Task3.1

for (a) education data set:

Combination of attributes with the most missing values: ('occupation', 'salary', 'credit\_card\_number')

Number of records with missing values for this combination: 254

for (b) merged data set.

Combination : ('marital\_status', 'occupation', 'credit\_card\_number')

Number : 1430

Task3 .2

Top 2 attributes with the highest number of missing values:

salary 2510

marital\_status 2240

For those attributes above in the merged dataset,I used the following imputation strategies based on correlation analysis:

1. Salary:

- Imputation Approach: Before imputing, I calculated the correlation between salary and other features (e.g., occupation, education) using Pearson correlation for numerical features and Label Encoding for categorical features. I selected features that showed strong positive or negative correlations with salary. I then used these features to build a Multiple Linear Regression model to predict missing salary values.

- Justification: Salary, being a continuous variable, is well-suited for regression-based imputation. The features I selected, such as occupation, had a meaningful relationship with salary. By using these in the regression model, I ensured that the imputed values reflect the broader trends and dependencies present in the dataset, leading to more accurate predictions.

2. Marital Status:

Imputation Approach: I calculated correlations between marital\_status and numerical features using Pearson correlation and evaluated associations with categorical features using Cramér’s V. The most relevant features (e.g., age, gender) were used to build a Multinomial Logistic Regression model for imputation.

Justification: As a categorical variable, marital\_status was imputed using logistic regression to predict category probabilities. This approach ensured realistic, data-driven imputation based on observed relationships.

Task3.3

Counts of Incorrect or Impossible Values per Attribute:

weight: 1648 incorrect or impossible values

blood\_pressure: 16815

cholesterol\_level: 1326

birth\_date: 10194

email: 1822

credit\_card\_number: 13815

salary: 2707

postcode: 16810

Judgment standard：each attribute should be based on logical constraints and domain-specific validation rules. for example

Weight:

Flags weight values below 2 kg or above 635 kg, which are not plausible.

Birth Date:

Ensures that birth\_date is not in the future and is consistent with the calculated age.

Credit Card Number:

Uses the Luhn algorithm to verify credit card numbers for validity.

Task3.4

for the incorrect or impossible values,i mainly take 3 steps:Standardization,Validation and Correction. for example:

postcode attribute: I identified issues with invalid formats not matching the expected pattern (e.g., four digits for Australian postcodes).

Actions Taken:

Standardization: Trimmed leading/trailing spaces and converted to uppercase if letters were involved.

Validation and Correction: Checked the format to ensure postcodes matched the four-digit pattern. Where errors like missing or transposed digits were found, I corrected them using available data such as suburb and state. For postcodes that could not be corrected, I marked them as NaN for further handling or exclusion.