IC 272: Lab3: Outlier detection, Standardization and Normalization of data

A dataset related to red variants of the Portuguese "Vinho Verde" wine is given. This dataset contains the values of different physicochemical tests from each samples of red wine [1]. The original goal of the dataset is to model wine quality based on physicochemical tests. The attributes of the dataset based on physicochemical tests are fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH value, sulphates, alcohol content and last attribute is on quality. Each expert graded the wine quality between 0 (very bad) and 10 (very excellent).

You are given with “winequality-red_original.csv” file. Write a Python program to do the following.

1. Read the data into a dataframe using pandas. Obtain the boxplot of all the attributes (exclude the attribute “quality”). Observe the number of outliers in each attributes and their values. Outliers are the values that do not satisfy the condition: **(Q1 - 1.5 * IQR) < X < (Q3 + 1.5 * IQR)** where, **IQR** is the **Interquartile range (= Q3-Q1), where Q1** and **Q3** are the lower and upper quartiles. Replace these outliers with the median of the attribute. Plot the boxplot again and observe the difference. Do you still get outliers? Why?
(You may use Q1=df.quantile(0.25) and Q3= df.quantile(0.75) in pandas)

2. Observe the range of the values in each attribute (Use the data obtained after outlier correction). Find the minimum and maximum values in each attribute.
i) Perform the Min-Max normalization of this data. (Make sure that you do not alter the attribute: “quality”)
ii) Perform Min-Max normalization to have the range of values between 0-20.

3. Use the data obtained after outlier correction. Find the mean and standard deviation of the attributes. Standardize each attribute (exclude “quality”) using the relation $Xnew=(X-\mu)/\sigma$ where μ is mean and σ is standard deviation. Compare the mean and standard deviations before and after the standardization.

4. Repeat steps 2 and 3 using **scikit-learn** instead of pandas. You can use the functions *StandardScaler* and *MinMaxScaler* in scikit-learn.
(Sample code can be found at:
https://python-data-science.readthedocs.io/en/latest/normalisation.html)