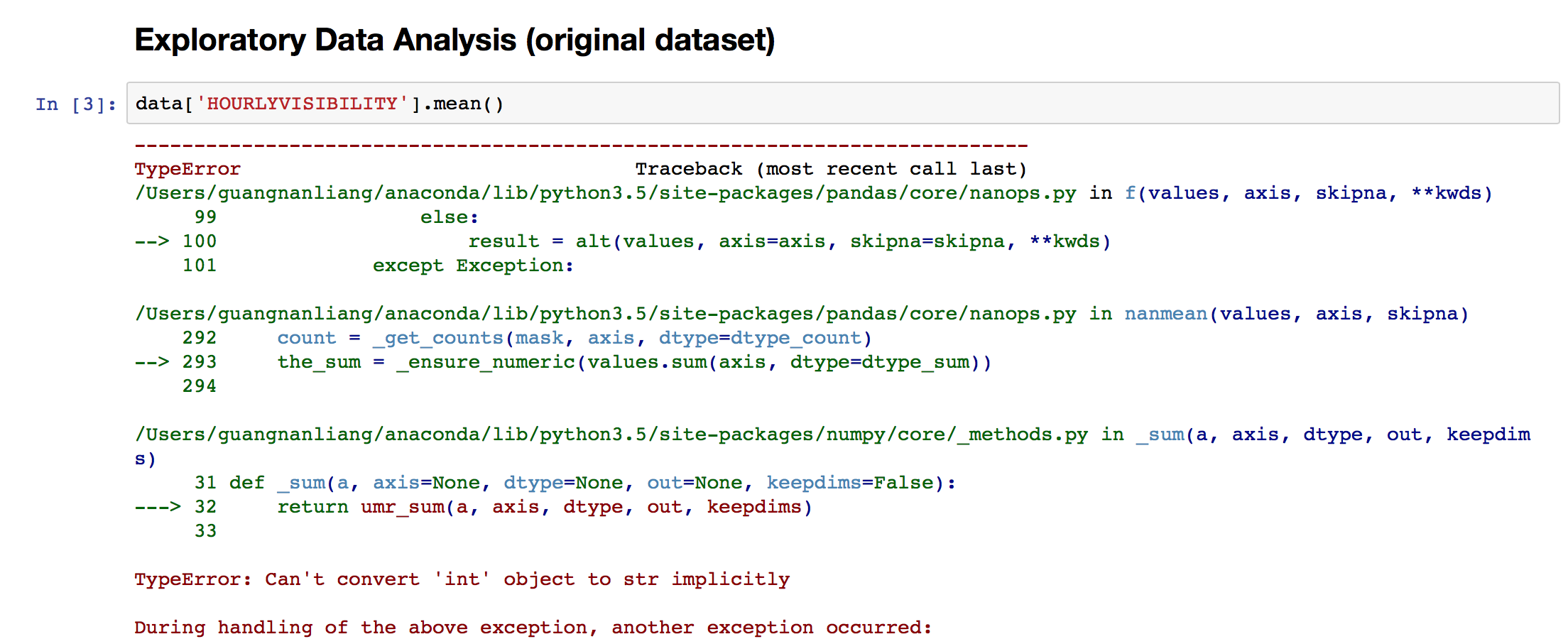
**Report of Data Analysis**

1. **Data Ingestion**

* **Using pandas. read\_csv () to read the original file**
* **Try analyzing raw data:**

The data without preprocessing cannot be analysis and visualization directly, there are too many key, type, and other errors.



