Assignment 1: Exploratory Data Analysis

In this assignment you will be conducting exploratory data analysis on several different datasets. This data can be found at /ulteosrv1/s0/meteo515/data. You may complete the assignment using your programming language of choice. Feel free to use built in functions but make sure you have read the documentation about these functions and are confident they are indeed conducting the calculations you intend. Please submit your assignment and the code used to generate any results by uploading the files to Canvas by the assignment due date. Your code should be well commented so that others can easily understand what has been done and marks may be removed from your assignment if this is not the case.

1. You have been provided data from the Penn State weather station going back to the late 1800's. This data is in the class data folder and is called SC_data.xlsx. The first column of the data is the date, the second is the maximum daily temperature (Tmax), the third is the minimum daily temperature (Tmin), and the forth is the total daily precipitation (PCP). Trace precipitation amounts are denoted by -1 and missing values are denoted by -99.

- Calculate a set of basic statistics on the raw Tmin, Tmax, and PCP data. This should include the mean, median, standard deviation, interquartile range, median absolute deviation, skewness, and Yule-Kendall index. Visualize the data through the production of schematic plots of each of Tmin, Tmax, and PCP. Include the schematic plot figure and a table of the basic statistics in your assignment.
- Repeat your analysis above but with missing values removed from the calculations. For precipitation, conduct the new analysis only for values > 0. Include the new schematic plots and table. How did removing the missing values influenced your calculations? How did removing non-precipitating days impact your precipitation distribution.
- Briefly describe what this analysis tells you about the distribution of the data and which statistics might be the most appropriate for each dataset.

2. In the class data folder, there is a standardized index of the NAO called nao.long.data.txt originally obtained from the ESRL at: https://www.esrl.noaa.gov/psd/gcos_wgsp/Timeseries/NAO/. The data is on a monthly timescale and is already standardized. Also in the class data folder is the unsmoothed AMO index called amon.us.long.data.txt originally obtained from the ESRL at https://www.esrl.noaa.gov/psd/data/timeseries/AMO/. This time series has been detrended but not standardized.

- In this question we will be examining both indices from 1900-2015 and so both sets of data can be imported and reduced to this time period. Compute the annual average timeseries both the AMO and the NAO. Normalize the AMO and re-normalize the NAO after having done the annual average. These new AMO and NAO indices will be used in the remainder of the question. Create a plot of the AMO and NAO indiex timeseries and produce a table with the same set of basic statistics for these datasets as was done in question 1.

- To examine the distribution of the data, we will employ a kernel density smoothing to each of the NAO and the AMO. Explore the sensitivity to the value of the smoothing parameter h. Include 5 figures with values of h that span the range of being too small, just right, too large. This can be decided upon subjectively. Describe your results.
- Compute and plot the autocorrelation function for both the NAO and the AMO. Try this calculation with two different types of correlation one being a Pearson correlation and the other being either a Kendall's Tau or a Spearman's rank. Compare and contrast the NAO and AMO autocorrelation functions as well as the two correlation methods used.