IC 272: Lab2: Data Cleaning – Handling Missing Values and Outlier Analyses

Deadline for submission: Sunday, 5 September 2021, 10:00 PM

You are given with two csv files. The "`landslide_data3_miss.csv`" is a file that contains some missing values. The "`landslide_data3_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. Following are the details of the attributes in the data:

- **dates:** date of collection of data.
- **stationid:** Indicates the location of the sensor.
- **temperature:** Atmospheric temperature around the sensor in Celsius.
- **humidity:** The concentration of water vapor present in the air (in $g.m^{-3}$).
- **pressure**: Atmospheric pressure in millibars (mb).
- **rain:** Measure of rainfall in ml.
- **lightavgw/o0:** The average light throughout the daytime (in lux units).
- **lightmax:** The maximum lux count by the sensor.
- **moisture:** indicates the water stored in the soil (measured between 0 to 100 percent).

Write a Python program (with pandas) to do the following on the data file "`landslide_data2_miss.csv`".

1. Plot a graph of the attribute names (x-axis) with the number of missing values in them (y-axis).

2. (a). Target attribute is "**stationid**", Drop the tuples (rows) having missing value in the target attribute. Print the total number of tuples deleted.

   (b). Delete (drop) the tuples (rows) having *equal to or more than* one third of attributes with missing values. Print the total number of tuples deleted.

3. After step 2, count and print the number of missing values in each attributes. Also find and print the total number of missing values in the file (after the deletion of tuples).

4. Experiments on filling missing values:
   a. Replace the missing values by mean of their respective attribute. (Use `df.fillna()` with suitable arguments.)
      i. Compute the mean, median, mode and standard deviation for each attributes and compare the same with that of the original file.
      ii. Calculate the root mean square error (RMSE) between the original and replaced values for each attribute. (Get original values from original file provided). Compute RMSE using the equation (1). Plot these RMSE with respect to the attributes.

   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.

ii. Calculate the root mean square error (RMSE) between the original and replaced values for each attributes. (Get original values from original file provided). Compute RMSE using the equation (1). Plot these RMSE with respect to the attributes.

Note: RMSE is computed between the replaced value and its corresponding original value. You are computing RSME for each attribute. Let Na be the number of missing values in attribute 'a'. Let $\hat{x}_i$ be the replaced value and x_i be the original value of i^{th} missing value. Then the RMSE for attribute 'a' is computed as:

$$RMSE = \sqrt{\frac{1}{Na}\sum_{i=1}^{Na}(\hat{x}_i - x_i)^2} \qquad \text{----- (1)}$$

5. **Outlier detection:**
   a. After replacing the missing values by interpolation method, find and list the outliers in the attributes "**temperature**" and "**rain**". Outliers are the values that does not satisfy the condition (Q1 – (1.5 * IQR)) < x < (Q3 + (1.5 * IQR)), where x is the value of the attribute, IQR is the inter quartile range, Q1 and Q3 are the first and third quartiles. Obtain the boxplot for these attributes.
   b. Replace these outliers by the median of the attribute. Plot the boxplot again and observe the difference with that of the boxplot in (5i). Do you still get outliers? Why?

Instructions:

- **Your python program(s) should be well commented. Comment section at the beginning of the program(s) should include your name, registration number and mobile number.**
- **The python program(s) should be in the file extension .py**
- **Report should be strictly in PDF form. Write the report in word or latex form and then convert to PDF form.**
- **First page of your report must include your name, registration number and mobile number.**
- **Upload your program(s) and report in a single zip file. Give the name as <roll_*number*>_Assignment2.zip. Example: *b20001*_Assignment2.zip**
- **Upload the zip file in the link corresponding to your group only.**

In case the program found to be copied from others, both the person who copied and who help for copying will get zero as a penalty.