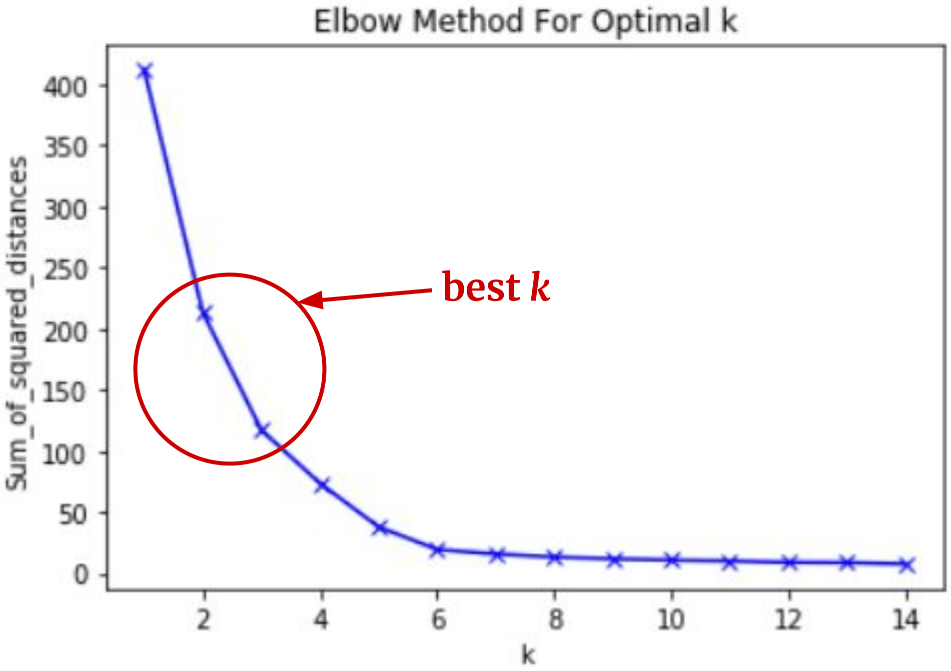
- *k* is the number of clusters we decide to assign to our athletes
- Choose *k* so that it best reflects the patterns in our athletes



Choose *k* by optimizing the variance in the data



Elbow plot



best *k*

Choosing *k* on our platform

- We decide *k* for the user using the elbow method, constraining *k* between 3 and 5

- Let users choose their own *k*

Other methods to choose *k*

- Plenty; most center around the idea of minimizing intra-cluster distances and maximizing inter-cluster distances

- More information here

Mahalanobis distance

- Corrects for effects from correlated features

- Same clustering procedure once pairwise distances are calculated

- Results are not so interpretable...

Cohorts

- Athlete chooses who he or she wants to compare with, and we perform cluster analysis on this cohort as opposed to on everyone

- With raw values, it is hard to say whether a centroid means “in range” or “out of range,” due to variability in demographics

- Cluster centroids are percentile values relative to everyone in the cohort

Outline

1. Data table basics
2. Terminology
3. Cluster analysis
4. Subject matter expert report

Report

- Integrates all of the tools and communicates everything to athletes, physicians, data scientists researchers

Thank you!