**Step to web scraping**

Step 1: Sign up to the website

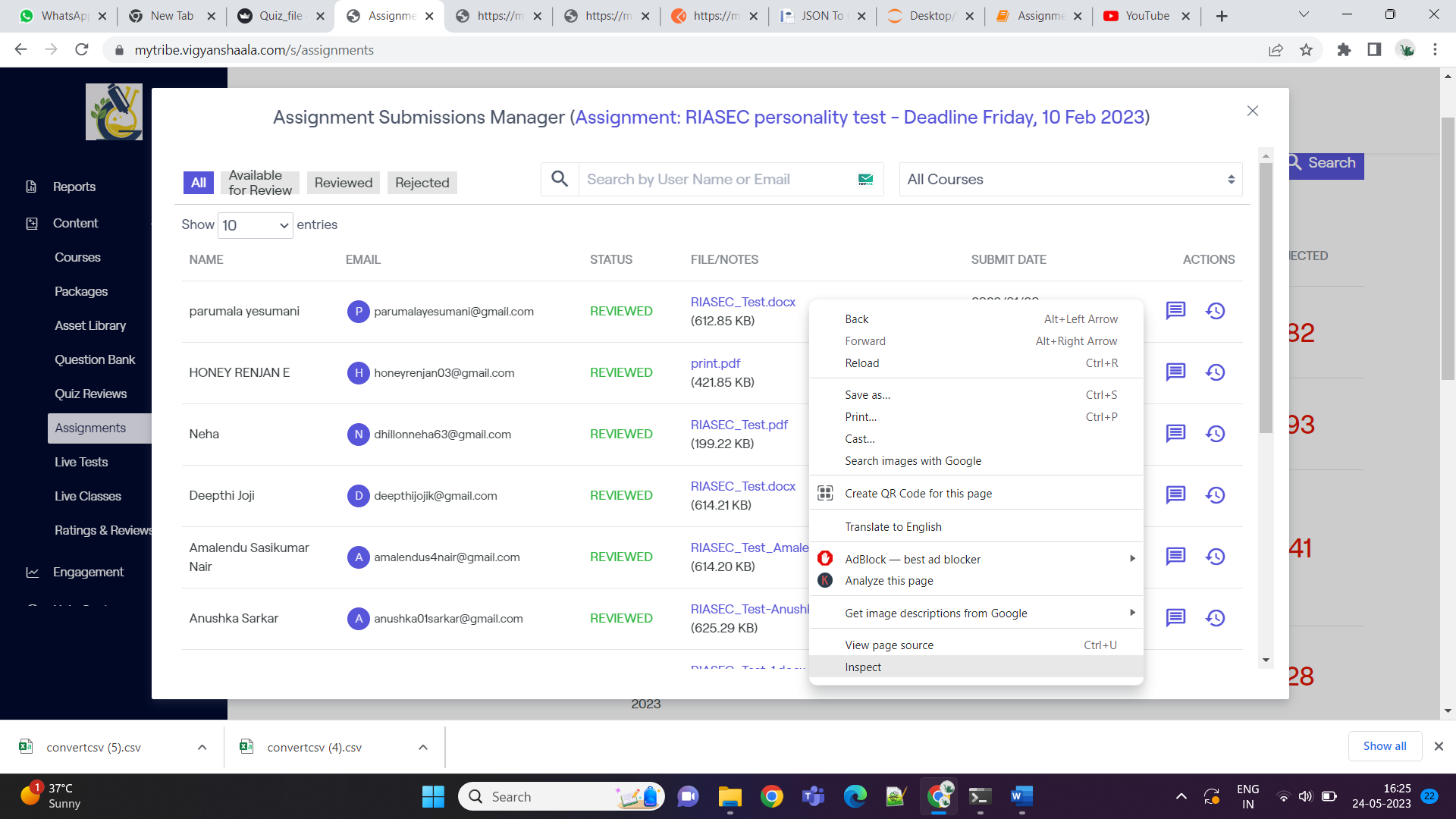
Go to the website: <https://mytribe.vigyanshaala.com>

Enter the given password (provided by the project manager) during the sign-up process.

Step 2: Go to the page you want to extract data from, for example, "Assignment review."

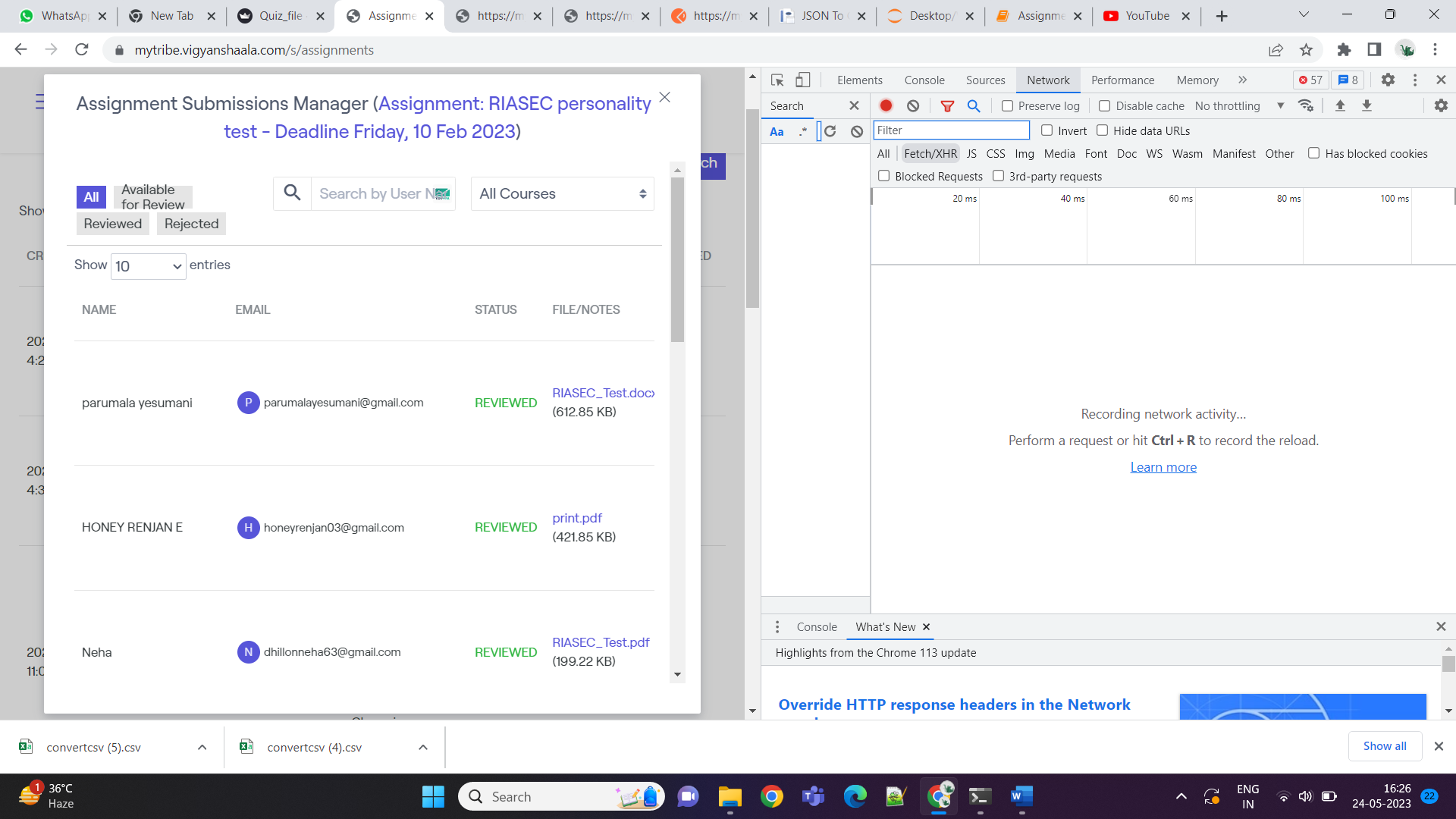
Step 3: Locate the Assignment file you want to extract data from and click on the desired data.

