

Statistical Computing - Exercises 04 - Ithaca Weather

The US government, via its agencies NOAA and NCEI, maintains the US Climate Reference Network, a collection of 139 high quality weather stations whose purpose is to monitor atmospheric and soil conditions. One of them is located right here outside of Ithaca, along Route 38.

We will be formatting and analyzing the hourly Ithaca data from 2022 and 2023, with the goal of building a model for predicting afternoon temperatures.

1. Read the dataset `datasets/CRNH0203-2022-NY_Ithaca_13_E.txt` into R as a data frame and add column names according to the **documentation**. Make sure you inspect your data frame and do any necessary cleaning.
2. Save the data frame to your computer using both `save` and `saveRDS`. What are the two file sizes, and how do the sizes compare to the size of the original .txt file? (You can check this in the terminal with the command `ls -lah`)
3. Create a data frame containing only the longitude and latitude. Save the result as both a .csv and a .RData file. How do the sizes compare? Is this surprising to you? Can you explain the results?
4. Create a new data frame that has one row for each day. Its columns should be the date, 3pm air temperature (use local standard time for all times and `T_CALC` for all air temperatures), 6am temperature, 7am temperature, 8am temperature, and 9am temperature.
5. Do a multiple regression of 3pm temperature on 6am, 7am, 8am, and 9am temperature and print out the summary. How accurate does the model say your predictions should be?
6. Use your model to predict 2023 3pm temperatures from morning temperatures. You'll have to read in the 2023 data and organize it the same way you organized the 2022 data. Report the root mean squared prediction error and make a plot of the model's prediction errors against day of year.
7. Build your own model for 3pm temperatures based on the 2022 data. You can use any information in the data, and any kind of model you like, but you must use only information that's available at the 9am hour or earlier (e.g. you cannot use today's 2pm temperature to predict today's 3pm temperature).
8. Test out your model on the 2023 data. Report the root mean squared prediction error and make a plot of the model's prediction errors against day of year.