First Group Assignment

This is your first of three homework assignments (in addition to the reading assignments). The value of this assignment is 6%. Each group will receive its own CUSTOMIZED DATASET to be analyzed using the R language and environment for statistical computing and graphics.

Context

Your team was recently hired for a project on critical infrastructure protection to enhance resilience against cyber threats to electric power grids, public water utilities and smart transportation systems. Such cyber-physical systems routinely rely on Supervisory Control for their continuous operation. Realizing that a security breach may be unavoidable, the risk mitigation approach taken here uses **anomaly-based online intrusion detection** based on monitoring and analyzing control signals streamed in real time from the continuous operation of a cyber-physical system. The sample dataset made available is extracted from supervisory control data describing electricity consumption for households. Extended versions of the this dataset will be studied using increasingly advanced analytic methods as we progress through the course project. This is the first building block.

Submission

Please complete the three tasks described below, create a PDF describing your solutions, and submit the PDF and also the R code of your solutions through the course page by 8 FEBRUARY 2021, 23:59 PST.

Data Exploration

The goal of this assignment is data exploration. The purpose of the data exploration phase is getting a better understanding of basic data characteristics. Besides the quality of the data, like completeness, validity, accuracy, consistency, availability and timeliness, this also includes aspects such as trends, seasonality, feature correlation and more. Technically, the electricity consumption data considered here form a *multivariate time series*[1] describing power consumption behaviour observed over time, one datapoint per minute for each listed variable. The time-dependent variables (also called *response*) are the following ones:

A. Global_active_power
B. Global_reactive_power
C. Voltage
D. Global_intensity
E. Submetering 1
F. Submetering 2
G. Submetering 3

[1] A multivariate time series has more than one time-dependent variable. Each variable depends not only on its past values but also has some dependency on other variables.

The data you need for this assignment is available from the course home page. From the dataset provided, you need to extract data spanning **one full week** from Monday to Sunday. The **week assigned to your group** is determined by your group number — e.g., Group 7 works with the data for the 7th week. On this data, complete all of the following tasks using R:

1. Compute the arithmetic and the geometric mean and the standard deviation for features A, B and C respectively. For features A and B compute the min and max values on weekdays and weekend days during day hours and night hours respectively. [1%]

   The command in R to read a ".txt" file is the following one:

   ```
   read.table(fileName, header = )
   ```

   In order to extract specific days from a time series you will need this command:

   ```
   as.POSIXlt(date, format = "")
   ```

   See also: https://stat.ethz.ch/R-manual/R-devel/library/base/html/strptime.html

2. Compute the correlation for each disjoint pair of the responses, A, B, C, D, E, F and G, using *Pearson's sample correlation coefficient*[2] as defined below. Represent the results of the correlation analysis in terms of a **correlation matrix**[3] and visualize the relevant part of the matrix using color-coding to show statistical significance.

   If we have a series of n measurements of two discrete random variables X and Y, written as x_i and y_i for $i = 1,2,...,n$, then the sample correlation coefficient can be used to estimate the **population Pearson correlation** r_{xy} between X and Y. The sample correlation coefficient is a measure of the linear correlation between X and Y, and can be written as

   $$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

   where n is the sample size; x_i, y_i are the sample points; and $\bar{x}$, $\bar{y}$ are the sample means of X and Y. [2%]

   The following command in R is used to calculate Pearson's correlation.

   ```
   cor(var1, var2, method = "")
   ```

---

[2] The Pearson correlation coefficient is a measure of the linear correlation between two variables X and Y. It has a value between +1 and −1, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation. It is widely used in the sciences.

[3] A correlation matrix is a table showing **correlation coefficients** between variables. Each cell in the table shows the correlation between two variables. A correlation matrix is used to summarize data, as an input into a more advanced analysis, and as a diagnostic for advanced analyses.

3. Focussing on **Global_intensity**, pick a fixed time window that shows a clearly recognizable power consumption pattern during weekdays and weekend days. A sensible choice for a specific time period is between 2 and not more than 6 hours. You need to explore different time windows by plotting the time series for Global_intensity in order to determine a good time window.

   Next, compute the **average** Global_intensity value for each data point in these two time windows over the five weekdays and also over the two weekend days. Assume you choose 7:30 AM to 12:30 PM as the time window. To create the new time series for weekdays, calculate the average of the values at the time 7:30 AM for all five weekdays, then the average values of 7:31 AM, 7:32 AM, etc.

   Finally, perform a linear regression based on the *least squares method* (LSM) and also a polynomial regression for each of the two time series you created. Plot the results of the two linear regressions in one diagram by overlaying two linear regression lines. Likewise, plot the two polynomial regression curves in one diagram. The goal is to visualize Global_intensity patterns in an intuitively interpretable graphical format. **[2%]**

   The commands in R for performing a least squares regression and polynomial regression are the following ones:

   Linear Fit

   fit_linear <- lm(y ~ x, data)

   Polynomial Fit

   fit_polynomial <- lm(y ~ poly(x, d, raw=TRUE, data) (d is the degree of the polynomial regression which is greater than 1)

Please submit a report describing your findings and rationale for your choices in one document and the R code as a separate file on the course page.

Thank you!