Step 4: Right-click and select "Inspect" from the context menu.



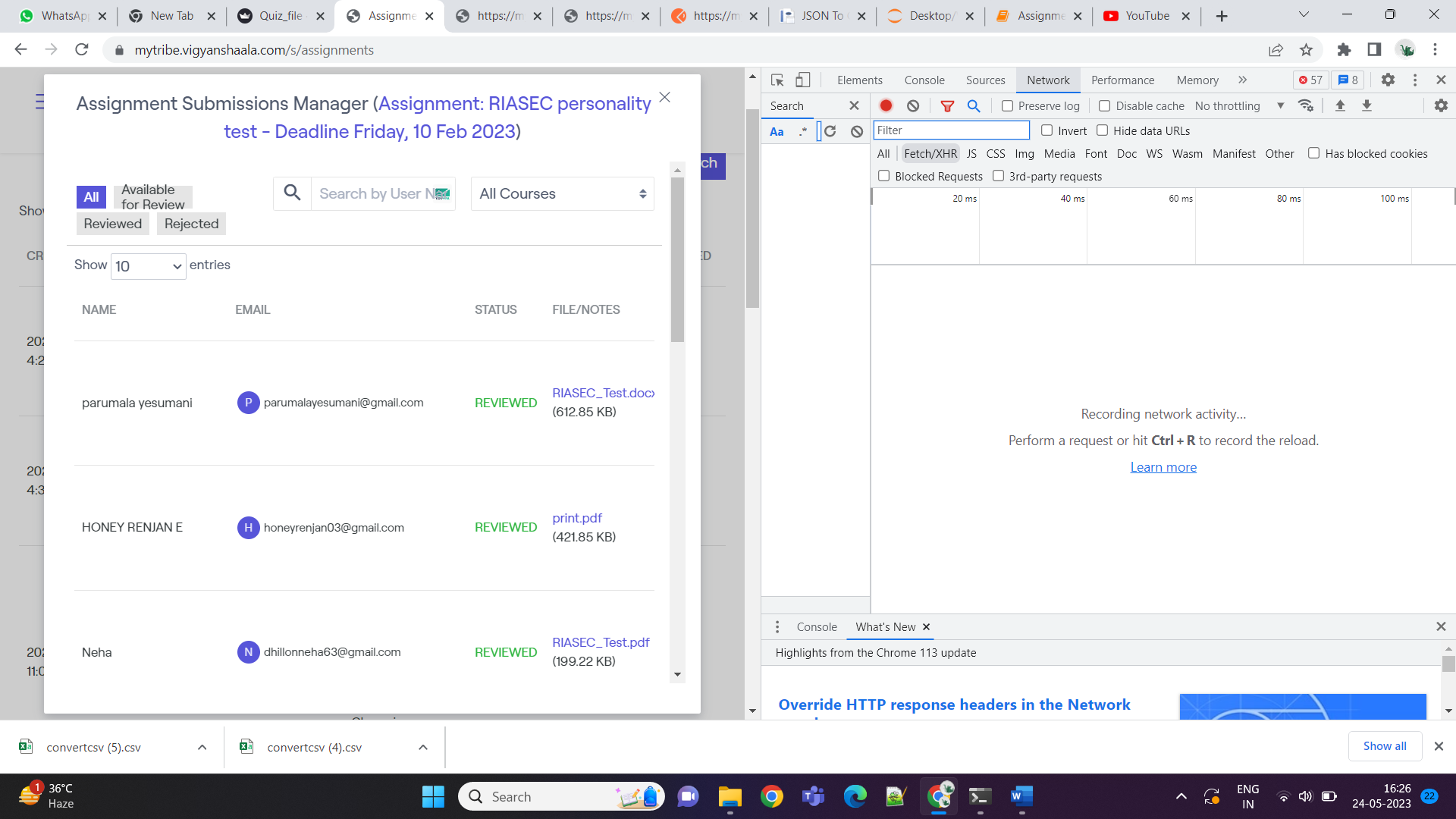


Step 5A: In the opened Developer Tools panel, navigate to the "Network" tab.