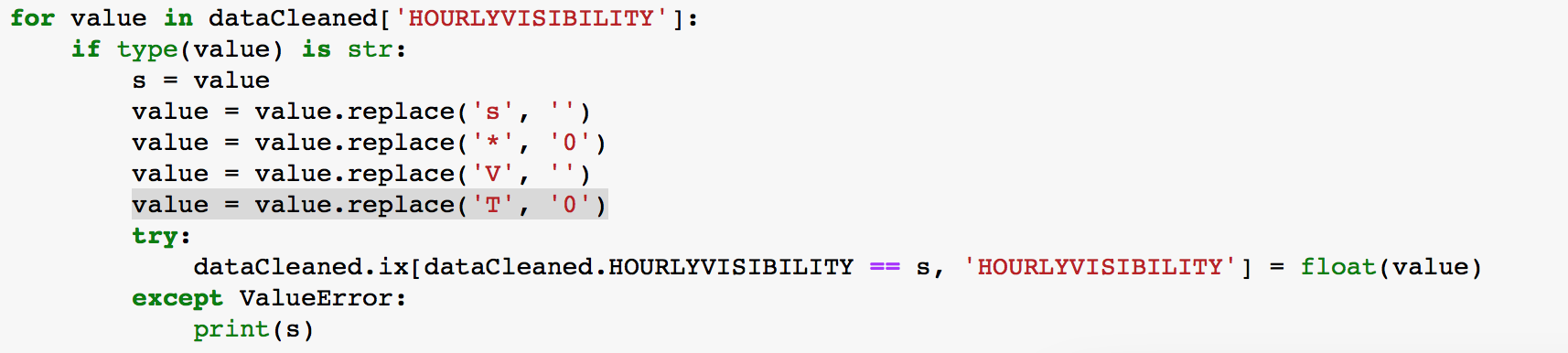
1. **Data wrangling (clean data)**

* **Remove SOD row**

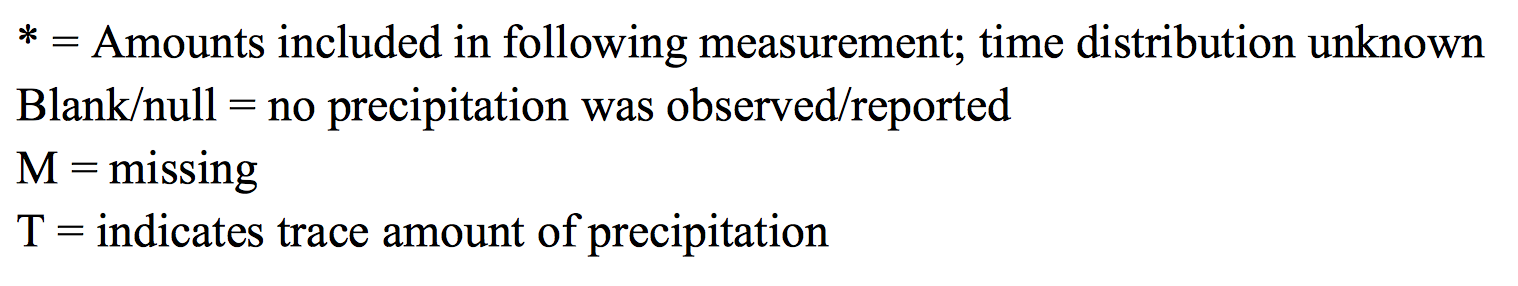
In our assigned station (CHESTER AIRPORT, CT US), the dataset includes SOD rows to summarize the whole day’s data, but the cells are empty. We remove these rows to avoid problems when calculating mean values.

* **Convert string into float**

The data type read into memory is string, so we need to convert them into correct type. Some numeric values also include special characters introduced in documentation. Thus, we need to replace these characters first.



special characters in dataset:



