# BTMA 531 Assignment 4

Due April 5, 2023 by noon on D2L

## Instructions

- You should create a single R script (or R Markdown if you wish) called {firstname}_{lastname}_Asgn4.r, which has the required code for all parts of the assignment.
- Make sure to use commenting (#) so that your R script file can be understood by someone else. Yo do not need to comment on what you are trying to do in each line, but it should be clear where the answer for each question is. Provide the written answers in comments (#) within the R script file.
- Make sure that your R script is executable from top to bottom on another computer. Make sure all requirements of questions are done using R (e.g. do not use calculator, excel, ... to calculate things).
- The assignment submission should be done through D2L dropbox. Upload the R file to the assigned dropbox folder in D2L.
- The purpose of the assignments is to help you learn through practice. You may want to use R help or search online for answers. However, note that this is an individual assignment. The work you submit should be 100% yours. Do not copy, share, or ask for files, chunks of code, or answers. Refer to the course outline for some examples of what to do and what not to do, and to Code of Student Conduct for more information on cheating and plagiarism. If you are note sure about a behavior, please ask.
- The focus is on the response to the question, and not on the approach you use to get there. However, unless otherwise stated, you should use only code to get to the response (that is, you cannot use a calculator for example to do some intermediate calculations, should not hard code numbers, and everything needs to be done in R).
- You are allowed to use online resources to find answers or chunks of code. However, make sure that you reference the source in your script using commenting (#). This is generally a good practice in coding, and for the purpose of this course, removes any confusion when several students use the same online source and therefore have similar codes in their scripts.

## Questions

**1** [50] Use the attached dataset. This is a revised dataset based on the Beijing Multi-Site Air-Quality Data Set at https://archive.ics.uci.edu/ml/datasets/Beijing+Multi-Site+Air-Quality+Data.

a) [7.5] Create a time-series object that includes only the temperature (*TEMP*, univariate). For frequency, use 12 (monthly data), there is no need to capture the start or end dates or to include the year, month, hour data. Then update the time-series by interpolating the missing values using *na.interp()* function from the *forecast* package.

b) [5] Plot the time-series for temperature. Then plot the moving average time-series using the moving average of the last 2 observations.

c) [7.5] Plot the ACF and PACF plots for the data. Is there any autocorrelation in the data? Test to see whether the time-series is stationary. Do we need to difference the data?

d) [10] Fit an ARIMA model to the data (you can use auto.arima). What orders are used for the AR and MA models, and is the data differenced? Is a seasonal model used? Use the model to make predictions for 24 months, and plot the predictions alongside the actual data.

e) [10] Decompose the time-series to its trend, seasonality, and random effects using the classical seasonal decomposition by moving averages. Then use the Holt-Winters model to forecast the temperature for the next 24 months. Plot the forecast along with the actual data.

f) [10] Inspect whether temperature has an impact on wind speed max ( *WSPM*) using time-series regression. You can remove the missing values from *WSPM* using *na.interp()* as well. Use the *auto.arima()* function to create the regression results. Does temperature impact WSPM at the 95% significance level?

**2** [20] Use the included “Groceries” dataset (within the arules package) for this question.

a) [10] Use the Apriori algorithm to find the association rules with a minimum support of 0.02 and minimum confidence of 0.4.

b) [10] Take one of the rules with size (the total number of items in the LHS and RHS) of 3, and explain its measures of support, confidence, and lift.

**3** [30] Use the attached “TextData” dataset for this question.

a) [7.5] Remove white spaces, stopwords, and numbers from the documents. Make the text all lowercase, and then stem the text.

b) [2.5] Create the document-term matrix. Find the frequent terms in all documents. Find highly associated terms (correlation more than 0.5) with two of the frequent terms (your choice).

c) [5] Create a word cloud of the terms.

d) [5] Cluster the terms using hierarchical clustering, with 5 clusters.

e) [5] Cluster the documents using K-means clustering, with K=5.

f) [5] Analyze the sentiment of all documents using the *syuzhet* package. Plot the sentiments as a bar plot.