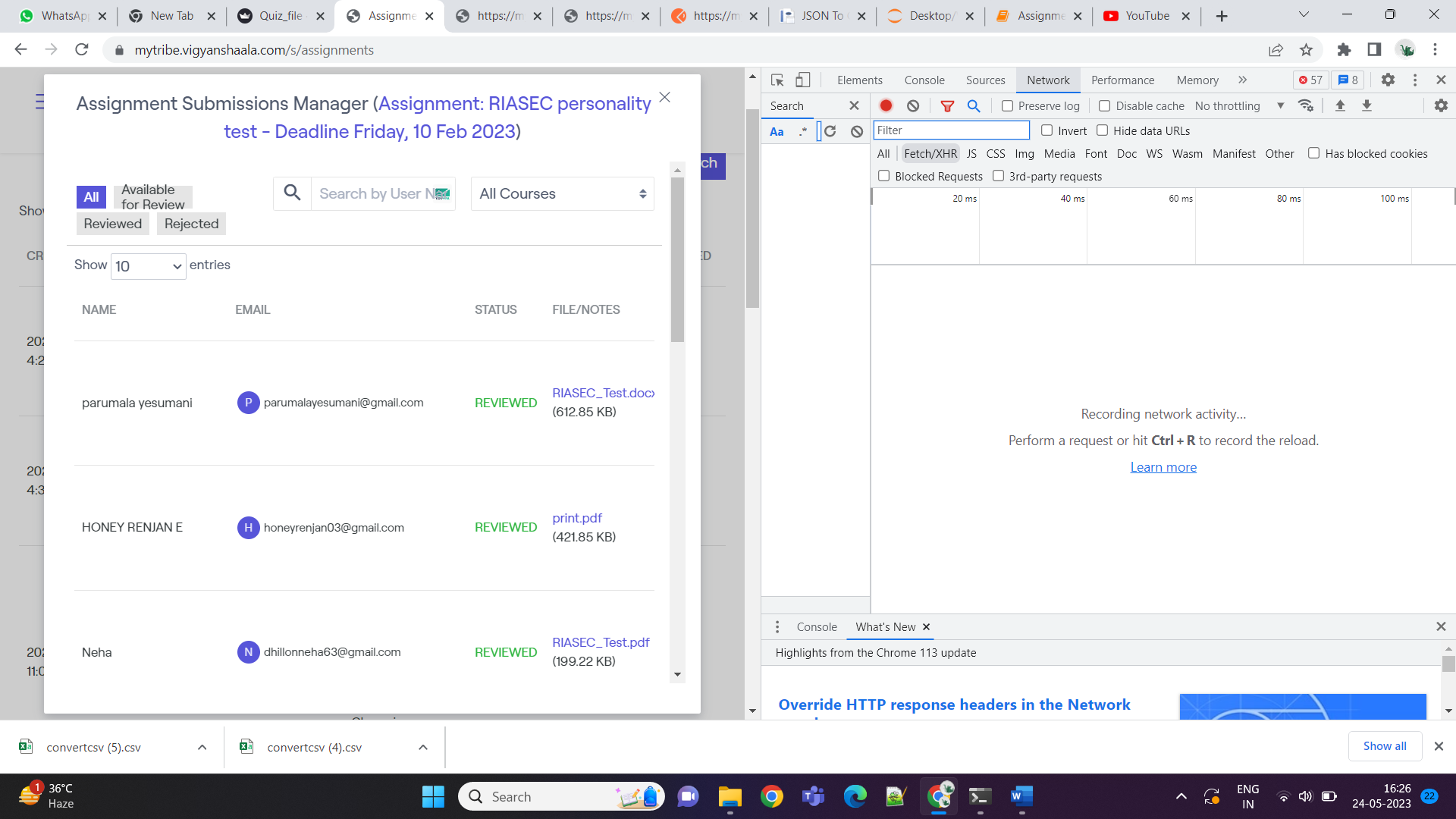




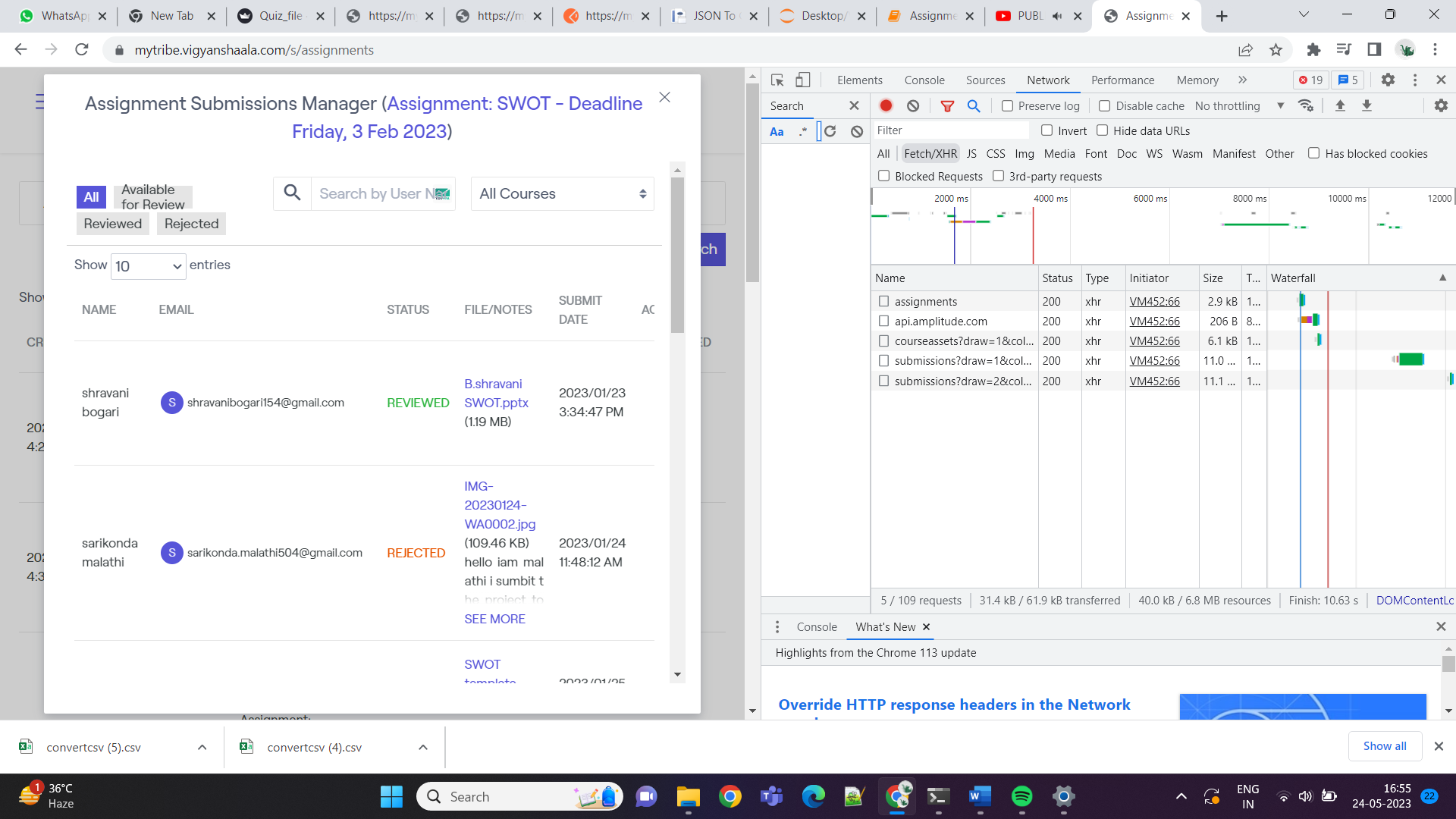
Step 5B: Now go and click Fetch/XHR



Step 6: Press Ctrl + R to refresh the page. This will start recording the network activity.



