CSC 1930.01: Data Science
Spring 2018 – Assignment 5
Dr. Ben Mitchell

Instructions

This assignment is focused on data cleaning and exploratory data analysis (EDA).

This is a paired assignment; the same rules apply as for previous paired assignments. There is no requirement that you work with the same partner as on the previous assignment; in fact, I encourage you to switch around who you work with, since this will be good practice for being a data scientist who needs to work with a diverse group of collaborators.

The goal of this assignment is to start with some 'raw' data, and then clean it up to produce some nice summary graphs. You can download the data from the standard place on Blackboard; the file is called 'temperaturesRaw.csv'. The data is similar to the temperature data we worked with before, except that this version contains data on a per-month basis rather than a per-day basis; it also spans a longer time range, and comes from three different weather stations (though all are in the greater Philadelphia area). It contains missing values, and also a number of readings that are clearly incorrect (the min and max values reported for the temperature columns should clue you in to this fairly quickly).

The goal is to generate a data table with a single (clean) row for each date; that is, you are trying to combine the three weather stations into a single 'average' reading for each month during the time period spanned. You should then generate several plots looking at the resulting data; this should minimally include a line graph showing how the temperatures (monthly Low, High, and Average) have evolved over the time period represented in the data, and box plot showing the distribution of the data from each of those three columns.

As always, the goal of this assignment is not just to produce the clean data, it's to document the process of how you did so. That means your submitted notebook should contain all the things you tried in the process of figuring out what was wrong and fixing it, along with markdown describing your thought process.