Pre-Class Discovery: Unsupervised Learning

Complete before the lecture. Bring your answers.

Data Science with Python – BSc Course

Instructions: Work through all six tasks before the lecture. There are no formulas here — just your intuition. Write directly on this sheet or bring a separate page with your answers. Total time: approximately 45 minutes.

Task 1: Sort These Companies into Groups

10 min

Below are 12 companies described by three numbers. Your job: sort them into groups of similar companies. You decide how many groups and what "similar" means.

ID	Company	Annual Return (%)	Volatility (%)	Market Cap (B USD)
A	SteadyBank	5.2	8.1	120
B	RocketTech	42.0	35.0	15
C	OldEnergy	3.1	12.5	85
D	SafeInsure	4.8	7.3	100
E	BioStartup	55.0	48.0	2
F	MegaRetail	8.5	11.0	200
G	CloudAI	38.0	30.0	25
H	GovBonds	2.0	3.5	–
I	CryptoVenture	60.0	55.0	5
J	PharmGiant	7.0	9.0	150
K	GreenSolar	28.0	26.0	10
L	DividendUtil	4.0	6.0	90

(a) How many groups did you create? List the companies in each group.

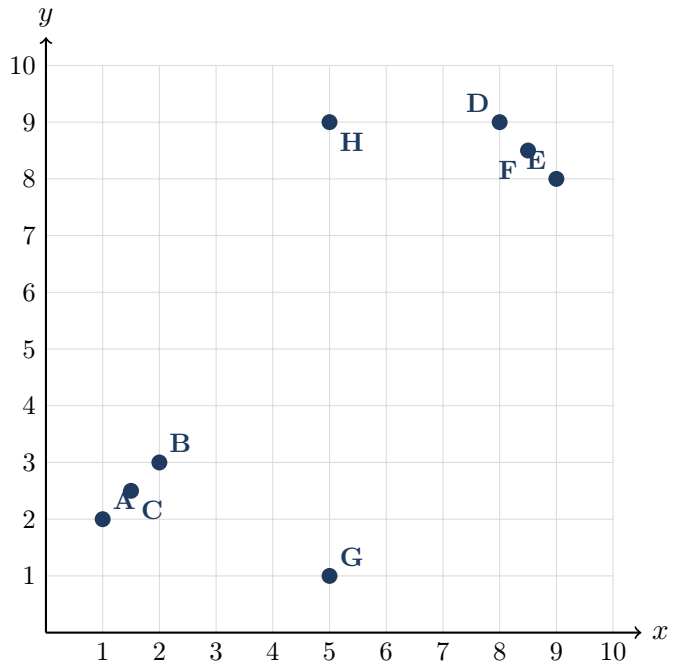
(b) What criteria did you use to decide which companies belong together?

(c) Could someone else look at the same table and create different groups? Why might that happen?

Task 2: Cluster by Eye

8 min

Eight data points are plotted on the grid below. Draw circles around the groups you see. How many clusters did you find?



(a) Draw circles around the groups. How many clusters do you see?

(b) Where did you put points G and H? Are they part of a cluster or on their own?

(c) What rule did your brain use to decide "these points belong together"? Can you describe that rule in one sentence?

Task 3: The Spotify Debate

5 min

Two friends are arguing about how Spotify should group songs into playlists:

- **Alex** says: "Group by *mood* — happy, sad, energetic, calm. A fast jazz track and a fast pop track both belong in the 'energetic' playlist."
- **Jordan** says: "Group by *genre* — pop, jazz, classical, hip-hop. Mixing genres in one playlist sounds terrible."

(a) Who do you agree with? Why?

(b) Could both be "correct"? What does this tell you about grouping data — is there always one right answer?

(c) If you had to pick *measurable features* from a song (tempo, loudness, key, lyrics sentiment, etc.), which features would support Alex's grouping and which would support Jordan's?

Task 4: Too Many Numbers

7 min

You measured five features for 50 students: **height**, **weight**, **age**, **shoe size**, and **arm span**.

(a) Which pairs of features do you expect to be strongly related (i.e., if you know one, you can roughly predict the other)?

(b) Could you describe each student with fewer than five numbers and still capture most of the important differences between students? Which features would you keep, which would you drop, and why?

(c) Imagine you had 500 features instead of 5. Would it still be possible to look at the data and spot patterns by hand? What would you need a computer to do for you?

Task 5: City at Night

8 min

Imagine you are flying over a city at night. You see lights below:

- Dense clusters of lights form **neighborhoods**.
- Some lights sit right at the **edge** of a neighborhood.
- A few isolated lights shine in **empty fields**, far from any neighborhood.

(a) In the space below, draw a bird's-eye view of a city with at least **three neighborhoods**, a few **edge lights**, and two or three **isolated lights** in empty fields.

(b) Label your drawing:

- Mark the lights at the dense center of each neighborhood as **Core**.
- Mark the lights at the boundary as **Border**.
- Mark the isolated field lights as **Noise**.

(c) What rule would you give a computer to decide whether a light is Core, Border, or Noise? Think about how many other lights are nearby and how close "nearby" means.

Task 6: Your Questions

5 min

After working through Tasks 1–5, you have seen several ideas: grouping items by similarity, reducing redundant information, handling outliers, and choosing features.

Write **three questions** you want answered during the lecture. These can be about anything — methods, applications, pitfalls, or connections to other topics.

Question 1:

Question 2:

Question 3: