Task 4: Additional Data Exploration and Cleaning

To ensure the quality of the cleaned dataset, aimed at exploring the relationships between education, employment, and health, I conducted several additional data exploration and cleaning tasks. These tasks address key data quality dimensions and help prepare the dataset for analysis.

1. Uniqueness and Duplicate Records:

- Issue: Duplicate records may exist in the merged dataset, leading to skewed results in the analysis.

- Action: I defined duplicates based on the same SSN and other key attributes. Using pandas’ `drop\_duplicates()`, I removed these duplicates, ensuring that only the most complete or recent record for each individual was retained.

- Justification: Removing duplicates ensures that each individual is uniquely represented, avoiding over-representation of specific individuals, which could bias the results.

2. Data Transformation for Categorical Attributes:

- Issue: Continuous BMI values need to be categorized to better analyze health patterns.

- Action: I created derived variables by converting the continuous BMI values into categorical ranges (e.g., `Underweight`, `Normal weight`, `Overweight`, `Obese`), following the WHO classification.

- Justification: This transformation allows for a clearer analysis of health trends by grouping individuals into meaningful BMI categories. It makes it easier to correlate health outcomes with education and employment factors.

3. Outlier Detection and Treatment:

- Issue: Some salary values were negative (e.g., -9999, -1280), which is unrealistic.

- Action: I identified these erroneous salary values and imputed them using a predictive model based on relevant attributes such as `education`, `occupation`, and `years\_of\_experience`.

- Justification: Imputing salary values based on predictors ensures the integrity of the data, maintaining realistic values and improving the dataset's accuracy for salary-related analysis.

4. Other Problems Worth Correcting:

- Email Validation: After merging the two datasets, some emails were in an invalid format (e.g., `123.gmail.com` instead of `123@gmail.com`). I noticed that when using the `combine\_first()` method, incorrect email formats were retained over valid ones. Therefore, I validated the email addresses before merging, ensuring that the correct email (i.e., the one containing an "@" symbol) was kept.

- Incorrect Weight Values: I found some weight values recorded as `-99`, which is clearly wrong. To address this, I calculated the correct weight using the available BMI and height values. This imputation ensured the weight values were consistent with the health metrics in the dataset.