Hyperiondev

TASK

Data Analysis – Preprocessing

Model-Answer Approach

Visit our website

Auto-graded Task 1

In this approach, we start by reading a CSV file using Pandas' **read_csv()** function, which returns a DataFrame – a two-dimensional labelled data structure with rows and columns. We then utilise the **head()** method to retrieve the first five observations from the DataFrame. By default, the head() method returns the first five observations. Subsequently, we identify missing values in each column using the **isnull()** function, followed by the **sum()** function to count the number of missing values in each column of the DataFrame. Finally, we classify the given scenarios according to the categories or types of missingness of data, which include *Missing Not At Random*, *Missing At Random*, *Missing Not at Random*, and *Missing Completely at Random*.

Auto-graded Task 2

In the first set of two scenarios, we determine whether standardisation or normalisation makes more sense.

Normalisation is chosen for the first scenario because the data features vary in scales. Normalisation scales the features to a range between zero and one, while preserving the relative differences. For the other scenario, standardisation works well, as the data assumes a Gaussian distribution. Standardisation transforms the features to have a mean of zero and a standard deviation of one. We then employ the **minmax()** function to normalise the "EG.ELC.ACCS.ZS" feature from the "countries" DataFrame. Subsequently, Seaborn's **histplot()** is utilised to plot a histogram showing the distribution of our column before and after the normalisation process. Following normalisation, the scales are observed to be between zero and one, indicating the success of the normalisation process.