# TW2 Data Preprocessing

### **Part 2:**

Submit a summary of your learning to Canvas. Your document should include:

* Full names of your team members who work on the assignment.
  + Xiaomei Xie
  + Lili Hao
* URL links to the notebook of each student on GitHub repo.
  + <https://github.com/xiaomeiX/TW2-Data-Clearning>
  + <https://github.com/lhaoSeattleu/TW2-Data-Clearning>
* A summary of what you learned from the teamwork assignment.

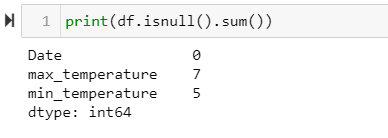
This dataset, namd daily-temperatures.csv, includes 3 columns and 3650 rows:

* Date
* Max\_temperature
* Min\_temperature

This dataset has missing values, mismatched data type values and outliers. We need to clean data and impute missing data before plotting the max temperature and min temperature.

## Missing values

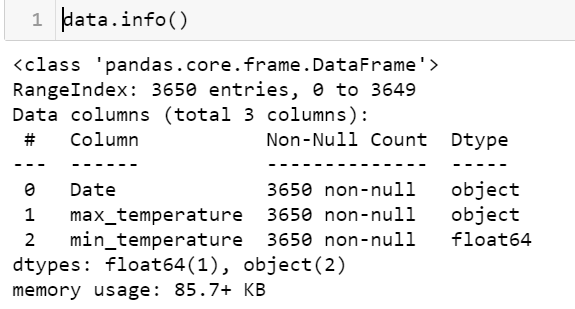
We first fill the null value with 0:



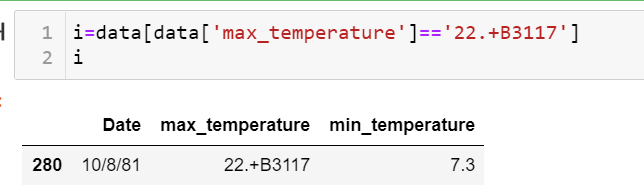


## Mismatched data type

The dataset also has invalid values which has mismatched data type in max\_temperature:



We find the index of the incorrect values.

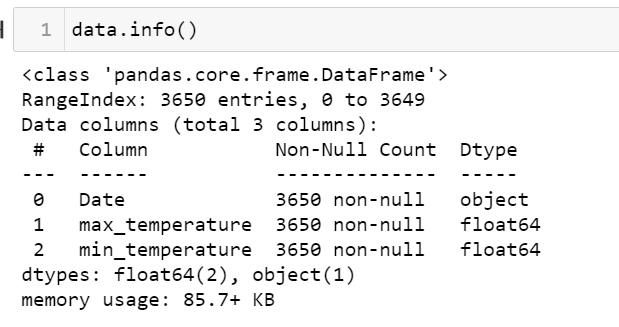


Then we replace the invalid value with an estimated float number

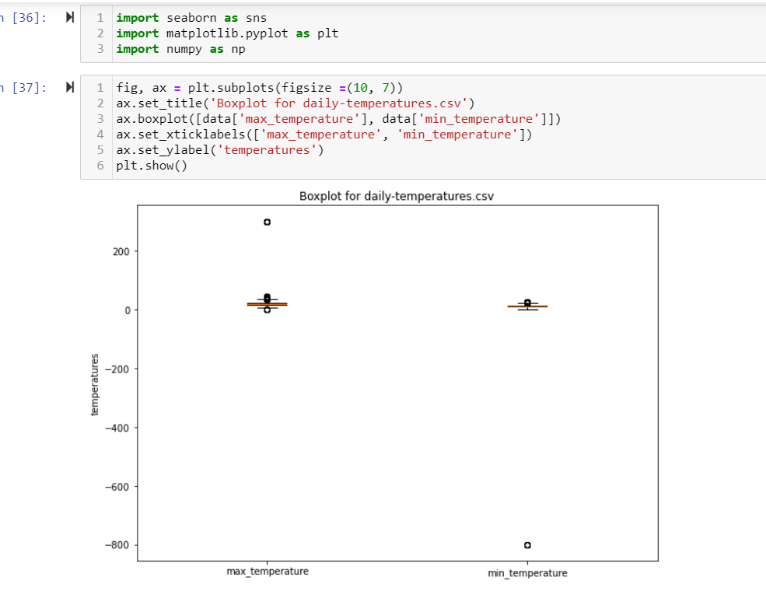


Then we covert the data type into a float.



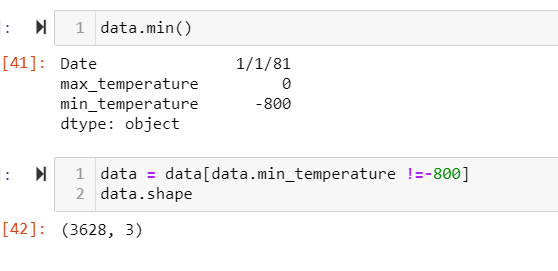
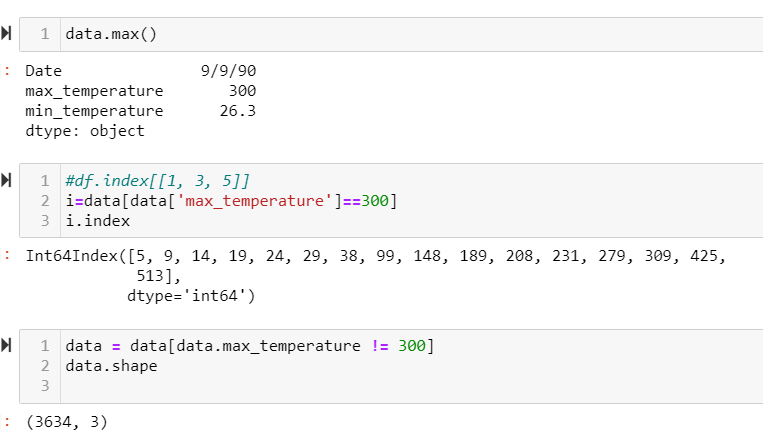


After we get the proper data type for the boxplot, we plot the max\_temperature and min\_temperature.



## Outliers:

After we plot the boxplot, we found outliers in both columns. We decided to delete all the outliers in both columns:



After we clean up and impute the data, the boxplot shows correctly:

