UNIVERSITY OF BRITISH COLUMBIA
Department of Statistics
Stat 443: Time Series and Forecasting
Assignment 1: Exploratory Data Analysis

The assignment is due on **Thursday, February 4** at **9:00pm**.

- Submit your assignment online on `canvas.ubc.ca` in the **pdf format** under module "Assignments".
- This assignment should be completed in **RStudio** and written up using **R Markdown**. Display all the R code used to perform your data analysis.
- Please make sure your submission is clear and neat. It is the student's responsibility that the submitted file is in good order (i.e., not corrupted).
- **Late submission penalty**: 1% per hour or fraction of an hour. (In the event of technical issues with submission, you can email your assignment to the instructor to get a time stamp but submit on canvas as soon as it becomes possible to make it available for grading.)

1. Rimouski is a city in Quebec, located in the Bas-Saint-Laurent region. Please go to the climate database of the Government of Canada (https://climate.weather.gc.ca/) and download the **monthly** temperature series for Rimouski (with station name "Rimouski" and climate ID 7056480) from January 1954 to December 2016.

   (a) Read in the data and create a time-series object for the monthly mean max temperature. Plot the series against time (label your graph) and comment on any features of the data that you observe. In particular, address the following points:
   - Does the series have a trend?
   - Is the series stationary?
   - Is there a seasonal variation? If so, what is the period of the seasonal effect?
   - Would an additive or multiplicative model be more suitable to decompose the series? Justify your answers.

   (b) Are there any missing values in the series? If so, identify the year and month of all the missing values. One commonly used imputation method is the LVCF (last value carried forward), which imputes the missing data with the last observed value. Is it appropriate to use the LVCF imputation method here? Justify your answer. If not, suggest an adapted version of the LVCF method that would be appropriate to use here. Apply the imputation to obtain a complete series.

   (c) Create training and test datasets. The training dataset should include all observations from the year 1954 to 2015. Use the observations of the year 2016 as the test set. You can use the command `window()` on a `ts` object to split the data.

   Using an additive model, decompose the (imputed) series into trend, seasonal, and error components. Use both moving average smoothing (R function `decompose()`) and the loess method (R function `stl()`). Plot both decompositions.

(d) Fit a linear model to the trend component (R function `lm()`) of the moving average decomposition. Does the linear model provide evidence of a trend at the 95% confidence level? Without doing any further analysis, would you use this trend component to make predictions? Justify your answer using the linear model results and/or the trend component plot.

(e) Predict the monthly mean max temperatures for the test dataset using the moving average decomposition. Compare your predictions under the following assumptions:

- There is a linear trend;
- There is no trend.

Compare your predictions with the test dataset graphically and compute the sample mean squared prediction error (MSPE), defined as the average of the squared distances between the predictions and observations in the test dataset. Which model has a smaller MSPE?

(f) In time series analysis, it is often assumed that the error term follows a Gaussian white noise process, i.e., $Z_t \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma^2)$. Generate the correlogram and Q-Q plot of the error component from the moving average decomposition, and discuss whether the Gaussian white noise assumption is appropriate.

2. The file `GSPC.csv` contains 2527 daily closing prices of the Global S&P500 index spanning a decade from January 2, 1985 until December 29, 1994.

(a) Create a time series object for the adjusted daily closing price of the S&P500 index and plot it over time. Comment on the features, including stationarity, trend and seasonality.

(b) Transform the original series to daily log-returns by taking the logarithmic difference, i.e., $X_t = \ln(S_t/S_{t-1})$, where S_t denotes the adjusted daily closing price at time t. Plot the daily log-returns over time and comment on the stationarity of the series. Why is it more convenient to work with the daily log-return series?

(c) Generate the correlogram for both the daily log-return series $\{X_t\}$ and the absolute value of daily log-returns $\{|X_t|\}$. Compare and comment on what you observe.