* **Transfer date type from string into datetime and Add extras columns**

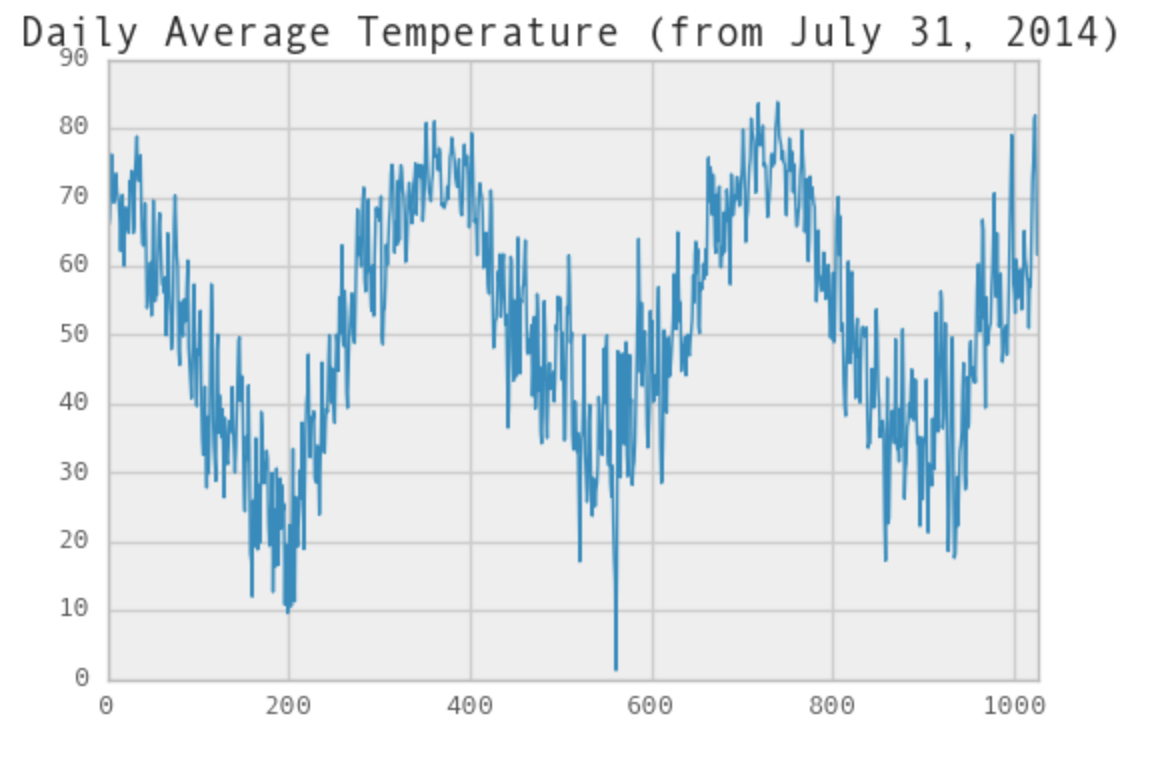
To group by time easily, we create 3 new columns to preset the information of year, month, and day.

* **Save cleaned data in disk**

After cleaning dataset, we save the cleaned dataset in the local disk, and for the later analysis, we just read the cleaned data.

