Assignment 1 (50 pts)

In this assignment you will be working with station data the Penn State weather station from 1980 to 2015. This data is in the class data folder and has a filename SC_data_1980_2015.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 (if applicable) are denoted by -99.

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. I should be able to re-create all of your results given the scripts you have provided me, this includes any data import commands. For your data loading, please ensure that your path is a relative path in the form '../../data/SC_data_1980_2015.xlsx'. 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. (18 pt) In this question we will explore the relationship between temperature and precipitation in State College. In this analysis you will only consider data in the winter months (December, January and February) will be considered in the analysis. Furthermore, any precipitation values less than or equal to $0.01 mm/day$ will be considered non-precipitating and thus set equal to 0.

a) Create a Venn diagram to represent the following situation. There are 4 categories of maximum daily temperature Tmax <30F, 30-40F, 40-50F, $\geq$ 50F forming 4 MECE events (T_1, T_2, T_3, T_4). Precipitation occurrence (P) has some intersection with each of these categories.

b) Compute the probability of each of these Tmax categories (ex. $Pr\{T_1\}$) as well as the conditional probabilities of precipitation occurring given Tmax in each category ($Pr\{P|T_1\}$, $Pr\{P|T_2\}$, $Pr\{P|T_3\}$, $Pr\{P|T_4\}$). Provide a table for each of the probabilities and conditional probabilities in your report.

c) Using the values computed above and applying the law of total probability, compute the $Pr\{P\}$. Note: from here it is easy to verify the law of total probability and your calculations using the data itself.

d) Using the values calculated in b) and applying Bayes theorem, compute $Pr\{T4|P\}$. Compare this to $Pr\{T4\}$, how does knowledge about the occurrence of precipitation change your expectation of the temperature being in this highest category?

2. (24 pt) In this question, you will explore the distribution of the data by computing some basic statistics and creating schematic plots of the distribution.

a) 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. For the calculations, ignore missing data and only consider precipitating days, defined as those having values greater than $0.01 mm/day$. Provide a table with these calculations in your report. Briefly describe what this analysis tells you about the distribution of the data and which summary statistics are appropriate for each dataset.

b) Visualize the data through the production of schematic plots. Create a figure (or subplot) for each variable (Tmin, Tmax, and PCP) that consists of 4 schematic plots, one for each season (DJF, MAM, JJA, SON). In these schematic plots you do not need to include a distinguish far out points from outside points, they may all be the same symbol. Describe any new insight you gain from this visualization that might help explain your summary statistics or understand the distribution in general. How does the data's distribution change across the seasons.

3. (18 pt) In this question, you will further explore the distribution of the Tmax data using histograms and kernel density functions. Create a frequency density histogram of the Tmax data and overlay it with a kernel density function both with the same value of bin width (w) / smoothing parameter (h).

a) Explore the sensitivity to the value of the w/h by utilizing sizes of 0.5, 1, 2, 5, and 10. Include all 5 figures with different values w/h.

b) Describe the data's distribution and compare the representation in the histogram vs. the kernel density function.

c) Discuss the impact of w/h on your results and explain. Describe issues with different bin sizes and which size you believe is most appropriate.