Data Generation

Chat-GPT Prompts

Generate python script to generate a realistic dataset of 8950 records for human resources. The dataset should include the following attributes:

1. Employee ID: A unique identifier.
2. First Name: Randomly generated.
3. Last Name: Randomly generated.
4. Gender: Randomly chosen with a 46% probability for ‘Female’ and a 54% probability for ‘Male’.
5. State and City: Randomly assigned from a predefined list of states and their cities.
6. 6.Hire Date: Randomly generated with custom probabilities for each year from 2015 to 2024.
7. 7.Department: Randomly chosen from a list of departments with specified probabilities.
8. Job Title: Randomly selected based on the department, with specific probabilities for each job title within the department.
9. Education Level: Determined based on the job title, chosen from a predefined mapping of job titles to education levels.
10. Performance Rating: Randomly selected from ‘Excellent’, ‘Good’, ‘Satisfactory’, ‘Needs Improvement’ with specified probabilities.
11. Overtime: Randomly chosen with a 30% probability for ‘Yes’ and a 70% probability for ‘No’.
12. Salary: Generated based on the department and job title, within specific ranges.
13. Birth Date: Generated based on age group distribution and job title requirements, ensuring consistency with the hire date.
14. Termination Date: Assigned to a subset of employees (11.2% of the total) with specific probabilities for each year from 2015 to 2024, ensuring the termination date is at least 6 months after the hire date.
15. Adjusted Salary: Calculated based on gender, education level, and age, applying specific multipliers and increments.
16. Be sure to structure the code cleanly, using functions where appropriate, and include comments to explain each step of the process.