Ran’s Individual Assessment

It was a nice team working experience on this EDA project. I am responsible for the data cleaning and distribution estimation. The data cleaning for this project is challenging, as all of the data is in txt format. Besides, we have to extract all the weekly data of 3 years.

The work I have done includes data input, data formatting, data visualization and model estimation.

Starting with the data input, from the downloaded folders containing weekly data for a year, I iterate through all the files and read them into corresponding variables. To make the data structure reusable, I put variables in the same year (specifically of Florida state since our topic is analyzing Florida’s cooling degree day) in a dataframe, and combine different years’ dataframes into a list. Since the txt files are not in standard table format, I read the whole content as text and skipped the first 16 rows containing redundant information, and split the lines on spaces to get the value of each column. The function keeps steps above until all files in all folders have been read. The related functions are read\_single\_dir, read\_single\_file and split\_columns.

Then I use histogram, boxplot, and kernel density to visualize the data. To my surprise, the data follows a multimodal distribution with two peaks, which is beyond our knowledge. So I did research online and found “modes” package to estimate multimodal distribution. Then I discussed with my teammates and we stared to work separately on each year data. At the end, we compared our analysis to study the difference among the distributions of cooling days per year.