Bajaj Health Programming Challenge
Data Engineering - Qualifier 1

Step 1:

You are given an attendance table that records students' attendance.

Table: attendance

student_id	attendance_date	status
101	2024-03-01	Absent
101	2024-03-02	Absent
101	2024-03-03	Absent
101	2024-03-04	Absent
101	2024-03-05	Present
102	2024-03-02	Absent
102	2024-03-03	Absent
102	2024-03-04	Absent
102	2024-03-05	Absent
103	2024-03-05	Absent
103	2024-03-06	Absent
103	2024-03-07	Absent
103	2024-03-08	Absent
103	2024-03-09	Absent
104	2024-03-01	Present
104	2024-03-02	Present
104	2024-03-03	Absent
104	2024-03-04	Present
104	2024-03-05	Present

Each student has an attendance_date entry marked as either 'Present' or 'Absent'.

Write a python code to find the students who were absent for **more than three consecutive days**. Your output should include:

- student_id
- absence_start_date (the first day of the absence streak)
- absence_end_date (the last day of the absence streak)

- Total absent days in that streak

The above output should **only have the latest absence streak** of the student.

Expected output from step1:

student_id	absence_start_date	absence_end_date	total_absent_days
101	2024-03-01	2024-03-04	4
102	2024-03-02	2024-03-05	4
103	2024-03-05	2024-03-09	5

Step 2:

Using the output from Step 1, write a Python program that:

1. **Join this data** with a students table that contains additional student information.

Table: students

student_id	student_name	parent_email
101	Alice Johnson	alice_parent@example.com
102	Bob Smith	bob_parent@example.com
103	Charlie Brown	invalid_email.com
104	David Lee	invalid_email.com
105	Eva White	eva_white@example.com

2. **Check for valid emails** (A valid email follows the pattern: something@domain.com).
    a. Email should have @.
    b. The @ character should be followed by domain.com (e.g. @gmail.com is valid, @gmail or @gmailcom is not valid)
    c. Characters before @ **should not have special character (except '_')** and should not start with number.
3. Add two new columns to the output of step 1:
   a. only valid parent emails against the student IDs(Valid email as per bullet point 2) (Column Name in output: **email**)
   b. a message in the below format only for students which have a valid parent email (Column name in output: **msg**):
   "Dear Parent, your child [Student Name] was absent from [Absence Start Date] to [Absence End Date] for [Total Absent Days] days. Please ensure their attendance improves."
4. The code should return the final dataframe as output.

Final expected output:

student_id	absence_start_date	absence_end_date	total_absent_days	email	msg
101	01-03-2024	04-03-2024	4	alice_parent@example.com	Dear Parent, your child Alice Johnson was absent from 2024-03-01 to 2024-03-04 for 4 days. Please ensure their attendance improves.
102	02-03-2024	05-03-2024	4	bob_parent@example.com	Dear Parent, your child Bob Smith was absent from 2024-03-02 to 2024-03-05 for 4 days. Please ensure their attendance improves.
103	05-03-2024	09-03-2024	5	None	None

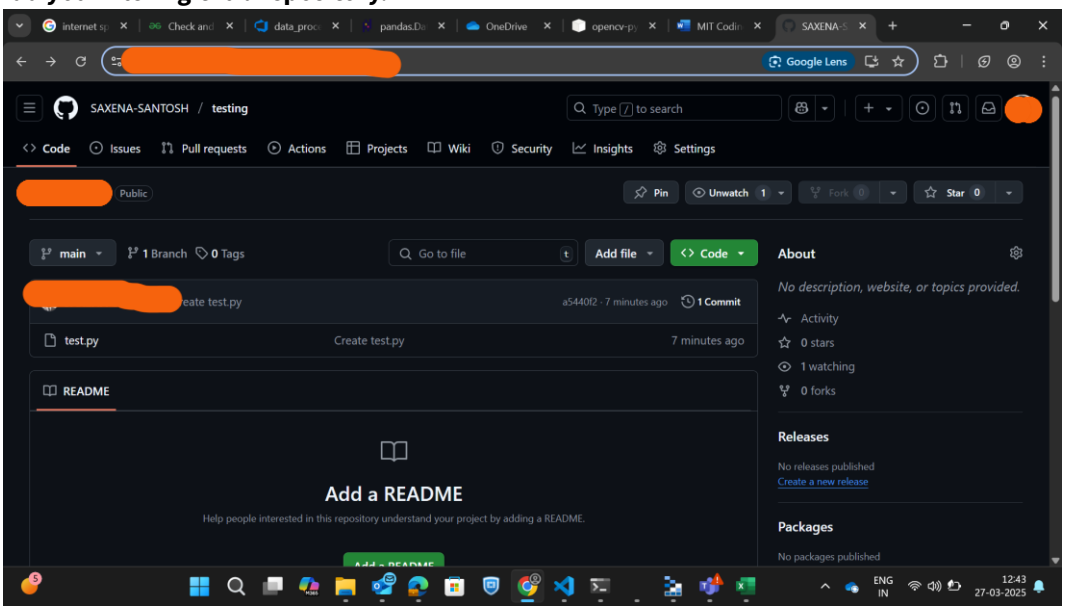
Expected columns : ['student_id', 'absence_start_date', 'absence_end_date', 'total_absent_days', 'email', 'msg']

