IC 272: Lab2: Data Cleaning – Handling Missing Values and Outliers

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 two csv files. The “winequality-red_miss.csv” is a file that contains some missing values. The “winequality-red_original.csv” is the original file without any missing values.

Make a copy of the file “winequality-red_miss.csv” as “winequality-red_miss-COPY.csv”. Write a python program (with pandas) to do the following using “winequality-red_miss-COPY.csv”. Missing values are interpreted as “NaN” in pandas.

1. Display the number of missing values in each attributes. Also find the total number of missing values in the file.
2. Delete any two integer values in attribute “*fixed acidity*” and replace any two integers values in attribute “*volatile acidity*” by “N/A”. Recalculate the number of missing values and observe the change.
3. Change any two integer values in attribute “*volatile acidity*” to “na”. Recalculate the number of missing values and observe the change. If your program could not detect “na” as missing value, make suitable changes in program to rectify it.

Now consider the file “winequality-red_miss.csv”. Write a python program (with pandas) to do the following in the dataset file “winequality-red_miss.csv”:

1. Count and display the number of tuples having one, two, three, four upto 12 missing value. Plot a graph for “number of missing values” (x-axis) vs “number of tuples” (y-axis).
2. Count and display the number of tuples having *equal to or more than* 50% of attributes with missing values.
3. (a). Delete (drop) the tuples having *equal to or more than* 50% of attributes with missing values.

(b). Target (class) attribute id "*quality*". Drop the tuple having missing value in the target (class) attribute.

4. Now, count and display the number of missing values in each attributes. Also find the total number of missing values in the file (after the deletion of tuples).
5. Experiments on filling missing values:
    a. Replace the missing values by median of their respective attribute. (Use `df.fillna()` with suitable arguments.)
        i. Compute the mean, median, mode and standard deviation for each attributes and compare with that of the original file "winequality-red_original.csv". Compare the box-plot for each attributes after filling the missing value with that of the original file.
        ii. Compare these replaced values with the actual values present in the original file. Calculate the root mean square error (RMSE) between the original and replaced values. (Get original values from "winequality-red_original.csv").
    b. Replace the missing values by propagating previous non-missing values in that attribute. (Use `df.fillna()` with suitable arguments.)
        i. Compute the mean, median, mode and standard deviation for each attributes and compare with that of the original file "winequality-red_original.csv". Compare the box-plot for each attributes after filling the missing value with that of the original file.
        ii. Compare these replaced values with the actual values present in the original file. Calculate the root mean square error (RMSE) between the original and replaced values. (Get original values from "winequality-red_original.csv").
    c. Replace the missing values in each attribute using linear interpolation technique. Use `df.interpolate()` with suitable arguments.
        i. Compute the mean, median, mode and standard deviation for each attributes and compare with that of the original file "winequality-red_original.csv". Compare the box-plot for each attributes after filling the missing value with that of the original file.
        ii. Compare these replaced values with the actual values present in the original file. Calculate the root mean square error (RMSE) between the original and replaced values. (Get original values from "winequality-red_original.csv").

Sample snippet is given below.

```
**********************************************
import pandas as pd
df = pd.DataFrame({"A":[12, 4, 5, None, 1],"B":[None, 2, 54, 3, None]})
 df=

      A     B
0  12.0   NaN
1   4.0   2.0
2   5.0  54.0
3   NaN   3.0
4   1.0   NaN

# to interpolate the missing values
df=df.interpolate(method ='linear', limit_direction ='forward')

df=
      A     B
0  12.0   NaN
1   4.0   2.0
2   5.0  54.0
3   3.0   3.0
4   1.0   3.0
```

Reference:

[1] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. "*Modeling wine preferences by data mining from physicochemical properties*," In Decision Support Systems, Elsevier, vol. 47, issue 4, pp. 547-553, 2009.