Knowledge grows

Python Code Challenge

Role: Data Science Intern (Digital Productions)

Please implement a data cleaning class in python. This class should have the following methods:

1) read_data(file_name)
    a. reads a csv file and stores the content (for example as pandas dataframe)
    b. implement appropriate exception handling (for example, if the file does not exist, or it's not a csv)
2) delete_outliers(x_percent)
    a. replace outliers with None
    b. remove for example the highest and lowest x percent of data
3) impute_missing_values(method)
    a. replace None values with appropriate given method
4) calculate_statistics()
    a. print appropriate statistics that describe the data
5) read_baseline(baseline)
    a. baseline should be a dict (for example: {'col_xyz': {'average': 5, 'min': 2, 'max': 6, 'std': 0.2}, 'col_abc': …} that describes per column statistics
    b. the statistics should be stored as class arguments
6) compare_to_baseline()
    a. function that compares the data with the baselines and prints a warning if the statistics of the data do not agree with the baseline
    b. inform the user if the data is similar to the baseline

Note: The focus of the challenge is to write clean and readable python code. Please feel free to add more methods that could be relevant for data cleaning. You can also add unit tests. In the interview, you will have to present the python code you wrote. Preparing the presentation is also part of the challenge.

Please use the attached data (csv file) to test your data cleaning class. It is hourly data provided by one of our chemical plants. The column names map to:

- 01FI1101E/PV.CV – consumption of ammonia in tons/hour
- 01FI1103/AI1/PV.CV\t– production of steam in tons/hour
- 01AI1923/AI1/PV.CV\t– ambient (outside) humidity in %
- 60PI0496/AI1/PV.CV\t– ambient (outside) pressure in mbara
- 00TI0538/AI1/PV.CV – ambient (outside) temperature in degrees Celcius
- 01HC1955/PID1/PV.CV – speed of compressor in rpms (revolutions per minute)

Please create your own baseline that makes sense in the context of the provided data.

Good Luck (:

YARA GmbH & Co. KG

Postal Address
Hanninghof 35
D-48249 Dülmen
Germany

Visiting Address
Hanninghof 35
D-48249 Dülmen
Germany

Telephone
+49 2594 798-0
Telefax
+49 2594 798-116

Registration No.
HRA3975
www.yara.de