ask2

2.1

Number of unique SSNs occurred in common in both datasets: 16005

Data only in the medical dataset: 3995

Data only in the employment dataset: 3185

2.2

I performed an outer join to include all records from both the medical and employment datasets, ensuring no data was lost. Records unique to one dataset were included with `NaN` values for fields from the other dataset.

Justification:

1. Data Completeness: Preserves all records for future analysis, even if they exist in only one dataset.

2. Real-World Scenarios: Reflects real situations where individuals may have medical or employment events, but not both.

3. Avoiding Bias: Ensures balanced representation by including individuals with only one type of event.

Task2.3

Number of unique SSNs with duplicate records in the medical dataset: 0

Those records in the employment dataset: 810

To handle duplicate records, I first identified all duplicate SSNs in both the medical and employment datasets. For deduplication, I consolidated these records by applying different strategies based on the type of data:

- Numerical attributes (e.g., BMI, salary) were handled by selecting the most recent value based on timestamps or averaging the values if appropriate.

- Categorical attributes (e.g., gender, education) were resolved by selecting the most frequent or consistent non-null entry.

- Textual attributes (e.g., clinical notes) were combined by concatenating all notes from duplicate records.

When conflicts arose, I prioritized the most accurate or complete records. After resolving conflicts, I merged duplicates into a single representative record for each SSN.

This approach ensured data integrity by preventing skewed analysis and redundancy, while retaining valuable information about each individual.

Task2.4

Inconsistency counts per attribute:

- first\_name: 0

- middle\_name: 2801

- last\_name: 0

- gender: 1631

- birth\_date: 0

- street\_address: 6597

- suburb: 6490

- postcode: 8358

- state: 2677

- phone: 8565

- email: 6878

To handle inconsistencies between the two datasets, I standardized text attributes (e.g., names, email, address) by converting them to lowercase and removing extra spaces. For names, I used a similarity threshold via the SequenceMatcher library, treating values as consistent if the similarity score was above 0.8. For gender, I mapped variations like "M", "F", "male", and "female" to standard values. Birth dates were handled with custom parsing to address potential formatting issues, such as "24:00" hours. Phone numbers and addresses were compared exactly. Inconsistencies were flagged, and counts were tracked for review.