Assignment

Disclaimer: Please be aware that the data used in this assignment is entirely fictitious and does not correspond to any real individuals or actual data.

A. Objective: Create a single CSV file containing the following 5 fields:

1. `patient_id` (Primary Key)
2. `enrollment_start_date` (Primary Key)
3. `enrollment_end_date` (Primary Key)
4. `ct_outpatient_visits`
5. `ct_days_with_outpatient_visit`

Each row in the file should represent a unique combination of `patient_id` and `enrollment_start_date` and `enrollment_end_date`. Field definitions are listed in Section C.

B. Prompt: To achieve the objective, follow the steps below and provide the required information in your answer sheet.

Step 1: Data Transformation

1. Access the `patient_id_month_year.csv` file stored in a public S3 bucket. You will need to use an API call to retrieve this file using these credentials:
    a. `aws_access_key_id='AKIAZLXG4RYJBLE4OTXT'`
    b. `aws_secret_access_key='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'`
2. This CSV file contains two fields: `patient_id` and `month_year`. Transform this dataset from `patient_id x month_year` level to `patient_id x enrollment_start_date x enrollment_end_date` level using Python.
3. Save the result as `patient_enrollment_span.csv`.

Answer 1: Report the number of rows in `patient_enrollment_span.csv`.

Step 2: Data Aggregation

1. Access the `outpatient_visits_file.csv` file stored in our public S3 bucket using an API call. This file includes three fields: `patient_id, date, and outpatient_visit_count`.
2. Using `patient_enrollment_span.csv` and `outpatient_visits_file.csv`, create a single CSV file with the 5 fields mentioned in the objective. Implement this using Python.
3. Save the result as `result.csv`.

Answer 2: Report the number of distinct values of `ct_days_with_outpatient_visit` in `result.csv`.

C. Variable Definitions:

Variable	Tables	Definition
patient_id	patient_id_month_year.csv outpatient_visits_file.csv patient_enrollment_span.csv result.csv	A unique identifier for each patient.
month_year	patient_id_month_year.csv	The month and year the patient was enrolled in a health program. For example, if patient_id = ID0001and month_year = 2023-04-01, then patient ID001 was enrolled in the health program for the entire month of April, e.g. from 2023-04-01 through 2023-04-30.
enrollment_start_date	patient_enrollment_span.csv result.csv	This date marks the beginning of a continuous period during which the patient was enrolled in the health program. It should be in a standard date format (e.g., YYYY-MM-DD). The date is inclusive, meaning that this date is counted as part of the enrollment period.
enrollment_end_date	patient_enrollment_span.csv result.csv	This date signifies the end of the same continuous enrollment period. It is also in the YYYY-MM-DD format and is inclusive, with this date being part of the enrollment period.
outpatient_visit_count	outpatient_visits_file.csv	The number of outpatient visits the patient had on a given date.
ct_outpatient_visits	result.csv	The number of outpatient visits a patient had within the enrollment period (between enrollment_start_date and enrollment_end_date).

ct_days_with_outpatient_visit	result.csv	The number of distinct days within an enrollment period (between enrollment_start_date and enrollment_end_date) when the patient had one or more outpatient visit.

D. Answer Sheet

Please submit your responses and code in a word document as follows.

Name:

Answers:

1. [Put Answer 1 here]
2. [Put Answer 2 here]

Code:

Please ensure that your scripts are well-commented to explain your logic.

```
[Paste your code for Steps 1 and 2 here or include the link to the
github repo]
```
