

BI Project

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 :

Step 1: Data Transformation

1. Access the `patient_id_month_year.csv` file:
<https://docs.google.com/spreadsheets/d/1UuYfyMDLckI-X7mbKS-DyXfl6PwV9FLgwwO-mcvNI7s/edit?gid=0#gid=0>
 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`.
- ➔ Report the number of rows in `patient_enrollment_span.csv`.

Step 2: Data Aggregation

1. Access the `outpatient_visits_file.csv` file:
<https://docs.google.com/spreadsheets/d/17fjOu72wVcfyqtkwZBcmi8ITglqj22HHL6TMtR763tY/edit?gid=0#gid=0>

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`.

➔ Report the number of distinct values of `ct_days_with_outpatient_visit` in `result.csv`.

C. Variable Definitions:

| Variable | Tables | Definition |
|-------------------------------------|--|---|
| <code>patient_id</code> | <code>patient_id_month_year.csv</code> <code>outpatient_visits_file.csv</code> <code>patient_enrollment_span.csv</code> <code>result.csv</code> | A unique identifier for each patient. |
| <code>month_year</code> | <code>patient_id_month_year.csv</code> | The month and year the patient was enrolled in a health program. For example, if <code>patient_id</code> = ID0001 and <code>month_year</code> = 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. |
| <code>enrollment_start_date</code> | <code>patient_enrollment_span.csv</code> <code>result.csv</code> | 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. |
| <code>enrollment_end_date</code> | <code>patient_enrollment_span.csv</code> <code>result.csv</code> | 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. |
| <code>outpatient_visit_count</code> | <code>outpatient_visits_file.csv</code> | The number of outpatient visits the patient had on a given date. |
| <code>ct_outpatient_visits</code> | <code>result.csv</code> | The number of outpatient visits a patient had within the enrollment period (between <code>enrollment_start_date</code> and <code>enrollment_end_date</code>). |

| | | |
|-------------------------------|------------|--|
| 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. |
|-------------------------------|------------|--|

