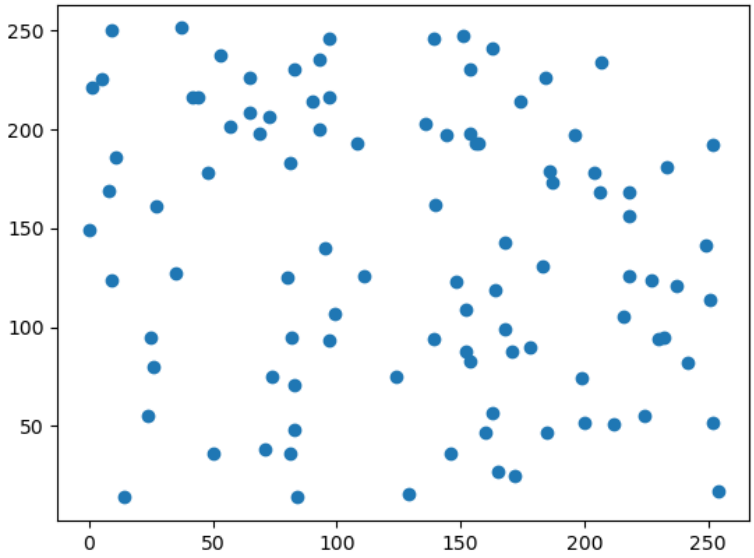
Ira A. Fulton Schools of Engineering
Arizona State University

CSE 572: Data Mining
Fall 2022 (Kerner)

Homework 3: Clustering

An important question in assessing cluster validity is whether we are finding real patterns in structured data or finding patterns in noise or random data. In this homework, you are given a dataset from an unknown source with unknown attributes. The figure below shows a scatter plot of the first two feature dimensions in the dataset.



You are asked to cluster the data into 3 clusters using K-means clustering. Your goal is to evaluate whether the clusters you find in the dataset are due to real structure in the data or just patterns in noise.

You can use any method we have discussed in class or in the Data Mining textbook to solve this problem. Show all of your work and then answer the question in the final Question cell of the notebook.

Submission

You will add your code to the notebook provided in the assignment instructions which contains starter code for loading the dataset (`cse572-homework3.ipynb`). Rename the notebook to `cse572-homework3-<lastname>.ipynb` and submit the following three deliverables:

1. a link to your Colab notebook (as a comment on the submission)
2. your .ipynb file (`cse572-homework3-<lastname>.ipynb`)
3. a pdf of the executed notebook (`cse572-homework3-<lastname>.pdf`)

Grading

Grading will be based on your code and the correctness of your response to the question. You may receive partial credit if you come to the incorrect conclusion but parts of your work are correct, so make sure to show your work.