Practical #1: Analyzing stock market indexes

Instructors: Nicolas Navet (nicolas.navet@uni.lu), Long Mai (long.mai@uni.lu)

The goal of the practical is to gain insights into stock market index data using descriptive statistics and correlation analysis. The data are the Open, High, Low, Close values of daily stock market indexes. We will focus in the following on a time series made up of the end-of-day closing values. The assignment is broken down into a number of questions:

1. Load the data file assigned to you into a data frame named "eod" (for end-of-day) using the read.csv2() function. *Hint: in R-Studio, you can change the current working directory in the menu "Session" so as to work in the directory where the data file is stored.*
2. Exploratory data analysis:
    a. Plot the column eod$Close. What do you observe?
    b. What is the number of entries in the data frame?
    c. What is the max variation between the min and max value of eod$Close?
3. What would be the gain or loss of an investor that would have invested 100€ in the index at the very beginning of the data and sold them at the very end? What would have been the highest value of its investment over the whole period? *Hint: consider only the Close values to answer that value even if it yields an approximate answer.*
4. The daily end-of-day return is defined as r_t = (Close_t-Close_t-1)/Close_t-1.
    a. Compute the daily end-of-day returns for your data. *Hint: one way to do that is using the diff() function but you can do differently.* The sequence of returns should be stored in a vector named "returns".
    b. What is the min and max return over the complete data set?
    c. How many returns are there above 4%?
    d. How many returns are there below -4%?
5. Plot the empirical distribution histograms of the returns (both positive and negative) with the "FD" heuristics for the number of bins, as well as 1) the boxplot of the positive returns and 2) the absolute value of the negative returns. What do you learn from that?
6. Divide the vector "returns" in three contiguous vectors of same size (possibly modulo 1 entry) named: "returns_1", "returns_2", and "returns_3". Check that the newly created vectors have the correct size before moving forward. Plot the autocorrelation within each of the three sets up to the lag 100. Overall, are the autocorrelations significant? Are the autocorrelations substantially different among the three sets? *Hint: you can gain insights into whether the autocorrelation is of random nature in the original data set by analyzing the autocorrelation of a shuffled version of the original data set (i.e., entries are re-ordered randomly).*
7. Compute the Pearson correlation coefficient between the returns of the index of your data file and the returns of the NYSE composite index (file "^NYA.txt") on the set of days in common over the whole period. *Hint: the first step is to discard the data that are not in both time series, this can be done by an R function written for that purpose.* Do the different time zones play a role?

Additional information:

Your report should include the answers to the questions and the corresponding R scripts. The report in pdf format must be submitted by email by October 24th, 2019 to both instructors (we will always acknowledge the good receipt). This is an individual project and solutions must not be shared amongst you. You can of course use all the external sources you need**.** Do not hesitate to ask us for guidance during the class or by email, our role is to help you achieve the objectives. **The weight of this assignment in the final grade will be 1.**

The grade will consider the correctness of the answers to the questions, the quality of the code and the report (writing, presentation). An individual oral presentation of the results may be organized.