Clustering Homework

All submissions:

- **QUESTIONS:** Edit readme with answers to the discussion questions
- **CODE:** Share clustering code:
    - Option 1: Link to online editor (Google Colab or Replit)
    - Option 2: Add **clusteringHW.py** file
    - Option 3: Add **clusteringHW.ipynb** file

Mild

- Complete the clustering algorithm for 1D data.
    - Put the code from section 5A into a loop that repeats a certain number of times - Section 5 Option 5B
    - Comment out 5A and disregard 5C
- Analyze the survey data.
    - Comment out the code in section 3A. Use the commented partial code in section 3B to plot each of the 5 data sets.
    - **Answer: "Q1: Which sets best show clustering?"**
    - Choose 1 of the sets that show good clustering.
    - Since these data sets are small and have repeated values, using `np.random.choice` creates a risk of matching centroid values. To avoid this, update the Section 4 line defining the initial centroids to use the `np.random.uniform` function instead of `choice`, using "low" as your lowest data value and "high" as the highest data value. (Try to use code to get these values instead of reading the data yourself!)
    - Test different k-values to best fit the data. Run your clustering algorithm and graph the results for different values of k until you feel that you have a "best" value.
    - **Answer: "Q2: What were the centroids? What k-value best fit the data?"**
    - **Answer: "Q3: What meaning can you extract from the clustering (what do these clusters tell us about the people in this class)? If you could not identify good clusters, what does that tell us about the data?"**
- Submission:
    - Python code that displays final clustered graph with centroids (axis/plot labels are appreciated!)
    - Readme file with Answers

Spicy

- Find a .CSV online and upload it into your project
    - You can find lots of good data from Kaggle
    - If using Colab, place it in the 'content' folder; otherwise, be sure to upload it to Github
    - **Answer: "Q1: What dataset .csv did you choose?"**
- Complete the clustering algorithm for 1D data.
    - Put the code from section 5 into a loop that repeats until the centroids or clusters stop changing
    - Comment out 5A and disregard 5B when running code in 5C

    - Print out number of iterations until convergence
- Run your finished clustering algorithm on the data you chose. Try different values for k until you're satisfied that you have the best k.
    - **Answer: "Q2: How many iterations did your program need to converge?**
    - **Answer: "Q3: What were the centroids? What k-value best fit the data?"**
    - **Answer: "Q4: What meaning can you extract from the clustering (what did the clusters tell you about the data's original context)? If you could not identify good clusters, what does that tell us about the data?"**
- Submission:
    - Python code that displays final clustered graph with centroids (axis/plot labels are appreciated!)
    - Readme file with Answers

Caliente

- Create a 2-dimensional clustering algorithm based on the 1D code along.
- Run the clustering algorithm on the iris data (in scikitlearn, provided code and k - see lesson slides).
- Optional: Create an algorithm to find the optimal k value by graphing the sum of squared differences.
    - **Answer: "Q1: How many iterations did your program need?**
    - **Answer: "Q2: What were the centroids? What k-value best fit the data?"**
    - **Answer: "Q3: What meaning can you extract from the clustering (what did the clusters tell you about the data's original context)? If you could not identify good clusters, what does that tell us about the data?"**
- Submission:
    - Python code that displays final clustered graph with centroids
    - Readme file with Answers