Model answer explanation

The model answer is one way to approach this problem and not the only solution. To create the model answer, we built the following steps into a data privacy pipeline:

- Remove the **customer_id** and **current_location** column. Removal of data (redaction) that does not have much informational value is a valid data privacy technique.
- Mask the **username** column to hide the real username.
- Replace the original **name** column with a fake name. Replacing real values for fake values is a valid data privacy technique.
- Mask the **email** column to hide the real email address.
- Add noise to the **date_registered** and **birthdate** columns to hide the real value. Adding noise protects the real value by adding random noise to the actual value.
- Categorise the **salary** and **age** columns into bins. This categorisation hides the original values and preserves the distribution.
- The **credit_card_provider** and **credit_card_expire** have been tokenised. This step converts the categorical value of the columns into a different random value while preserving the distribution.
- The **credit_card_number** and **credit_card_security_code** have been masked.
- The **employer** and **job** columns have also been tokenised to preserve the original distribution.
- **Residence** and **address** have been replaced with fake values.