Note:

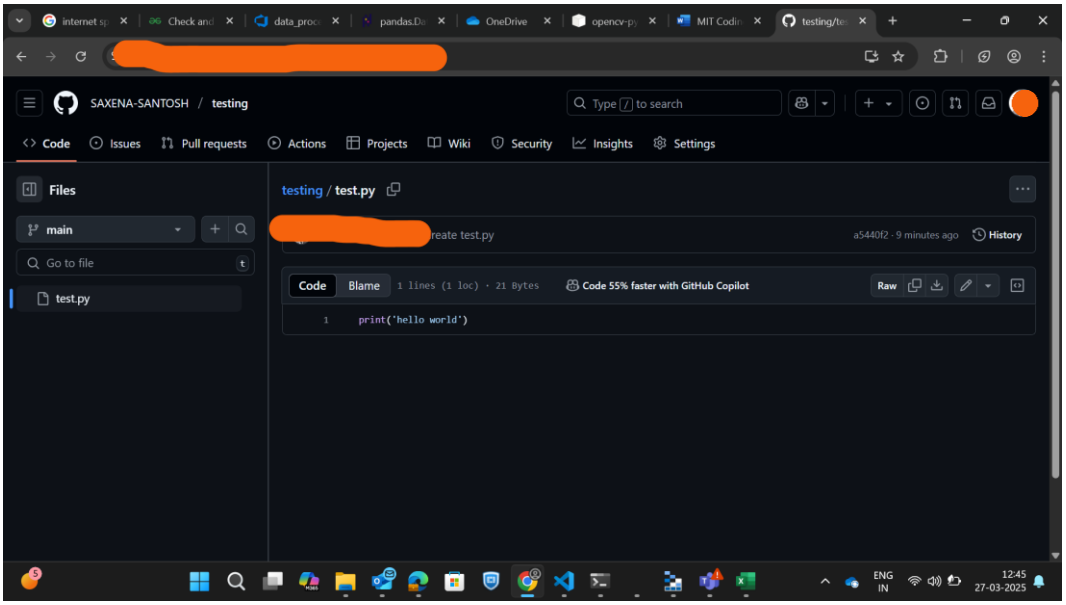
- **Ensure that the column names are strictly matching with the expected output.**
- **You are requested to edit the run() function in the python file sent to you.**
- **run() function should return pandas dataframe only.**
- **File name should be your ROLLNO.py. For Ex. 12345.py**
- **Submission Link: https://forms.office.com/r/4ctAErYK4B**
- **Any deviation from the expected output may lead to disqualification.**

Upload Instructions:

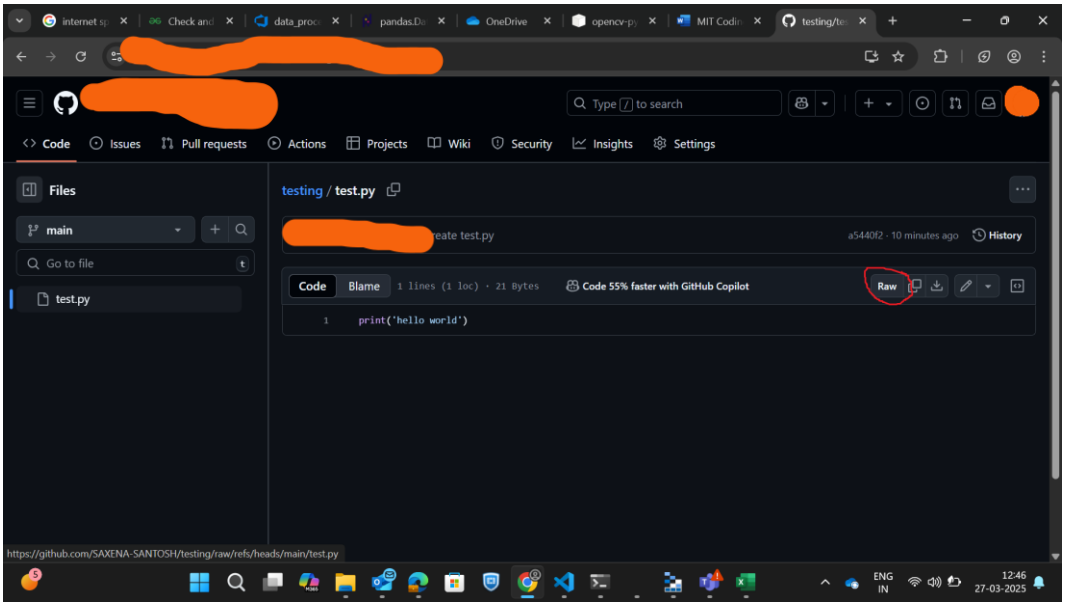
1. **Instructions to upload a file**
    a. **Add your file in github repository.**



b. Go to the uploaded file.



c. Click on raw button.



d. Copy the url and submit it in the form.