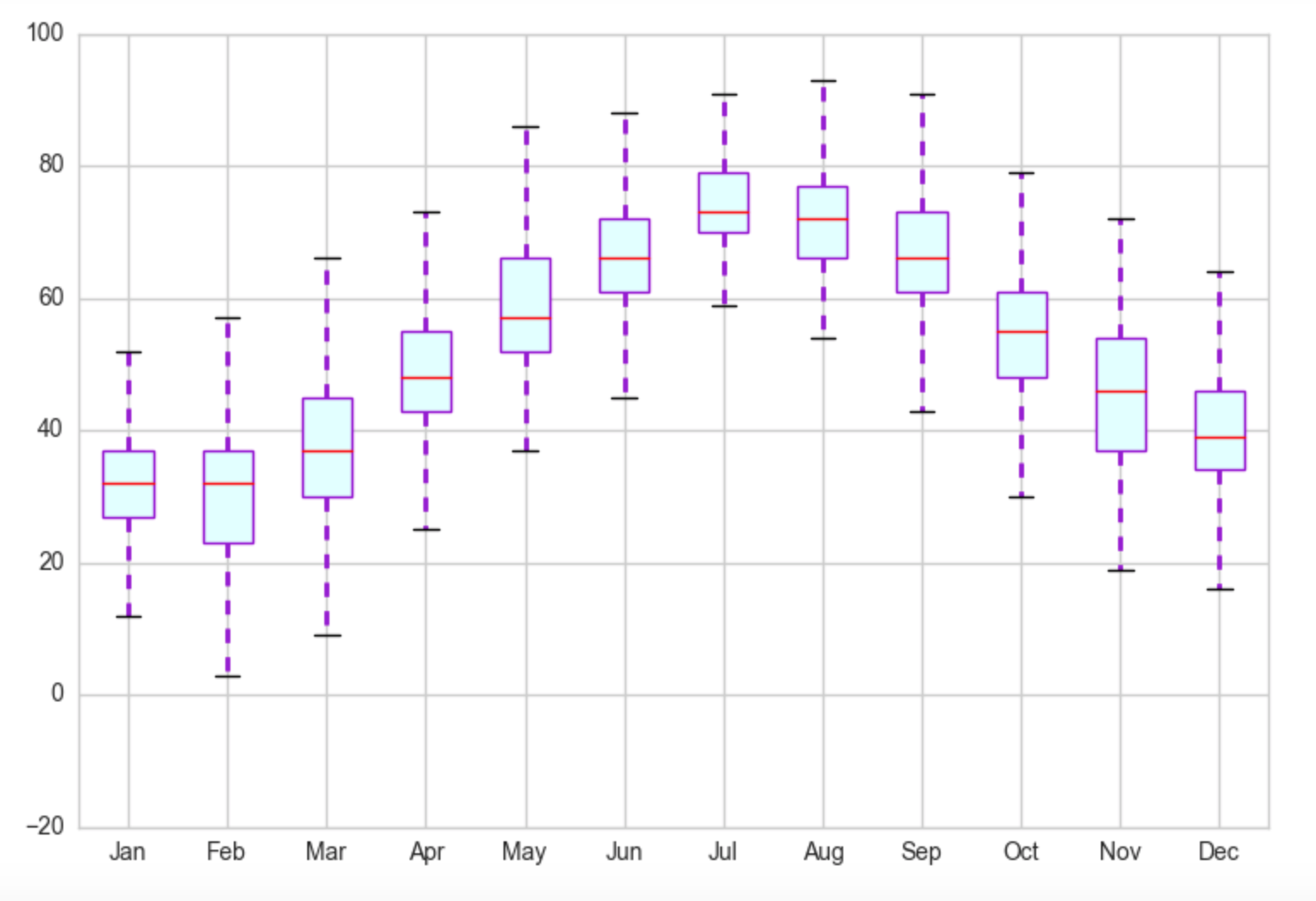
1. **Visualization:**

* **Daily temperature**



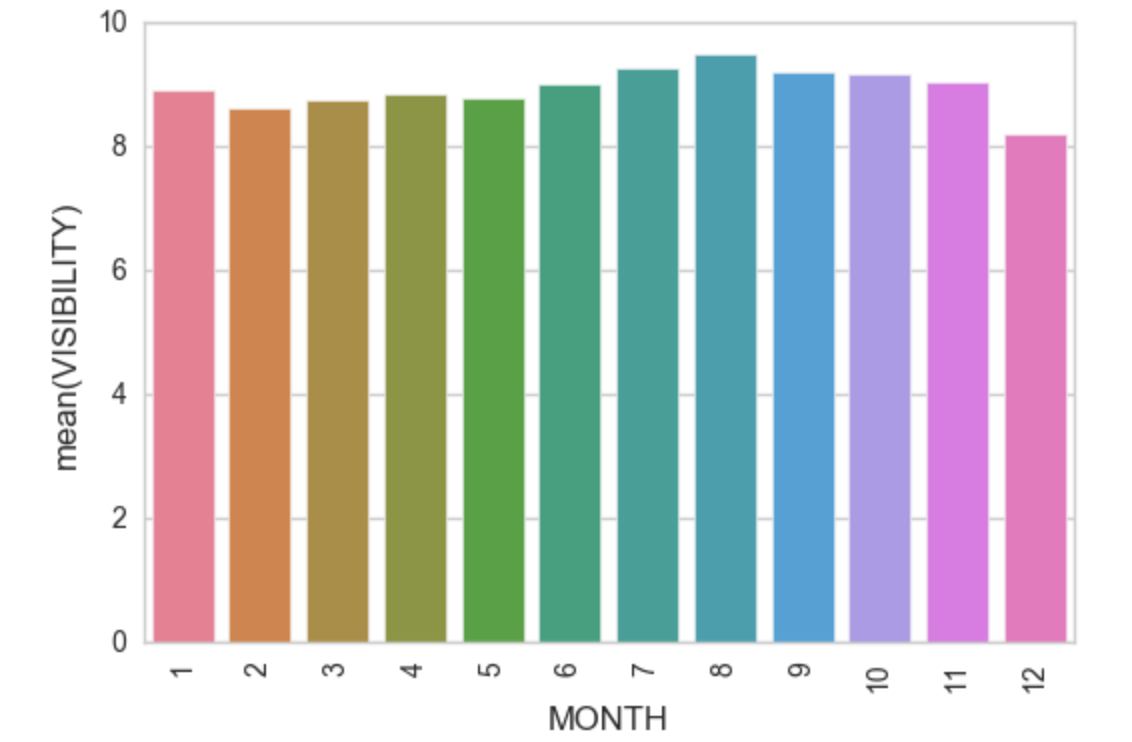
This graph groups the hourly dry bulb temperature by day, and presents the average temperature fluctuates with date.

