CARNEGIE MELLON UNIVERSITY
DATA ANALYTICS (COURSE 18-899)
ASSIGNMENT 1

You should submit, using Canvas, a report in the form of a PDF document (Student_ID-Name-DAassignment1.pdf). Include a cover-sheet on the assignment with your name and the required details. Number the pages, graphs, tables and answers carefully to correspond with the questions. Each answer should be supported by Matlab or R or Python code, graphs and calculations. The submission deadline is 23:59 Rwandan Time (CAT) on **Monday 27 January 2020**. If you prefer to use R and Python for this assignment, the report should provide a list of the non-built-in libraries you used in your code.

1. Download historical daily weather data for France. For example, the analysis could be based on the weather in Paris by using the **attached CSV file**. 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:
http://clients.rte-france.com/lang/an/visiteurs/vie/vie_stats_conso_jour.jsp
Save it as a csv file and load it into your computer.

4. Synchronize 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 modeling approach is not over-fitting? Describe two approaches that could be used to prevent over-fitting?