

Module 5B Questions:

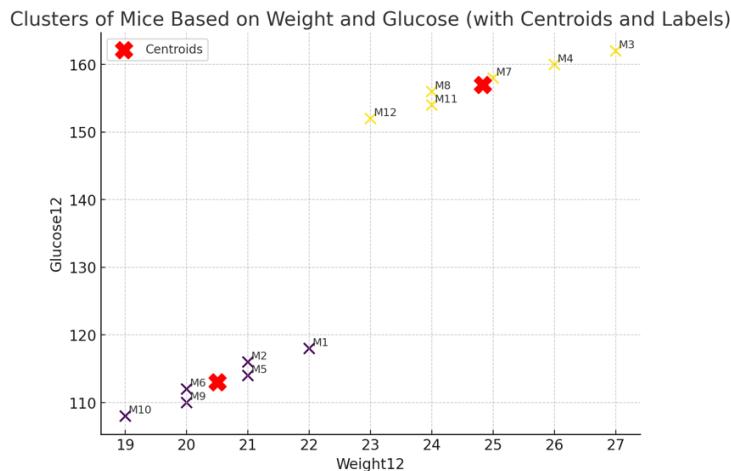
1. Researchers used **weight (Weight12)** and **glucose (Glucose12)** to identify natural clusters among the 12 mice.

Two clusters emerged:

- **Cluster A:** Mice with lower weight and lower glucose
- **Cluster B:** Mice with higher weight and higher glucose

Using this clustering outcome, answer the following:

- i. What do these two clusters suggest about the natural grouping of mice based on physiological traits (weight and glucose)?
- ii. Do these clusters appear to align more closely with **diet** (Chow vs. HFD) or **strain** (B6, BALB, CAST)? Explain what patterns in the data support your interpretation.
- iii. If a mouse falls near the boundary between clusters, what would that imply about its metabolic profile?
- iv. Why might clustering based on *continuous traits* (like weight and glucose) reveal different groupings than clustering based on *categorical labels* (like strain or diet)?



2. You want to analyze gene expression data from different strains of mice to understand how their genes respond to varying environmental conditions. You use k-means clustering to group these genes into 3 clusters based on their expression patterns (low, medium, high). Explain – at a very high-level conceptual level, no math necessary, but you might want to use the website that was referenced in the K-means section! – how K-means decides which genes belong to which cluster? What are centroids and what is their role in this process? What happens if two data points are equi-distant from two centroids? What happens if there is a lot of overlap with the data points?