Pre-Class Discovery: Solutions and Discussion Notes

Instructor copy – do not distribute before the lecture.

Data Science with Python – BSc Course

Task 1: Sort These Companies into Groups

Model answer

A natural grouping with three clusters:

- **Stable / low-risk:** A (SteadyBank), C (OldEnergy), D (SafeInsure), F (MegaRetail), H (GovBonds), J (PharmGiant), L (DividendUtil). Low returns (2–9%), low volatility (3–12%), large market caps.
- **Growth / high-risk:** B (RocketTech), G (CloudAI), K (GreenSolar). High returns (28–42%), high volatility (26–35%), mid-cap.
- **Speculative / very high-risk:** E (BioStartup), I (CryptoVenture). Extreme returns (55–60%), extreme volatility (48–55%), micro-cap.

(b) Criteria used: Most students group by return and volatility together (risk-return profile). Some also consider market cap as a third dimension.

(c) Different groups: Yes — a two-cluster split (safe vs. risky) or a four-cluster split (separating mega-caps from mid-caps within the stable group) are both defensible. This foreshadows the "choosing K" problem in K-Means.

Common misconceptions

- Students sometimes group alphabetically or by industry name rather than by the numerical features. Redirect: in unsupervised learning, the algorithm only sees numbers.
- Some insist there is one correct grouping. Key point: clustering is subjective — different criteria yield different (valid) groupings.

Task 2: Cluster by Eye

Model answer

(a) Two clear clusters: {A, B, C} in the bottom-left and {D, E, F} in the top-right. Points G and H are ambiguous.

(b) G sits alone at (5,1) — far from both clusters. H sits alone at (5,9) — closer to the D/E/F cluster but still separated. Reasonable answers:

- 2 clusters with G and H as outliers/noise
- 3 clusters: {A,B,C}, {D,E,F,H}, and G as noise
- 4 clusters: {A,B,C}, {D,E,F}, {H}, {G}

(c) The intuitive rule is "points that are close together belong in the same group." This is essentially what distance-based clustering algorithms formalize.

Discussion note

This task directly sets up three lecture topics:

1. K-Means uses Euclidean distance (the "closeness" students described)
2. DBSCAN can label G as noise automatically
3. The number of clusters depends on the method and parameters

Task 3: The Spotify Debate

Model answer

(a) Both are correct. This is the central insight: clustering depends on which features you choose.

(b) Yes, both can be correct simultaneously. There is no single "true" clustering — the result depends on the question you are asking and the features you measure. This connects to the lecture point that feature selection and scaling determine clustering outcomes more than the algorithm choice.

(c) Feature mapping:

- **Alex (mood):** tempo (BPM), loudness (dB), valence (happiness score), energy, danceability
- **Jordan (genre):** instrument distribution, harmonic complexity, lyrics topic, typical time signature

Common misconceptions

- Students often pick a side and argue one is "better." Push back: in unsupervised learning, the notion of "correct" depends on the use case, not on the data alone.

Task 4: Too Many Numbers

Model answer

(a) Strongly correlated pairs:

- Height and arm span (nearly 1:1 proportional)
- Height and shoe size (taller people tend to have larger feet)
- Weight and height (correlated but with more spread)

Age is largely independent of the body measurements for a student cohort of similar age.

(b) Yes, you could reduce to about 2–3 features:

- A "body size" composite (captures height, arm span, shoe size)
- Weight (partially independent of height)
- Age (independent of body proportions)

This is exactly what PCA does: it finds the directions of maximum variance and compresses correlated features into fewer components.

(c) With 500 features, manual inspection is impossible. You would need a computer to identify which features are redundant, compress them, and visualize the result in 2–3 dimensions. This motivates PCA, t-SNE, and UMAP from the lecture.

Discussion note

Use this task to introduce the curse of dimensionality and the idea that most real-world datasets have redundant features.

Task 5: City at Night

Model answer

(a) Drawings should show 3+ dense patches of dots (neighborhoods), a few dots at the edges of those patches, and 2–3 isolated dots far from any patch.

(b) Labels:

- **Core:** Lights in the dense interior of a neighborhood, surrounded by many other lights.
- **Border:** Lights at the edge of a neighborhood, near some core lights but not surrounded on all sides.
- **Noise:** Isolated lights in empty fields, far from any neighborhood.

(c) A computer rule might be: "A light is Core if it has at least [some number] other lights within [some distance]. A light is Border if it is not Core but is within [that distance] of at least one Core light. Everything else is Noise."

This is precisely the DBSCAN algorithm with parameters MinPts and epsilon.

Common misconceptions

- Students sometimes think Border and Noise are the same. Clarify: border points are near a cluster but not dense enough to be core. Noise is truly isolated.
- Some students struggle with the two-parameter rule. Emphasize: you need both a distance threshold *and* a count threshold to distinguish density from isolation.

Task 6: Your Questions

Typical student questions and where they are answered

- "How do you know how many clusters to use?" — Lecture slides on elbow method, silhouette score, and BIC.
- "What if the clusters are not round?" — DBSCAN and GMM sections.
- "How do you know if the clustering is good?" — Internal validation metrics (silhouette, inertia) and external validation when labels exist.
- "Can you combine clustering with prediction?" — Pipeline section (StandardScaler + PCA + KMeans).
- "What happens with outliers?" — DBSCAN noise label, GMM low-probability points.
- "How is this different from classification?" — Opening section: no labels vs. labeled data.

Instructor tip

Collect student questions at the start of the lecture. Revisit them at the end to check whether the lecture answered each one. Unanswered questions make good exam review topics.