Homework 1
- grades should be on Blackboard
- if you have specific questions about your grade, feel free to reach out to Lyla or myself
- if you would like to boost up your grade, keep in mind that there will be two extra credit assignment available to be announced towards the end of the course

- general comments:
1. please double check that your homework is going to show up properly on the course website (formatting, filenames)
-> if I have to fix things, if you pull request breaks thing, it's not the end of the world -> we will deduct 5%
-> if you are facing challenges with Github pulls, reach out to Lyla or myself

2. as we start working with more different types of data, we will want to make sure that we're being clear with our language descriptors (categorical, ordinal, quantitative) and visually channels (color, is it hue or is it saturation) -> our data visualizations will become more complex
-> as our data visualizations become more complex, it will important to be purposeful in our choices -> critically about how do these choices enhance saliency

Reflection cards
- we are keeping track of them
- we will review those question
-> some of these questions will address themselves as we continue to build on previous lessons

Kmeans clustering

-> we will have more time for the hands on component

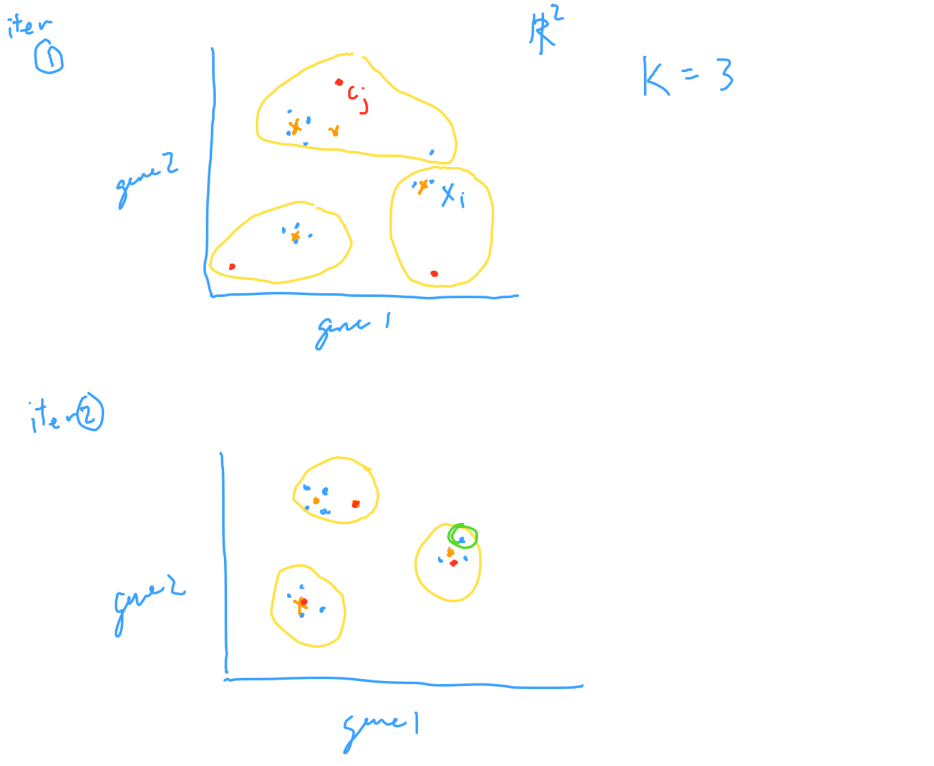
-> I'm guiding you to analyze and look at the data in a particular way
-> you should critically think about every step that you're taking, why you're taking those steps
-> in the future, next week, I will give you all a new dataset
-> confident that you know what you're doing and can figure out what you try to look at
-> real world has no right answer to how to do these analyses -> up to you to make choices, using data visualizations to evaluate whether your approach is reasonable

Learning objectives:
- what is kmeans -> apply to our data
-> how can we use data visualization to evaluate the quality of our analyses pipeline?

Goal: find transcriptionally distinct groups of cells -> cell-types
(excitatory neurons, inhibitory neurons, microglia, oligodendrocytes, astrocytes)

Kmeans clustering is one of many clustering algorithms
-> partition observations (cells) into k clusters (groups, cell-types) based on attributes on those observations
-> k must be specific a priori

Algorithm:



1. randomly place k centroids (c) → stochastic

2. repeat to until convergence
(convergence means that cluster assignments no longer change):
- for each observation (x)
  - find the nearest centroid

$$\arg\min_j D(x_i, c_j)$$

↳ euclidean

  - assign the observation to that cluster (centroid)

- for each centroid, recompute to find a new centroid based on the observations in that cluster

$$c_j' = \frac{1}{n_j} \sum x_i$$

(mean)

x_i: ~~be categorical~~ ordinal ?

$\mathbb{R}^2$

gene ?

x_i quantitative → x_i be PCs ? ←

tSNE ? ←