Ebola data cleaning and preparation

This exercise aims to clean and prepare a dataset using Python.

After having uploaded the dataset in Python, I explored the data through the commands head(), columns(), and info(), describe(), and isnull().sum to understand the dataset contents, identify the essential columns, and check for missing values.

Then I started the cleaning process, removing the rows without essential information in the columns EbolaResult_cn, DOAdm_cn, and Age_cn to obtain a cleaned dataset with complete data. I created a new dataset called ebola_cleaned.

Then I moved on to patients' symptom columns. These columns contain Boolean values (True and False). In some rows, there were missing values (NaN), which I decided to substitute with False, assuming that the absence of data implies the absence of symptoms.

For the numeric columns, which are: Temp_cn, Systolic_cn, Diastolic_cn, Hrate_cn, Rrate_cn, Age_cn, Weight_cn, and Height_cn, I converted them into numeric types to transform invalid values into NaN. Then I replaced the NaN values with the column's median. I chose the median because it is not influenced by outliers, and I can maintain the data distribution without deleting too many rows.

After cleaning, I checked the missing values to ensure that all principal columns were complete.