Data Cleaning – Handling Missing Values and Outlier Analyses

You are given with two csv files. The "`pima_indians_diabetes_miss.csv`" is a file that contains some missing values. The "`pima_indians_diabetes_original.csv`" is the original file without any missing values. This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases, USA. The objective of this dataset is to predict whether a patient has diabetes based on diagnostic measurements. Following are the details of the attributes in the data:

- `pregs` (Pregnancies) : Number of Pregnancies
- `plas` (Glucose) : Plasma glucose concentration a 2 hours in an oral glucose tolerance test.
- `pres` (Blood Pressure) : Diastolic blood pressure (mm Hg).
- `skin` (Skin Thickness) : Triceps skin fold thickness (mm).
- `test` (Insulin) : 2-Hour serum insulin (mu U/ml).
- `BMI` (BMI) : Body mass index (weight in kg/(height in m)^2).
- `pedi` (Diabetes pedigree function) : Diabetes pedigree function.
- `Age` (Age) : Age of patients (years).
- `Class` (Outcome) : Class variable (0 or 1). 0 indicates not presence of diabetics and 1 indicates presence of diabetics. 268 patients out of 768 are 1, the others are 0.

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

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

2. (a). 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 and also print the row numbers of the deleted tuples (with respect to `pima_indians_diabetes_miss.csv`).
   (b). Drop the tuples (rows) having missing value in the target (class) attribute. Print the total number of tuples deleted and also print the row numbers of the deleted tuples (with respect to `pima_indians_diabetes_miss.csv`).

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 xx_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:

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

5. **Outlier detection:**
    i. After replacing the missing values by interpolation method, find the outliers in the attributes "`Age`" and "`BMI`". 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.
    ii. 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 or**
- **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.