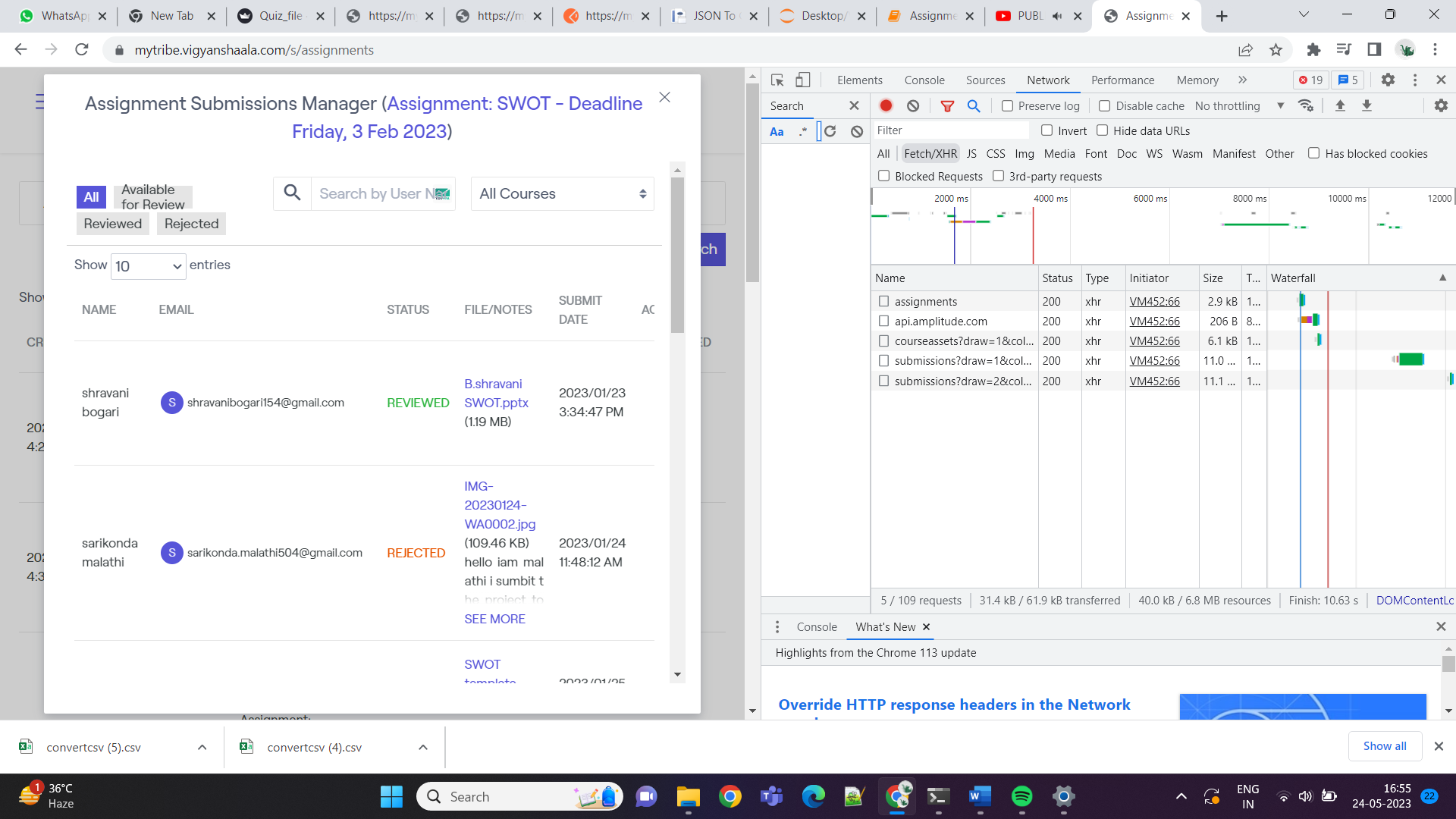
Step 7: Select the specific file you want to extract data from and choose the "All" option. This will include all entries (reviewed, rejected, and under review) in the recorded JSON file.





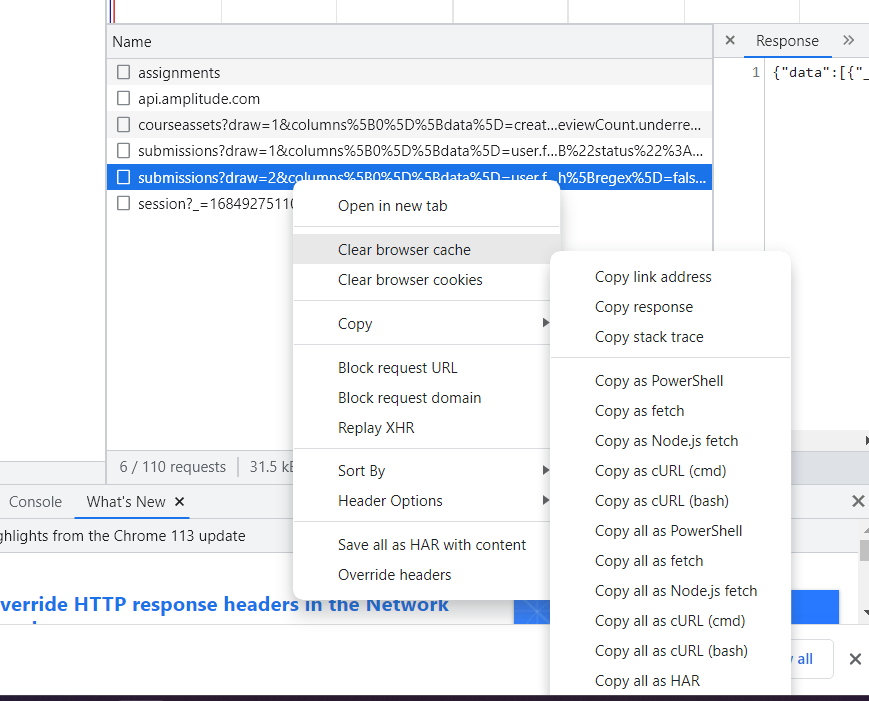
Step 8: A new submission file will be created, containing JSON data. Always select the most recent data that you want to extract. So, whenever you go to another page or make changes in current one. Always new data will get collected.





