CARNEGIE MELLON UNIVERSITY
DATA, INFERENCE & APPLIED MACHINE LEARNING (COURSE 18-785)
ASSIGNMENT 2

INSTRUCTIONS

- Submissions should be made via canvas.
- **Single** Python/MATLAB code file(.ipynb or .m) **[Do not Submit checkpoints for .ipynb]**. In addition, each line of code should be documented by text. This demonstrates that the code is unique and owned by the student
- Assignment report(.pdf) with full evidence that the assignment was completed by the student and demonstrate a full understanding of each step in the process including textual descriptions of each result (statistics, table, graph etc) represents and insights that can be gained
  - Indicate the libraries you have used in your code at the beginning of the report (After the title page)
- Data files (as given)

Submission process:

1. Put source code **file and data files** in a single folder
2. Name of the folder should be the same as your andrew ID
3. **Zip this folder and attach the zipped file on assignment submission page (CANVAS)**
4. After attaching zipped file, click on "Add Another File" from assignment submission page and **attach your report**
5. Submit your assignment

N.B. This new process will allow us to compile your reports in **Turnitin** to check for plagiarism.

Specific reasons for a submission being classified as incomplete include:

- Failure to correctly name your folder with your Andrew ID, report, and code file with andrewID_DIAML_AssignmentNo. For example, mcsharry_DIAML_Assignment1, mcsharry_DIAML_Assignment2 and mcsharry_DIAML_Assignment3.
- A missing report describing the steps, results, and insights
- A missing dataset required for running the code
- A missing code file such as .ipynb or .m file
- An error in the file path needed to run the code

The student is responsible for checking that their submission is complete. Students will lose 10% as for late submission even if the submission is repaired during the 24 hours after the deadline has passed, and receive 0 for the assignment if it is not repaired.

The submission deadline is **Eastern Time (ET) on Monday, 19 September 2022 17:59 / Rwandan Time (CAT) on Monday, 19 September 2022 23:59.**

No.	Question	Format	Value
1	Using the World Bank Indicators, download data for "GDP per capita (current US$)" and "Malnutrition prevalence, weight for age (% of children under 5)." What kind of relationship do you expect? Make a scatter plot of malnutrition prevalence against GDP per capita (using all available years and countries). What kind of relationship do you see? Make a graph for the geographical regions (six regions excluding North America and use a different color for each). Make a graph for income levels (four income levels and use a different color for each). Carefully label all graphs and provide legends.	Three Graphs	20%
2	Using Quandl, download data, synchronize the time stamps and plot time series for the prices of Wheat, Crude Oil and Gold in $ on the same graph. Indicate the maximum and minimum prices in all three time series using coloured dots. Use a legend to explain each one.	Graph	20%
3	Download "CO2 emissions (metric tons per capita)" data from the World Bank Indicators. Select the emissions for all countries in 2010 and calculate summary statistics. Provide a table giving the mean, median, standard deviation, 5, 25, 75, and 95 percentiles. Repeat the same process for "School enrolment, primary (% net)."	Two Tables	20%
4	The World Bank Indicators provide variables called "Fertility rate, total (births per woman)" and "GDP per capita (current US$)". Make a scatter plot of Fertility rate versus GDP per capita for all countries in 2010. Produce cumulative distribution functions for the fertility rate variable using data from 1990 and 2010 respectively. Use vertical lines to indicate the mean and median. Use a legend to explain which is the mean and which is the median. Have fertility rates changed over this twenty-year period?	Two graphs	20%
5	Download the data for the "Happy Planet Index" from hpi data and "Corruption Perceptions Index" from https://www.transparency.org/en/cpi/2016/index/nzl. Both datasets are available as excel spread-sheets. Find matching countries for both indices and make a carefully labelled scatter plot of HPI against CPI to demonstrate the relationship using ranks in both cases. Are there any countries that stand out as being unusual?	Graph	20%