CARNEGIE MELLON UNIVERSITY
DATA ANALYTICS (COURSE 18-787)
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)
- 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. Only deadline extension requests submitted at least 48 hours before the deadline will be considered and please make sure to follow the guidelines stated on Piazza.

The submission deadline is **on Monday 13, Feb, 2023 16:59 Eastern Time (ET) /**

Monday 13, Feb, 2023 23:59 Rwandan Time (CAT) .

1. Intraday on-shore wind power generation measured every hour for one year is available from the csv file WindGeneration.csv. Load the data into your computer and produce a graphic showing the time series of the wind generation over time. Is there evidence of annual seasonality?

2. Plot the change in wind generation over time as a percentage of the maximum generation. Is there evidence of annual seasonality?

3. Consider positive and negative ramps in wind power generation, x(t), as a percentage of the maximum, over the hourly timescale. An hourly ramp is therefore defined as r(t,d) =100*[x(t+d)-x(t)]/max(x) where d=1 for an hourly sampling period. Construct empirical cumulative distribution functions (CDF) for both the positive and negative ramps and plot these with the probability on a vertical logarithmic axis. Plot the CDF for a normal distribution with mean-zero and standard deviation from the observations. Is the normal distribution a good model for wind power extremes?

4. National power system operators are tasked with the challenge of balancing supply and demand. They need to understand the variability in wind generation over different timescales. Investigate variability over timescales from one hour to one day by plotting the 1%, 5%, 95% and 99% percentiles. This can be achieved using distributions of the ramps r(t,d) with d =1,2,…,24.

5. Calculate and plot the autocorrelation of wind generation for lags over 10 days. Comment on the structure of the autocorrelation.

6. Calculate and plot the autocorrelation of change in wind generation for lags over 10 days. Include horizontal lines to detect statistically significance values ($p<0.05$). Is there any evidence of diurnal seasonality? Might it be more appropriate to model the change in wind generation than the wind generation?

7. Use a variance ratio test to investigate the structure of the wind generation time series. Can the null hypothesis of a random walk be rejected? Is there evidence of either mean-reversion or mean-aversion?

8. Estimate the optimal window for a simple moving average. Is there a simple benchmark that improves on the persistence benchmark?

9. Evaluate the mean-Absolute-error (MAE) performance of the persistence benchmark forecast over forecast horizons from one hour to one day. Plot MAE as a percentage of the maximum generation for the persistence benchmark.

10. Loop over the number of parameters to use in an ARIMA model for describing wind generation and use information criteria (AIC and BIC) to find the optimal ARIMA model.