Introduction to Data Science
Homework: Clustering

1. In this problem, we are going to use the utilities dataset. This dataset provides different attributes of utility companies in different states.
    a. Apply k-mean algorithm with two different k (your choice), visualize the value of your cluster centroids for different features and discuss which k is better.
    b. Apply hierarchal clustering on the data. Use two different methods to measure the distance between clusters (single, average, complete). Visualize your trees. Pick the one you think is more reasonable. Justify your choice. Pick two different k (that you chose in section a) to cut the tree.
    c. Compare the result (for the best k) in section a and b. Are they similar?

2. Use voting and political party dataset. Apply k-means on all the voting data (obviously k=2). Now check the real political parties of each sample in each cluster. Report that in each cluster, how many are democrat and how many are republican. Discuss your result.

Description of Data:

- **Utilities**:
  # x1: Fixed - charge covering ration (income/debt)
  # x2: Rate of return on capital
  # x3: Cost per KW capacity in place
  # x4: Annual Load Factor
  # x5: Peak KWH demand growth from 1974 to 1975
  # x6: Sales (KWH use per year)
  # x7: Percent Nuclear
  # x8: Total fuel costs (cents per KWH)

- **Voting**:
  • The file "p2 congress 1984 votes.csv" contains a matrix $X \in \{-1, 0, 1\}^{435\times16}$ indicating the votes of 435 U.S. House of Representatives congress members on 16 key issues in the congress of 1984. Here, -1, 0 and 1 denote reject, neutral and accept votes respectively. Here is the list of attributes that they voted on:
  # handicapped-infants: 2 (y,n)
  # water-project-cost-sharing: 2 (y,n)
  # adoption-of-the-budget-resolution: 2 (y,n)
  # physician-fee-freeze: 2 (y,n)
  # el-salvador-aid: 2 (y,n)
  # religious-groups-in-schools: 2 (y,n)
  # anti-satellite-test-ban: 2 (y,n)
  # aid-to-nicaraguan-contras: 2 (y,n)

Introduction to Data Science
Homework: Clustering

mx-missile: 2 (y,n)
immigration: 2 (y,n)
synfuels-corporation-cutback: 2 (y,n)
education-spending: 2 (y,n)
superfund-right-to-sue: 2 (y,n)
crime: 2 (y,n)
duty-free-exports: 2 (y,n)
export-administration-act-south-africa: 2 (y,n)

• The file “p2 congress 1984 party affiliations.csv“ contains a vector Y of size 435 x 1 indicating the party affiliations (Republican or Democrat) of 435 congress members in the congress of 1984