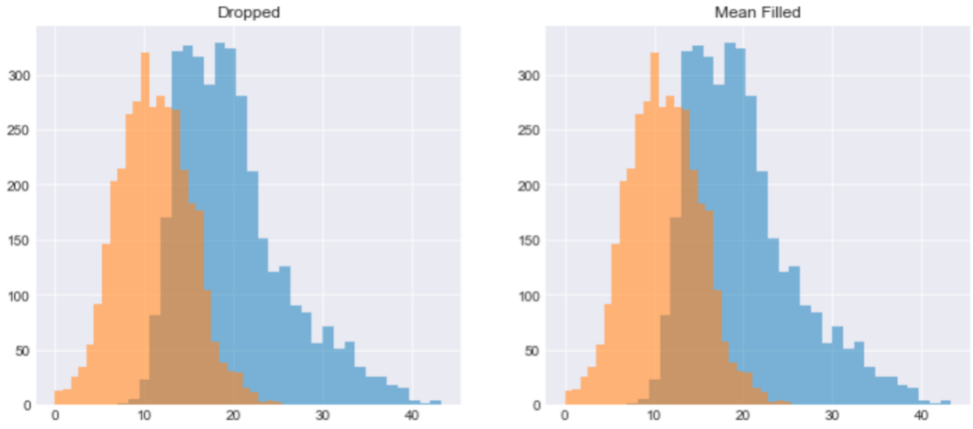
Mia Rodgers: https://github.com/miamrodgers/4310-ML/blob/main/ME2_preprocessing/data_cleaning.ipynb
Alex Larsen: https://github.com/alarsen123/ML-HW/blob/main/ME2_preprocessing/ME2_preprocessing/data_cleaning.ipynb
ME2

This assignment was focused on data cleaning, we used the daily-temperatures dataset. To start, I printed the info of the dataset. I noticed that there were some null objects so I printed the sum of null values in the dataset. There were 7 null max temperatures and 5 null min temperatures. Also, the max temperature dtype was non-numeric, so I converted that column using to_numeric() so it could be graphed.

Next, I printed the 5-number summary. The maximum of max_temperature was 300, and the minimum of min_temperature was -800, which are clearly outliers. Then, I printed the value counts of max_temperature over 100 and min_temperature under -100. There were 16 max_temperatures that were 300 and 7 min_temperatures that were -800, so I removed those observations.

Since there were some null values, we decided to see if there was a difference between the distributions when we dropped the null values and when we filled them with the mean values.



Based on the graph, we saw that there is no difference between the two different methods for this data set.