Task 2 Data Anonymization

Steps followed for Task-2

Data Anonymization and Privacy Techniques Applied

1. **Removal of Non-Essential Data:**
    - The customer_id and current_location columns were removed to ensure data privacy, as they provided minimal informational value.
2. **Username Anonymization:**
    - The username column was anonymized while retaining only the first and last letters. The middle characters were replaced with asterisks (*), using the following formula:

        =LEFT(F2,1) & "*****" & RIGHT(F2,1)
3. **Name Replacement with Fake Data:**
    - The name column was replaced with randomly generated fake names to protect user identities. This was achieved using the RANDBETWEEN and INDEX functions:

        =INDEX($I$2:$I$9424, RANDBETWEEN(1, COUNTA($I$2:$I$9424)))
4. **Email Address Masking:**
    - The email column was partially masked to conceal the actual address while retaining the first letter and domain:

        =LEFT(L2,1) & "****" & MID(L2, FIND("@",L2)-1, LEN(L2) - FIND("@",L2) + 2)
5. **Date Noise Addition:**
    - Random noise was introduced to date_registered and birthdate to obscure exact values while preserving the overall pattern:

        =TEXT(DATE(YEAR(C2)+RANDBETWEEN(-2,2), MONTH(C2), DAY(C2)), "YYYY-MM-DD")
6. **Salary and Age Binning:**
    - The salary and age columns were categorized into predefined bins using VLOOKUP for anonymization while maintaining distribution:

        =VLOOKUP(G2, Salary_Age_Bins, 2, TRUE)
    - Here, Salary_Age_Bins is a lookup table mapping original values to categorized ranges.
7. **Credit Card Provider and Expiry Tokenization:**
    - The credit_card_provider and credit_card_expire columns were tokenized to replace real values with random but structured alternatives. A pivot table was used to create mappings, and VLOOKUP was applied:

=VLOOKUP(H2, Provider_Token_Mapping, 2, FALSE)

8. **Credit Card Number Masking:**
    - The credit_card_number column was fully masked, displaying only the last four digits for partial visibility:

        =REPT("*", LEN(J2)-4) & RIGHT(J2,4)

9. **Credit Card Security Code Masking:**
    - The credit_card_security_code column was masked for security purposes by replacing digits with asterisks:

        ="***"

10. **Employer and Job Tokenization:**

- The employer and job columns were tokenized to maintain original distribution while hiding actual details:

    =VLOOKUP(K2, Employer_Job_Token, 2, FALSE)

11. **Residence and Address Replacement:**

- The residence and address columns were replaced with randomly generated fake values, similar to the name replacement approach:
- =INDEX($I$2:$I$9424, RANDBETWEEN(1, COUNTA($I$2:$I$9424)))