Homework 2

Deadline: 2018.03.20 (Tuesday) 23:59

Problem 1: Olympic Data Analysis [hw21.ipynb]

Your task is to analyze an olympics dataset (**olympics.csv**), which was derrived from the Wikipedia entry on All Time Olympic Games Medals, and does some basic data cleaning. Use this dataset to answer the questions below. You are required to write some functions, and write some code to execute these functions. Note that this time you are required to do this problem using Pandas in ipynb.

(1) Which country has won the most gold medals in summer games?
The answer is: `United States`

(2) Which country had the biggest difference between their summer and winter gold medal counts?
The answer is: `United States`

(3) Which country has the biggest difference between their summer and winter gold medal counts relative to their total gold medal count? Only include countries that have won at least 1 gold in both summer and winter.

$$\frac{Summer\ Gold - Winter\ Gold}{Total\ Gold}$$

The answer is: `Bulgaria`

Problem 2: Census Data Analysis [hw22.ipynb]

Your task is to analyze a census dataset (**census.csv**) from the United States Census Bureau. Counties are political and geographic subdivisions of states in the United States. This dataset contains population data for counties and states in the US from 2010 to 2015. For a description of the variable names, please refer to **co-est2015-alldata.pdf**. Note that this time you are required to do this problem using Pandas in ipynb.

(1) Which state has the most counties in it?
(Hint: consider the sumlevel key carefully! You'll need this for future questions too.)
The answer is: `Texas`

(2) Only looking at the three most populous counties for each state, what are the three most populous states (in order of highest population to lowest population)?
(Hint: Use CENSUS2010POP)

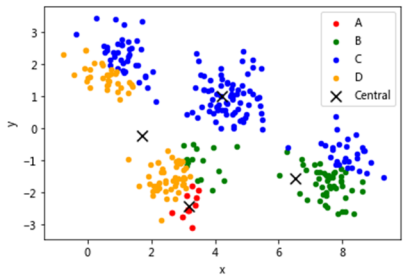
The answer is: `['California', 'Texas', 'New York']` if you do not consider SUMLEV
or: `['California', 'Texas', 'Illinois']` if you consider SUMLEV
Note that both answers are correct.

(3) Which county has had the largest change in population within the five year period?
(Hint: population values are stored in columns POPESTIMATE2010 through POPESTIMATE2015, you need to consider all six columns.)
e.g. If County Population in the 6 year period is 100, 120, 80, 105, 100, 130, then its largest change in the period would be |130-80| = 50.
The answer is: `Texas` if you do not consider SUMLEV
or: `Harris County` if you consider SUMLEV
Note that both answers are correct.

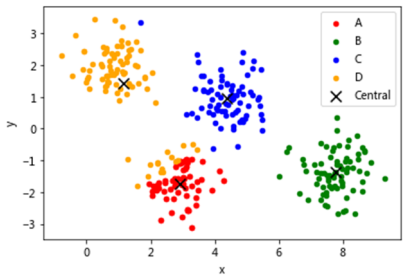
Problem 3: K-means Clustering Implementation [hw23.ipynb]

Your task is to use Python, along with numpy and Pandas, to implement the well-known clustering algorithm, K-means, based on a synthetic dataset **cdata.csv**. This dataset contains two data columns, "X" and "Y", and one "cluster" column (1, 2, 3, and 4). In implementing K-means, you need to use "X" and "Y" as **features** for clustering while the "cluster" column is for your validation. Note that it is not necessary to perfectly clustering all of the data points into clusters. Also note that the "cluster" column is not used in clustering.

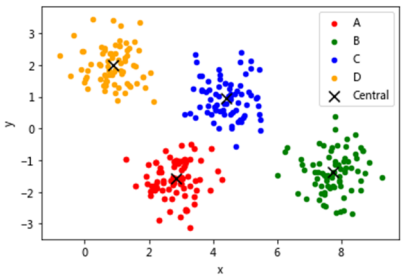
(1) Randomly select data points as the initialized centroids. By default, please set K=4. Report and plot the process until convergence. The centroids also need to be plotted. An example is shown below. Note that it may not have 3 rounds (it can be 4 or 5 rounds, depend on initialized centroids).



Round 1

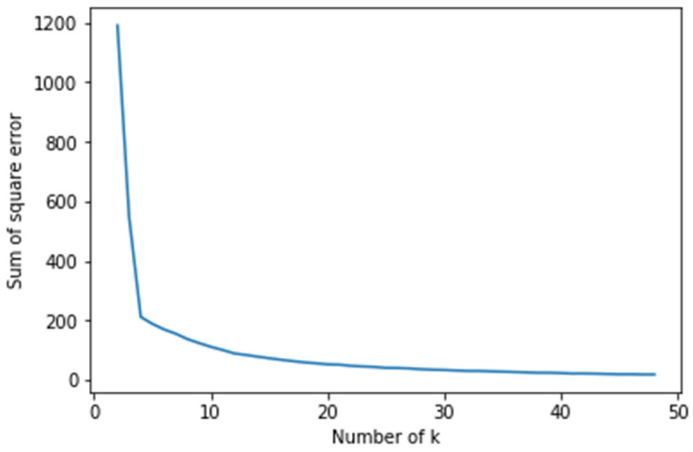


Round 2

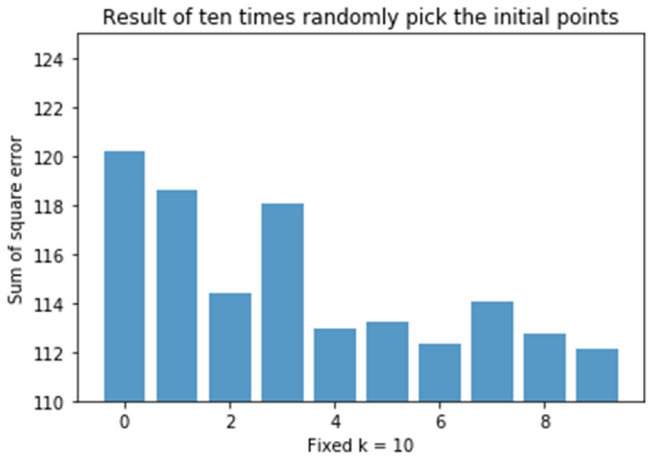


Round 3

(2) Re-execute your K-means clustering algorithm by changing K from 2 to 50 (from 2 to 10 is also okay). Plot the K value (x-axis) vs. the value of Sum of Squared Error (SSE) (y-axis) as below. Note that it is reasonable and acceptable if the curve is 凹凸不平. ☺



(3) Try 10 times of randomly initialized centroids, and plot their SSE values (y-axis) as below.



Note that in your codes for these problems,
you need to write some comments to describe the meaning of each part.

How to Submit Your Homework?

Submission in NCKU Moodle. Before submitting your homework, please zip the files (**hw21.ipynb**, **hw22.ipynb**, **hw23.ipynb**) in a zip file, and name the file as "學號_hw2.zip". For example, if your 學號 of your team are H12345678, then your file name is:

"H12345678_hw**2**.zip" or "H12345678_hw**2**.rar"

When you zip your files, please follow the instructions provided by TA's slides to submit your file using NCKU Moodle platform http://moodle.ncku.edu.tw .

Have Questions about This Homework?

Please feel free to visit TAs, and ask/discuss any questions in their office hours. We will be more than happy to help you.