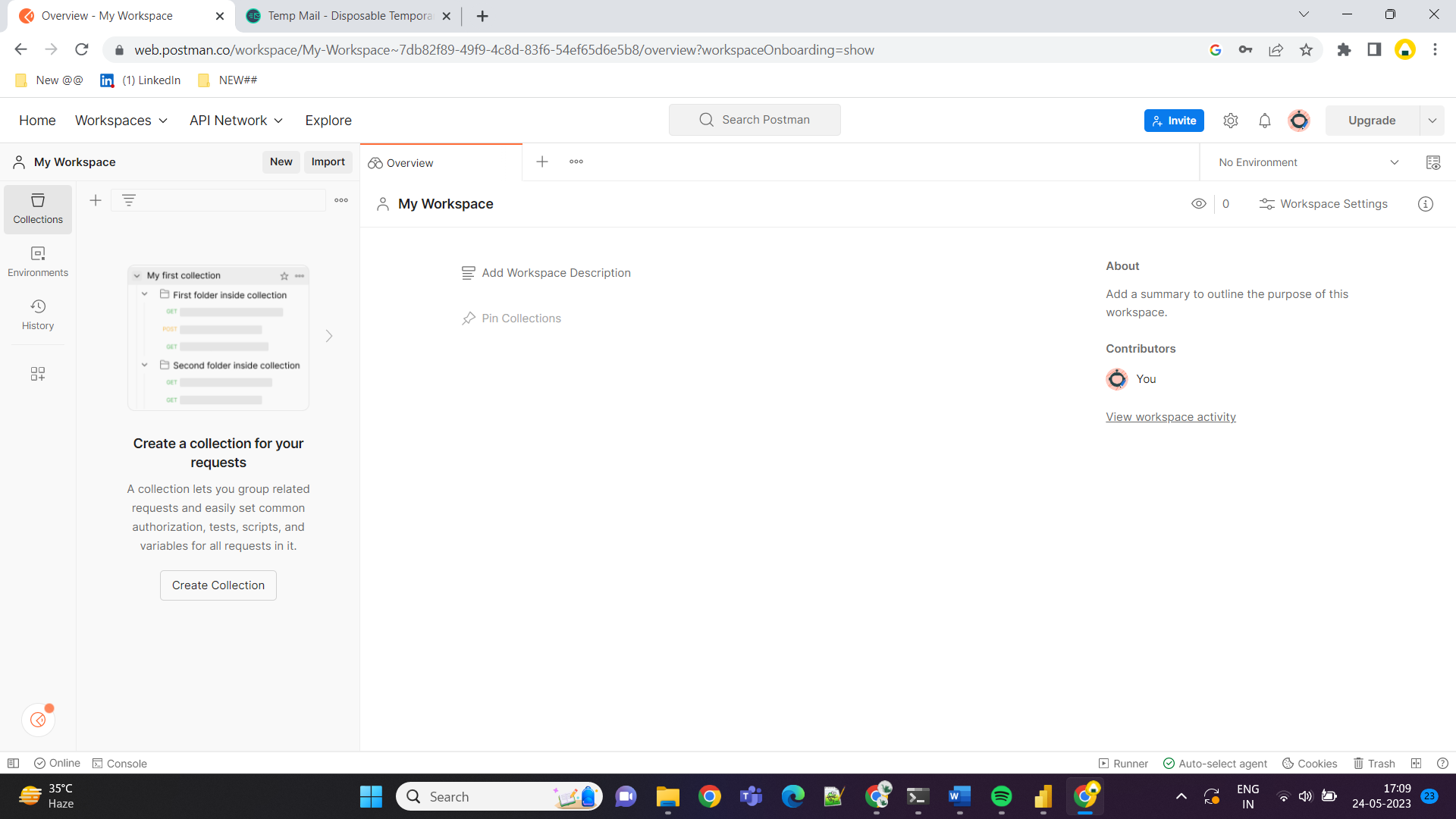


Step 9: Right-click on the latest entry in the submission data and choose "Copy." Then, select "Copy as cURL (bash)."





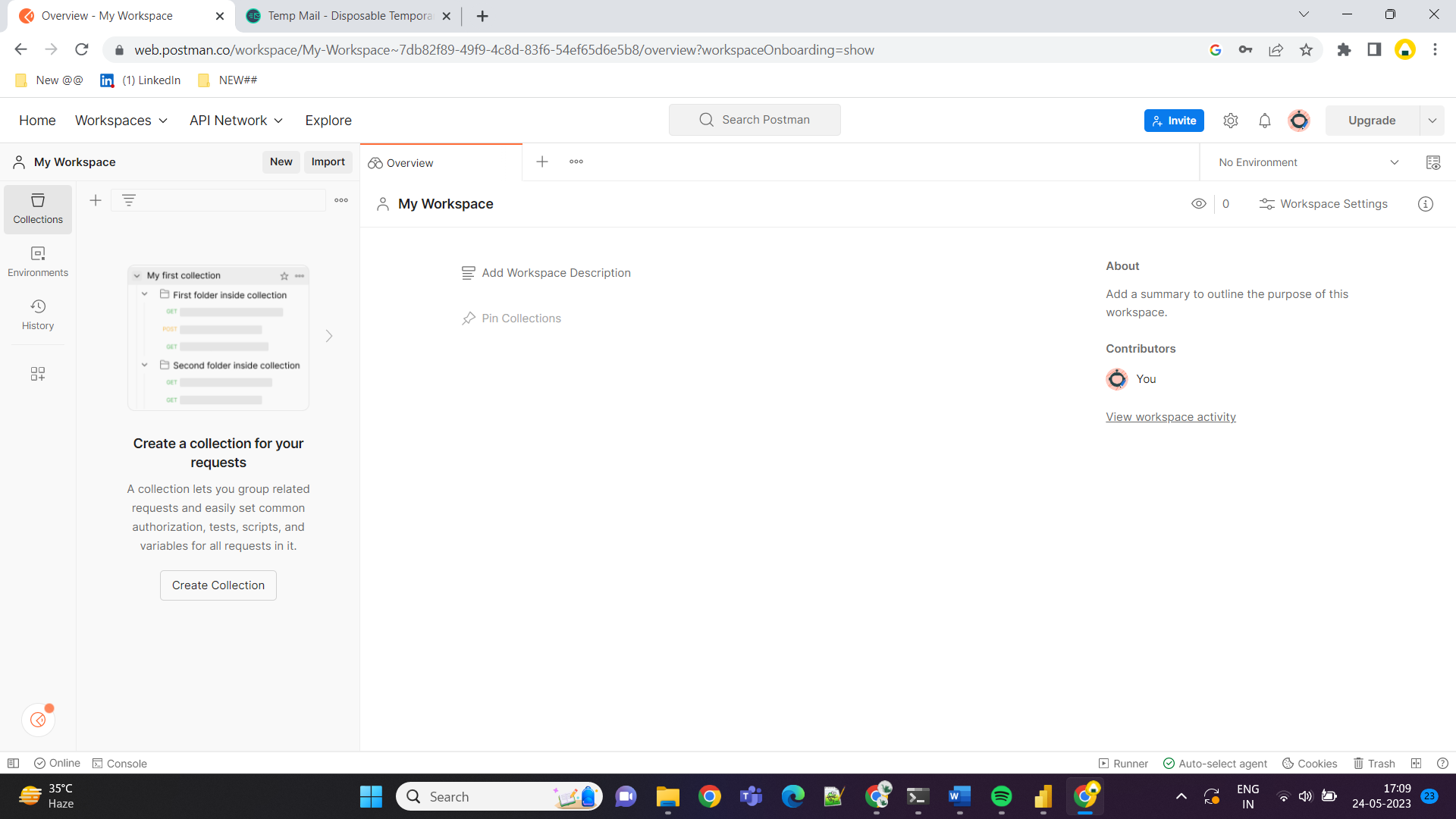
Step 10: Open Postman (<https://www.postman.com>) And create an account if you haven't already. Once logged in, you will see the dashboard of your workspace.



Step 11: Click on "Import" in Postman and paste the data copied in Step 9 as cURL (bash).

