Principles and Applications of Data Science
Homework #1

Due: Apr 22, 2020

This assignment is to practice how to use Jupyter/IPython notebook and run a toy example (covid19-state-dataset). Please follow the steps below to have the work done on notebook. The generic template (HW1-covid19_state_data.ipynb) of this homework is provided on the **i-school(Plus)** (`https://istudy.ntut.edu.tw/learn/index.php`) platform of school.

Step 1
: Use Pandas (`https://pandas.pydata.org/`) to load COVID-19 State Data Set (`https://www.kaggle.com/nightranger77/covid19-state-data/data`) as the dataframe.

Step 2
: Get 20 data items as sample randomly and show them.

Step 3
: Show 10 data items which the Deaths are more than 100 as sample randomly.

Step 4
: Sort the data by GDP and present the top 20 data items.

Step 5
: Show the simple statistical information (mean, std, min, max, quartile1, quartile2, quartile3).

**Use matplotlib (`https://matplotlib.org/`) to show 2D images about data.

Step 6
: Plot the distribution of two classes: 1. GDP $<$ 58000, and 2. GDP $\leq$ 58000 in COVID-19 State Data using different colors and different marker where x-axis is the Pollution and y-axis the Mortality-rate.

Step 7
: Show the proportion of three classes below in COVID-19 State Data using pie chart:

Class 1 Mortality-rate $<$ 0.02

Class 2 Mortality-rate between 0.02 and 0.03

Class 3 Mortality-rate $>$ 0.03

About submitting this homework

- Please upload your homework project named as `HW1-covid19_state_data-SID.ipynb` to **i-school(Plus)** (`https://istudy.ntut.edu.tw/learn/index.php`) platform .
- The **deadline** is the **midnight of Apr 22**, 2020 and **Late work** is not acceptable.

- Honest Policy: We encourage students to discuss their work with the peer. However, each student should write the program or the problem solutions on her/his own. Those who copy others work will get 0 on the homework grade.