CARNEGIE MELLON UNIVERSITY
BIG DATA SCIENCE (COURSE 18-788)
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_BDS_AssignmentNo. For example, mcsharry_BDS_Assignment1, mcsharry_BDS_Assignment2 and mcsharry_BDS_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 27, March, 2023 17:59 Eastern Time (ET) /**

Monday 27, March, 2023 23:59 Rwandan Time (CAT) .

QUESTIONS

1. Load the two historical monthly datasets for each of the thirty districts in Rwanda. The files **RwandaDistrictRainfall.csv** and **RwandaDistrictVegetation.csv** provide measurements of rainfall and the enhanced vegetation index obtained from satellite imagery.

2. Plot two time series graphs, one for rainfall and the other for the vegetation index. Each of these two graphs should have a 6x5 subplot of all the districts. Be sure to label them appropriately. (30 graphs in total but displayed in a 6x5 format.)

3. Calculate the mean, median, minimum and maximum of the rainfall for each month of the year (this implies 12 monthly values and not 180 months!). Plot these four variables against time (month of the year) on a graph. Make a second graph showing the same statistical quantities for the vegetation index. Label the graphs to indicate each of the four quantities.

4. Calculate the correlation coefficient, C, for rainfall between each pair of districts. Make a graph to show the correlation values versus the distance, d, measured in km between the pair of districts. Fit a model of the form, $C(d) = C_0 \exp(-ad)$ by estimating C_0 and the decay constant a. Plot this curve on the graph to show how quickly the correlation declines with distance. Also provide your estimate of C_0 and the decay constant a.

5. Synchronize the dates corresponding to both time series and make a scatter plot of vegetation index against rainfall for the same months. Again use different colors and symbols to distinguish between the different districts and create a figure legend.

6. Can you create a new feature by transforming the rainfall time series that provides better predictions of the vegetation index? One idea is to use a delayed time series to test rainfall in month t-k against vegetation index in month t. What is the optimal value of k months for each district and is there a consensus? Use the correlation to inform your decision. (The range is 12. Test the delay when k ranges from 0 to 12)

7. Another idea for a new feature is to take moving averages of rainfall over the last n months. What is the optimal value of n for each district and is there a consensus? Use correlation to inform your decision. Provide a graph instead of a table showing correlation for each n months. Consider the moving average window, n, ranging from 1 to 12.

8. Is there any evidence for using a quadratic model to describe how the vegetation index varies with rainfall (or any of the above features: delayed rainfall and simple moving average rainfall)? What about higher order nonlinear relationships? Kindly limit consideration to cubic for the highest order. (Use the table format for the answer suggested below).

The performance metrics used (Adjusted R-squared, RMSE and R-squared) should be provided in the format of a table as shown below:

Variable	Linear	Quadratic	Cubic
Rainfall			

Delayed Rainfall			
SMA Rainfall			

9. For each district, use cross-validation to select the best transformation, combining moving averages and delays of the monthly measurements. Provide tables as suggested. Please remember that it makes sense to first consider the SMA and then a delay afterwards when combining the moving averages and delay approaches.

Adjusted R-squared, RMSE and R-squared - (Each should have its table as shown below)

Variable	Linear	Quadratic	Cubic
Delayed SMA			
Delayed Rainfall			
SMA Rainfall			
Rainfall			

10. Describe the optimal model that you would recommend for predicting the vegetation index. Consider linear, nonlinear and nonparametric models. Report the performance of this model using an appropriate measure such as the coefficient of determination, R2, or the RMSE? (Plot the graphs with the fitted models with vegetation against the rainfall feature. We will either use the rainfall, delayed rainfall, SMA rainfall or Delayed SMA depending on which feature was better) Note: Use the best variable from Q9 to help make this conclusion and combine all the information in a format as suggested below:

Variable	Linear	Quadratic	Cubic	Non-parametric 1	Non-parametric 2