Exercise 4: Data Processing and Imputation

Consider the diabetes dataset shared in our Blackboard folder for *Exercise 4*. (Exercise 4 is related with lecture topic missing data and imputation). The file is called `diabetes.csv`. Download this file and move it next to your notebook file that you are executing - i.e. next to this file. The data is famous in machine learning and data analysis. It contains medical data of women native to the Phoenix, Arizona area, collected between 1965 and 1975.

The dataset includes information about `768` female patients, where each patient is described by `8` different medical attributes:

1. *Pregnancies*: the number of times the patient has been pregnant.
2. *Glucose*: the patient's plasma glucose concentration, measured in milligrams per deciliter (mg/dl).
3. *Blood Pressure:* the patient's diastolic blood pressure, measured in millimeters of mercury (mm Hg).
4. *Skin Thickness:* the thickness of the patient's triceps skinfold, measured in millimeters (mm).
5. *Insulin:* the patient's insulin level, measured in milli-international units per milliliter (mu U/ml).
6. *BMI:* the patient's body mass index, calculated as weight in kilograms divided by the square of height in meters (kg/m^2).
7. *Diabetes Pedigree Function:* a score that represents how likely the patient is to develop diabetes based on family history.
8. *Age:* the patient's age in years.
9. The class label 0 or 1 indicating whether each patient developed diabetes within 5 years of the measurement (1) or not (0).

Here, an example code is provided to get started your own code for the exercise 4. In code you will be able to import all relevant libraries and load the given dataset and learn something about it.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn import linear_model
from sklearn import datasets

df = pd.read_csv("diabetes.csv")
df.head()

# A simple df.info() is ran for a quick and abstract check for
missing data in any of the variables. This lists the number of non-
null values and the datatype of each variable.

df.info()
```

Task 1. Data Exploration and Identification of Anomalies:

1.1 Examine Attribute Distributions

- Perform descriptive statistics `(.describe())` and visualize the distributions of `Plasma Glucose Concentration, Diastolic Blood Pressure, Triceps Skinfold Thickness, 2-Hour Serum Insulin, and Body Mass Index.`
- Identify potential issues in the dataset (e.g., incorrect values, missing values, extreme outliers, or unexpected distributions).

1.2 Identify Missing and Anomalous Values

- Investigate whether any values appear as missing (e.g., zeros where they should not be possible).
- Implement an appropriate method to detect missing values and anomalies programmatically.

1.3 Suggest and Implement a Strategy for Handling Anomalies

- Propose and implement an appropriate method to handle anomalies (e.g., replacing zeros, removing extreme outliers, transforming the data).
- Justify why your chosen method is appropriate for this dataset.

Task 2. Consider If multiple variables in the dataset contain missing values, can Regression Imputation be applied directly to impute one of them, given that the predictor variables also have missing data? If **not**, explain why and provide a programming solution using the given dataset. If **yes**, justify your reasoning with supporting evidence.

Task 3. Building and Evaluating a Regression Imputation Model

After completing Task 2, proceed with implementing a **linear regression model** to impute the missing values.

3.1 How can you use the model to predict values?

3.2 Show difference between randomly imputed values and the values provided by the regression imputation.

3.3 Use some visualisation methods and show how good you imputed the missing values. Whether you negative impacted the distribution? [Hint: compare different variables distributions in the original data and in the imputed data]

For Reading: Notes on Regression Imputation

Regression imputation is a method for imputing missing data by using a regression model to predict the missing values based on the available data for other variables. This method can be useful when there are strong correlations between the missing variable and other variables in the dataset. Regression imputation assumes that the missing values are missing at random (MAR) or missing completely at random (MCAR). If the missing values are not MAR or MCAR, regression imputation may not be appropriate and other imputation techniques may need to be used. Additionally, regression imputation assumes a linear relationship between the variables, so it may not work well if the relationship is nonlinear or if there are interactions between variables.

You would apply regression imputation when you have a dataset with missing values and want to estimate the missing values based on the available data. This method can be especially useful when the missing values are related to other variables in the dataset, as it allows you to leverage the information from those variables to make more accurate predictions.