Problem Set 3

Data Science 602: Data Analysis and Machine Learning

Spring 2022

1. **Statistical data visualization** In the first assignment, problem 3, you identified an "interesting" pattern in the weather or citations dataset. Recall that such discovered knowledge should be novel, useful, and non-trivial. Develop a explanatory visualization to present the knowledge you discovered. (You may choose a different pattern from the one you used in the first homework assignment if you wish.)

2. **Data preparation** This problem uses the weather dataset from previous problem sets. Prepare a dataset to predict the observed temperature from the following predictors:

   (a) The non-temperature fields from the observation

   (b) The temperature recorded in the prior observation

   Beginning with the weather dataset:

   (a) Add the temperature from the prior reading as a new feature. That is, for each observation at time t_k, $k > 0$, the new feature should have the value of the temperature reading at time t_{k-1}. For the first observation ($k = 0$), the value should be missing because the prior temperature is unknown.

   (b) Because the observed temperature is the target variable, remove the current temperature from the data frame, and save the values into a matrix $\mathbf{y}$

   (c) Treat missing values in the dataframe so that the output dataset contains no missing values. In the notebook, explain your rationale for treating missing values.

   (d) Remove the non-numeric date field, and convert the dataframe to a numpy matrix, $\mathbf{X}$

   (e) Scale the numpy array using a `StandardScaler`.

   Show the first few rows of the resulting matrices $\mathbf{X}$ and $\mathbf{y}$.