CARNEGIE MELLON UNIVERSITY
DATA ANALYTICS (COURSE 18-787)
ASSIGNMENT 1

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)
- Using ChatGPT for any assignment is not allowed as it could lead to being flagged for plagiarism.
- 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 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_DA_AssignmentNo. For example, mcsharry_DA_Assignment1, mcsharry_DA_Assignment2 and mcsharry_DA_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 **on Monday 30, January, 2023 16:59 Eastern Time (ET) /**

Monday 30, January, 2023 23:59 Rwandan Time (CAT) .

1. Download historical daily weather data for France from https://canvas.cmu.edu/files/7555321/download?download_frd=1 . Load the data into your environment for use. Fill any gaps in the data using linear interpolation.

2. Calculate the correlation matrix between all the weather variables. Make a graphic to show the correlation matrix as a heat-map.

3. Download historical daily electricity consumption data for France from:
https://canvas.cmu.edu/files/7555322/download?download_frd=1
Save it as a csv file and load it into your computer.

4. Synchronise the dates corresponding to both time series and make a scatter plot of energy consumption against mean temperature.

5. Fit a quadratic model to the energy versus temperature. Plot the quadratic fit as a line on top of the scatter plot.

6. Based on the empirical analysis, what is the optimal temperature coinciding with minimal consumption? Use the quadratic fit and verify visually.

7. Use a stepwise approach to find an optimal multivariate linear regression model using the weather variables to forecast consumption. Which variables are selected? What is the coefficient of determination, R^2?

8. Increase the number of explanatory variables by also considering squared terms for each weather variable. Use a stepwise approach to obtain a new model. Which variables are selected? What is the new R^2 value and is this an improvement?

9. Consider the day of the week effect by including dummy variables for the day of the week in the multivariate regression. Which days of the week are selected for the new model? What is the new R^2 value and does this improve the model?

10. Can you be sure that this modelling approach is not over-fitting? Describe two approaches that could be used to prevent overfitting?