* **Monthly temperature**

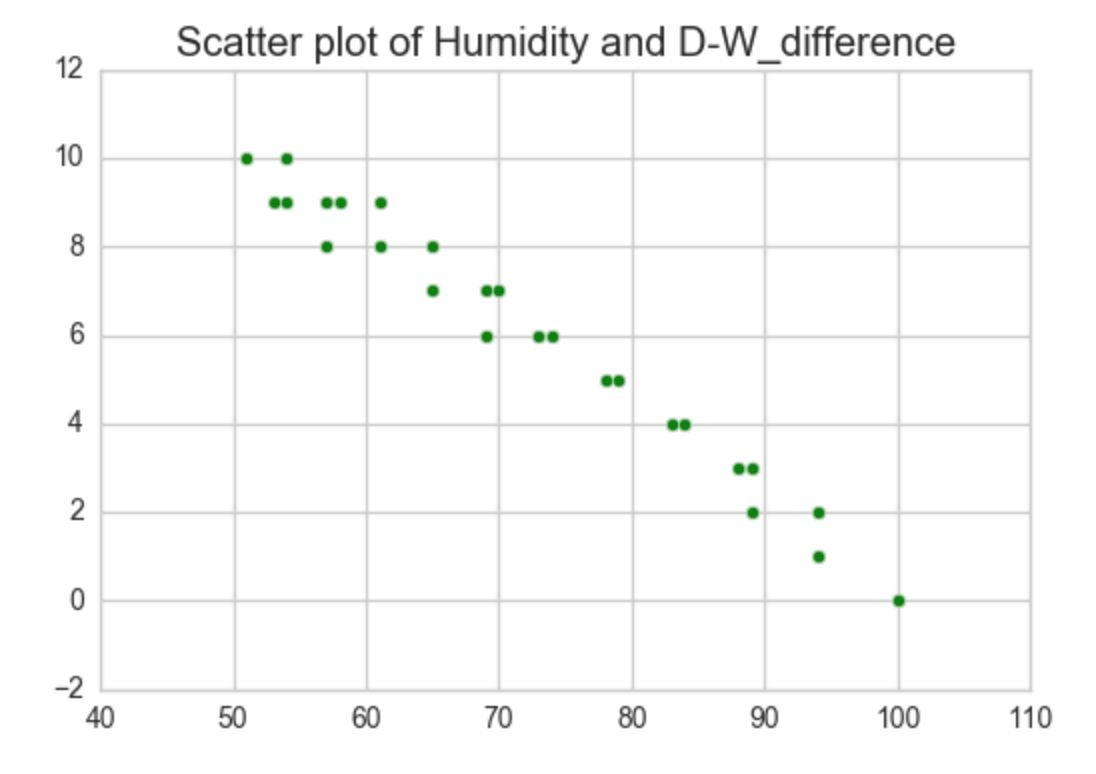


In this graph, we can get the max, min, and median temperature of every month, and compare the monthly temperature.

* **Monthly mean visibility**



* **Relationship among DryBulbTemp, WetBulbTemp, and Humidity**



we pick up the temperature data between 70F to 80F and humidity data, and find that the difference between drybulbtemp and wetbulbtemp is inversely proportional to humidity.