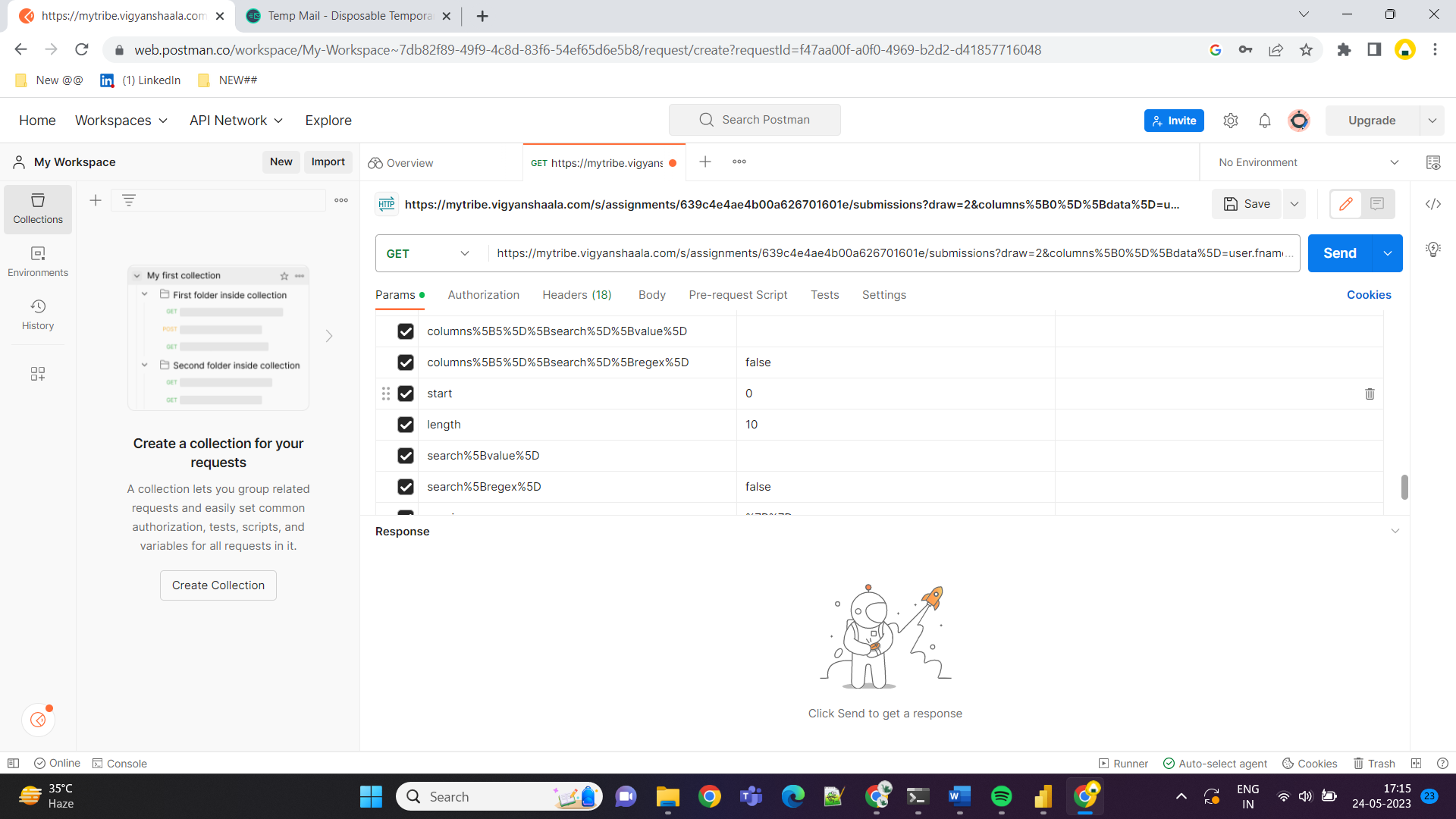
NOTE: While copy data in step 10 copy data of Copy as curl(bash) not

~~Copy as all curl(bash)~~. As if we select copy as all curl(bash)then at end in postman we cannot see data or attributes. So, it’s necessary to choose right file.



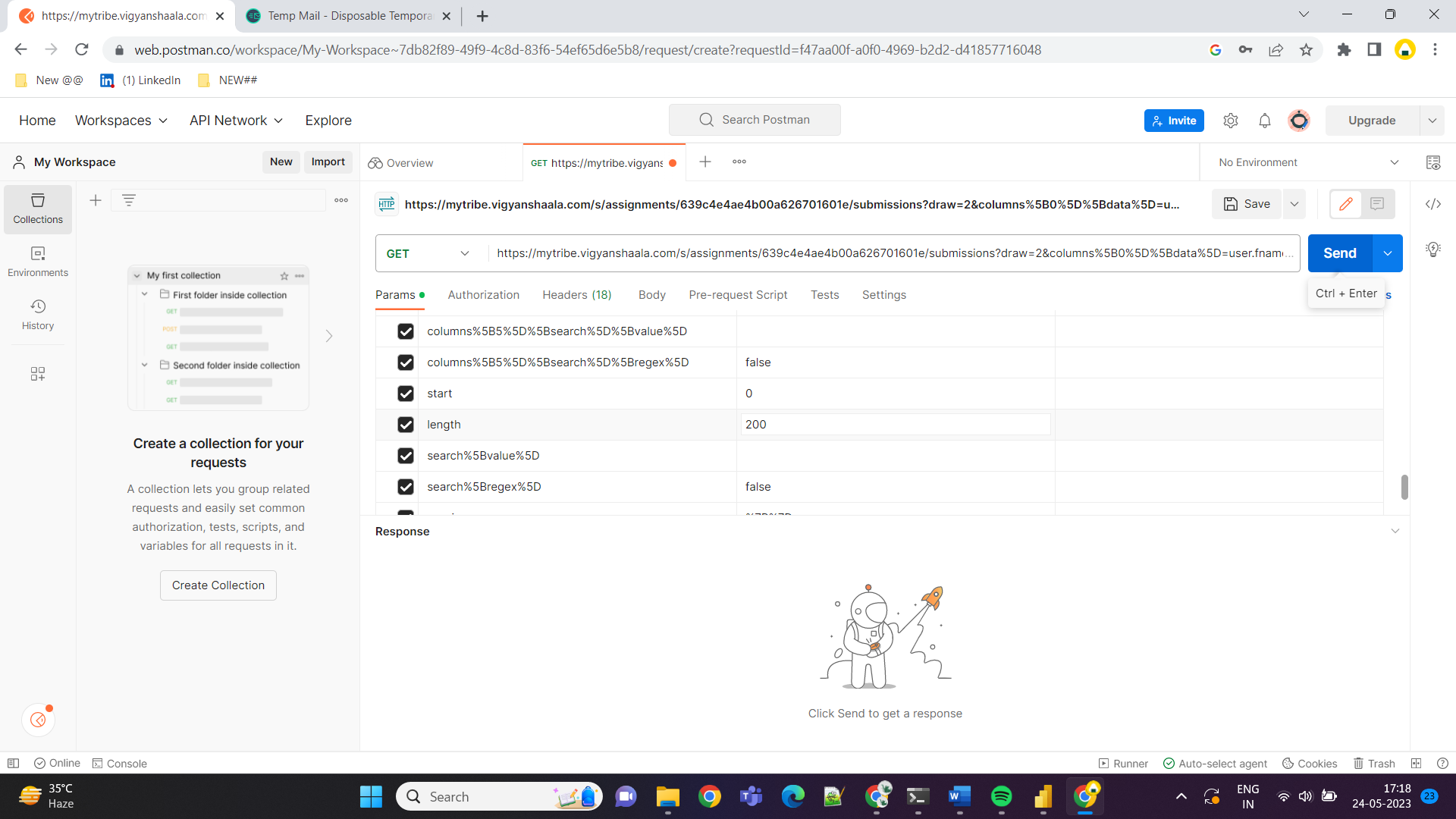


Step 12: Modify the "length" attribute to the number of rows you want to extract or a rounded-up number. For example, if you estimate there are around 145 rows, you can set it to 200 to include all rows.



