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

The file “pima_indians_diabetes_original.csv” is given to you. This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of this dataset is to predict whether a patient has diabetes based on diagnostic measurements. Write a python program (with pandas) to do the following.

1. Read the data into a dataframe using pandas. Obtain the boxplot for the attributes “BMI”,”pres” and “pedi”. 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?

2. Observe the range of the values in these 3 attributes (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.
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 these 3 attributes 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.