IC 272: Lab2: Data Cleaning – Handling Missing Values

You are given with two csv files. The “landslide_data2_miss.csv” is a file that contains some missing values. The “landslide_data2_original.csv” is the original file without any missing values. This dataset contains the readings from various sensors installed at 10 locations around Mandi district. These sensors give the details about the factors like temperature, humidity, pressure etc. Write a python program (with pandas) to do the following on the data file “landslide_data2_miss.csv”.

1. Count and display the number of tuples having one, two, three, four upto 8 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 “*stationid*”. 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. 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 for each attribute. (Get original values from original file provided).

   b. 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. 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 for each attributes. Plot these RMSE with respect to the attributes.

6. **Visualizing the data.** Consider the attribute “temperature”. Obtain a histogram plot for this attribute for following cases.
    a. Histogram after filling the missing values by “0” (df.fillna(0)).
    b. Filling the missing values by median.
    c. Filling the missing values by interpolation.
    d. Plot the variation in this attribute for a selected “stationid” with respect to time index (Day). Note that you have data from multiple “stationid”.