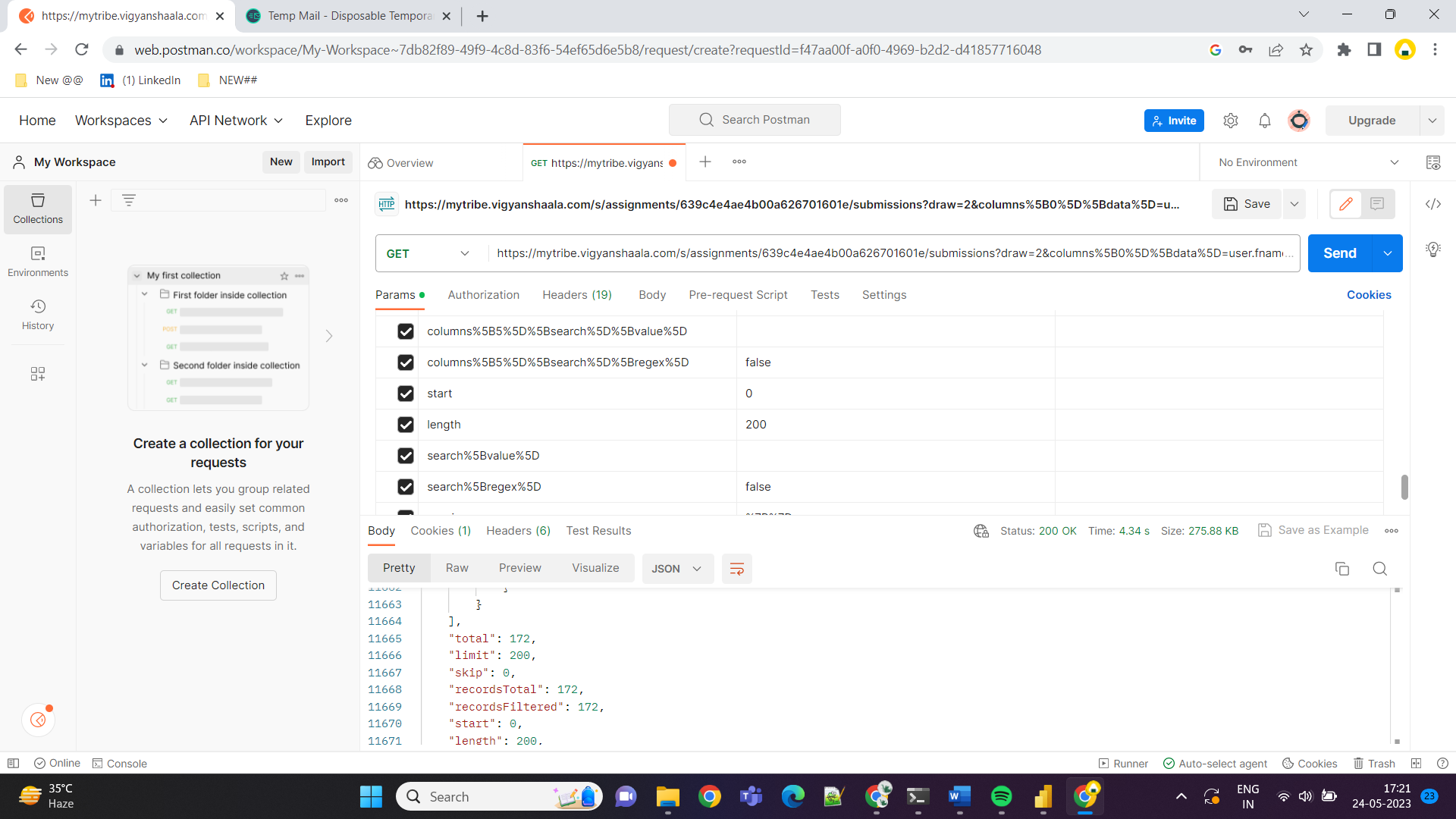


Step 13: Click on "Send" after adjusting the length value.





Step 14: The response will display JSON data. Copy the entire JSON data.

