For the two attributes with the highest number of missing values in the merged dataset—salary and marital\_status—I applied the following imputation strategies, guided by correlation analysis to select relevant attributes:

1. Salary:

- Imputation Approach: Before imputing, I calculated the correlation between salary and other features (e.g., `occupation`, `education`, `years\_of\_experience`, and `current\_age`) 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 `years\_of\_experience` and `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 other numerical features using Pearson correlation and evaluated associations with categorical features using Cramér's V. I selected the most relevant features (e.g., `age`, `gender`, `education`, and `occupation`) and used them to build a Multinomial Logistic Regression model for imputation.

- Justification: Marital status is a categorical variable with multiple levels, making logistic regression appropriate for predicting probabilities of categories such as single, married, or divorced. The correlation analysis ensured that the most predictive attributes were included, allowing for realistic, data-driven imputation that aligns with observed patterns in the dataset.