Name: Rishika A

**Answers**:

1. The number of rows in patient\_enrollment\_span.csv as Answer 1: 3105
2. The number of distinct values of ct\_days\_with\_outpatient\_visit in result.csv: 33

**Code:**

**step1\_transform.py:**

import pandas as pd

from datetime import timedelta

from dateutil.relativedelta import relativedelta

# Load the monthly enrollment data CSV file, then convert the month\_year column into actual datetime objects

df = pd.read\_csv("patient\_id\_month\_year.csv", parse\_dates=["month\_year"])

#sort data

df = df.sort\_values(['patient\_id','month\_year'])

#Generate spans, which holds each patient's continuous enrollment Interval

enrollment\_spans = []

#Process one patient at a time

for patient\_id, group in df.groupby('patient\_id'):

group = group.sort\_values('month\_year').reset\_index(drop=True) # Make sure the data is sorter and indexed from 0

start\_date = group.iloc[0]['month\_year'] # start the span with first month

end\_date = start\_date

#Loop through rest of the months for this patient

for i in range(1, len(group)):

current\_date = group.iloc[i]['month\_year']

#If the month is less than or equal to 31 days after the last month, continue the span

if (current\_date - end\_date).days <= 31:

end\_date = current\_date

else:

#if there's a gap, end the current span and start a new one

span\_end = end\_date + relativedelta(months=1) - timedelta(days=1)

enrollment\_spans.append([patient\_id, start\_date, span\_end])

#start a new span

start\_date = current\_date

end\_date = current\_date

#After finishing all months for this patient, Add the final span

span\_end = end\_date + relativedelta(months=1) - timedelta(days=1)

enrollment\_spans.append([patient\_id, start\_date, span\_end])

#Convert the list of spans into a dataframe

result\_df = pd.DataFrame(enrollment\_spans, columns=["patient\_id", "enrollment\_start\_date", "enrollment\_end\_date"])

#Save the result as CSV

result\_df.to\_csv("patient\_enrollment\_span.csv", index = False)

print("Answer 1:", len(result\_df))

#Prints the number of rows in patient\_enrollment\_span.csv as Answer 1: 3105

**Step2\_aggregate.py**

import pandas as pd

#Load generated Enrollment\_span.csv file from step1 transformation and parse the data

enrollment = pd.read\_csv('patient\_enrollment\_span.csv', parse\_dates=['enrollment\_start\_date','enrollment\_end\_date'])

#Load outpatient visit data

no\_of\_visits = pd.read\_csv("outpatient\_visits\_file.csv", parse\_dates=['date'])

final\_data = [] #list to accumulate final results

# Iterate over each enrollment span

for \_, row in enrollment.iterrows():

patient\_id = row["patient\_id"]

start\_date = row["enrollment\_start\_date"]

end\_date = row["enrollment\_end\_date"]

# Filter the visits for this patient within the enrollment window

patient\_visits = no\_of\_visits[

(no\_of\_visits["patient\_id"] == patient\_id) &

(no\_of\_visits["date"] >= start\_date) &

(no\_of\_visits["date"] <= end\_date)

]

# Total visits and number of unique visit days

total\_visits = patient\_visits["outpatient\_visit\_count"].sum()

unique\_days = patient\_visits['date'].nunique()

#Append the result for this enrollment period

final\_data.append([

patient\_id,

start\_date.date(),

end\_date.date(),

total\_visits,

unique\_days

])

#Transform the results to a Dataframe

result\_df = pd.DataFrame(final\_data, columns=[

"patient\_id", "enrollment\_start\_date", "enrollment\_end\_date", "ct\_outpatient\_visits", "ct\_days\_with\_outpatient\_visit"

])

#Save the result as CSV

result\_df.to\_csv("result.csv", index='False')  
# Print the number of distinct values in days-with-visit column, here Answer: 33

print("Answer2:", result\_df["ct\_days\_with\_outpatient\_visit"].nunique())

**How to run Code**:

Step 1:

Load Input file: patient\_id\_month\_year.csv  
Save the above Script as step1\_transform.py  
Run the code as: python3 step1\_transform.py  
Output File: patient\_enrollment\_span.csv  
Answer: 3105

Step2:

Load CSV File from first step - patient\_enrollment\_span.csv and also load outpatient\_visits\_file.csv in your working folder  
Save the above script as step2\_aggregate.py  
Run as:   
python3 step2\_aggregate.py  
Answer: 33