First Group Assignment

This is your first of three homework assignments (in addition to three reading assignments). The value of this assignment is **5%**. Please use the Household Electricity Consumption Dataset available on the course page. This data 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 critical infrastructure routinely relying on *supervisory control* for its continuous operation. Realizing that a security breach may be unavoidable, the novel risk mitigation approach taken here uses **behaviour-based online intrusion detection** by 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 term project. This is the first building block.

Submission

Please complete the tasks described below, create a PDF describing your solutions, and submit the PDF and R code of your solutions through the course page by end of day (23:59) on Thursday, JANUARY 30, 2025.

Data Exploration and Preparation

The goal of this assignment is data exploration and preparation. The purpose of this phase is getting a better understanding of the 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 represents a *multivariate time series*[1] describing the power consumption behaviour observed over time, one datapoint per minute. 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.

On the data provided, complete all of the following tasks using R:

1. The goal of this task is to apply statistical methods to detect point anomalies. You will learn to handle missing data through interpolation and identify anomalies using the concept of standard deviation, specifically focusing on the use of **Z-scores**.

   The uploaded dataset, Household Electric Power Consumption data, comprises several features related to power consumption in a household.

   Apply **linear interpolation** to fill in all missing (NA) values for each feature in the dataset. This method estimates missing values using a linear function based on the non-missing data points adjacent to the NAs (ensure that the interpolation is performed separately for each feature).

   After addressing the missing values, the next is to **detect point anomalies** in the dataset. For this purpose, you will use the concept of standard deviation as a measure to identify outliers.

   Calculate the Z-score for each data point in each feature. The Z-score is a statistical measure that describes a data point's relationship to the mean of the group of data points. A data point is considered a point anomaly (or outlier) if its Z-score is more than 3 standard deviations away from the mean (in mathematical terms, a data point can be considered anomalous, if $|z| > 3$ and it implies that the data point is significantly different from the dataset's average behaviour).

   Upon identifying the anomalies, calculate the percentage of data points that are considered point anomalies based on the criterion above for each feature and for the entire dataset. This will give you insight into the proportion of the data that behaves unusually. **[2%]**

From your preprocessed dataset 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). 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 as defined below. Represent the results of the correlation analysis in terms of a **correlation matrix**[2] 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**[3] 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. **[1%]**

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

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

3. Focussing on **Global_intensity** only, determine representative time windows, one for day hours and one for night hours respectively, that illustrate a typical power consumption pattern over a time period of several hours.

   Next, compute the average Global_intensity value for each data point in these two time windows over the five weekdays and also over weekend days (assume you choose 7:30 AM to 5 PM as the day time window for weekdays. To create the new time series for day time of weekdays, calculate the average of the values at the time 7:30 AM for all five days, then the average values of 7:31 AM, etc.)

---

[2] 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] 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.

Finally, perform a linear regression based on the *least squares method* (LSM) and polynomial regression for each of the resulting four time windows and represent the results (four linear regression lines, four polynomial regression curves) graphically in two diagrams as an illustration of Global_intensity behaviour. **[2%]**

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

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 the report and the code **through the course page** by JANUARY 30, 2025.

Thank you!