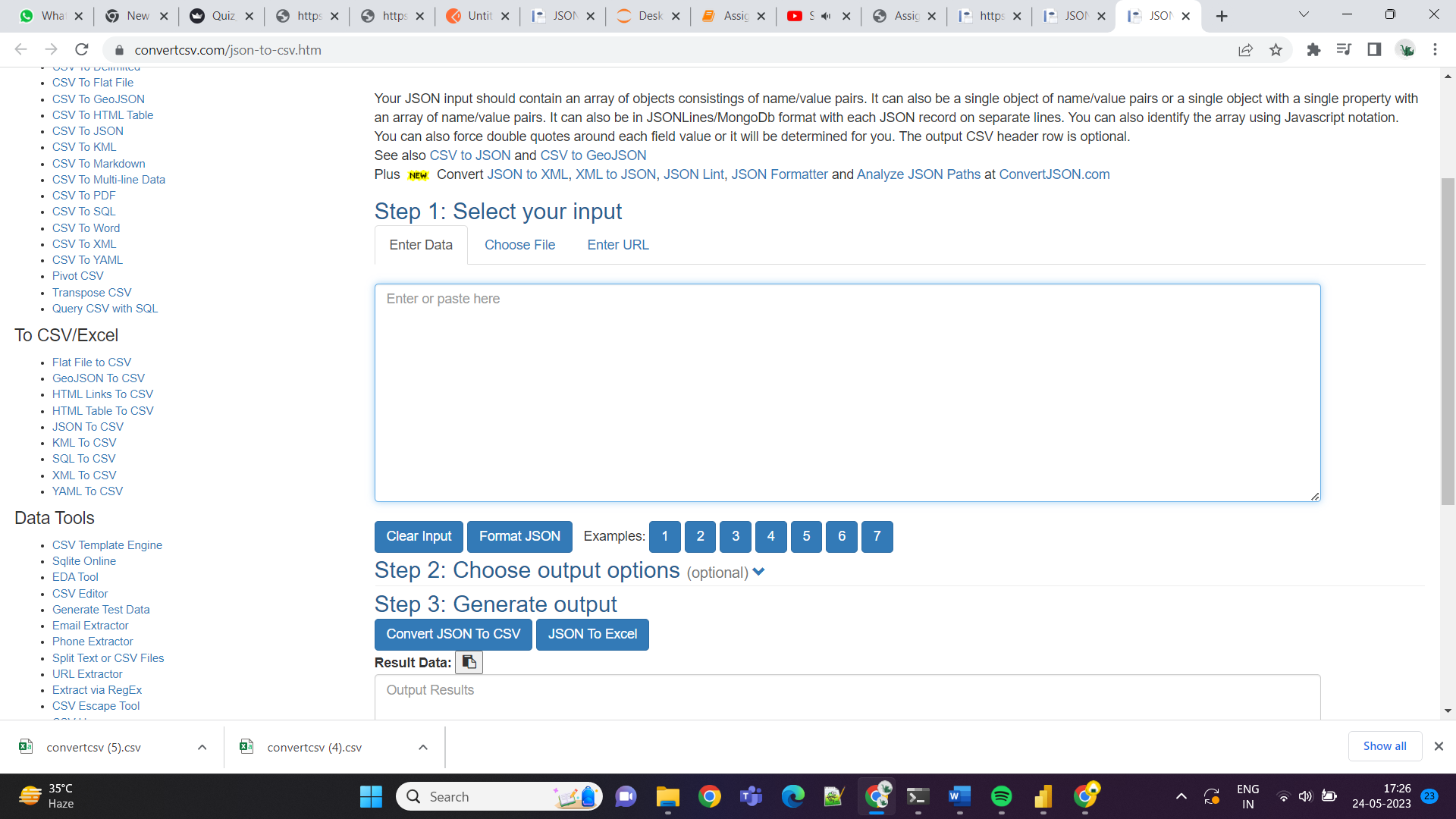




Step 15: Go to a website that can convert JSON data to CSV format, such as

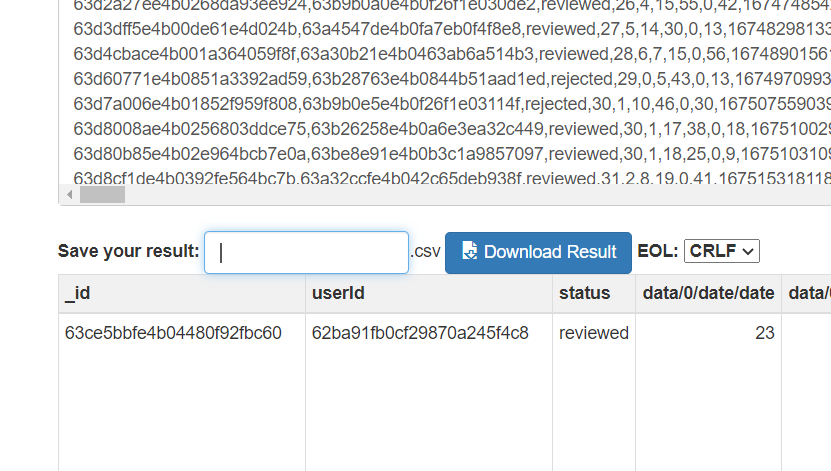
<https://www.convertcsv.com/json-to-csv.htm>

Paste the copied JSON data into the provided field.



Step 16: Customize the output by changing the column names as desired. Use the assignment title or other important fields as column names.

Note: Make sure name while giving name should start from Assignment and also don’t take After hyphen. for example, if name is Masterclass Assignment: Planning & Applying for Masters in India & Abroad then while renaming put Assignment: Networking: Build a World-Class LinkedIn Profile.



Step 17: Click on "Download Result" to obtain the converted CSV file. Save the file in your desired folder.

Step 18: Repeat Steps 8-17 for extracting data from other files, if necessary.

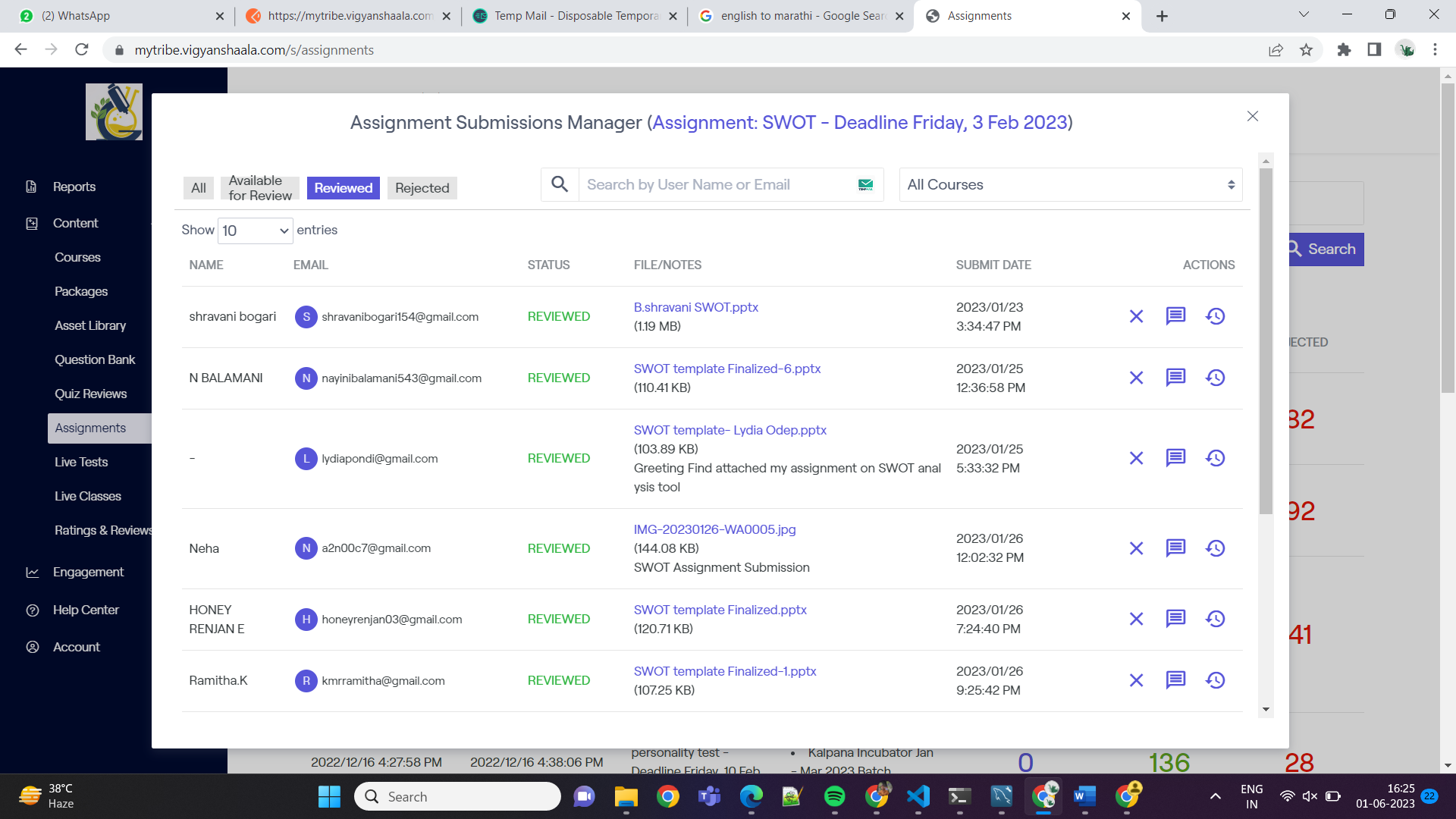
Step 19: Review and putting files into define folder

Open the downloaded CSV file and review the extracted data to ensure its accuracy and completeness. And then paste the files in given location.

**Giving name to files:**

Change the Name as desired. When customizing the column names in the CSV output, it's important to consider the information you want to include while keeping the naming consistent. Like we have to start name form Assignment and if you find anything Deadline or Date then don’t take it while giving name. Here's an example based on the provided file names:

From the original file names:





Assignment: Critical thinking - Deadline Friday, 10 Mar 2023

Assignment: Masterclass - Academic CV and Industry Resume

Assignment: Career Action Plan (CAP) - Deadline Friday, Feb 24, 2023

Assignment: Masterclass 1 - Career Exploration

Assignment: SWOT - Deadline Friday, 3 Feb 2023

To maintain consistency and remove unnecessary details, you can modify the names to start with "Assignment" and exclude the deadline information:

Modified names:

Assignment: Critical thinking

Assignment: Masterclass - Academic CV and Industry Resume

Assignment: Career Action Plan (CAP)

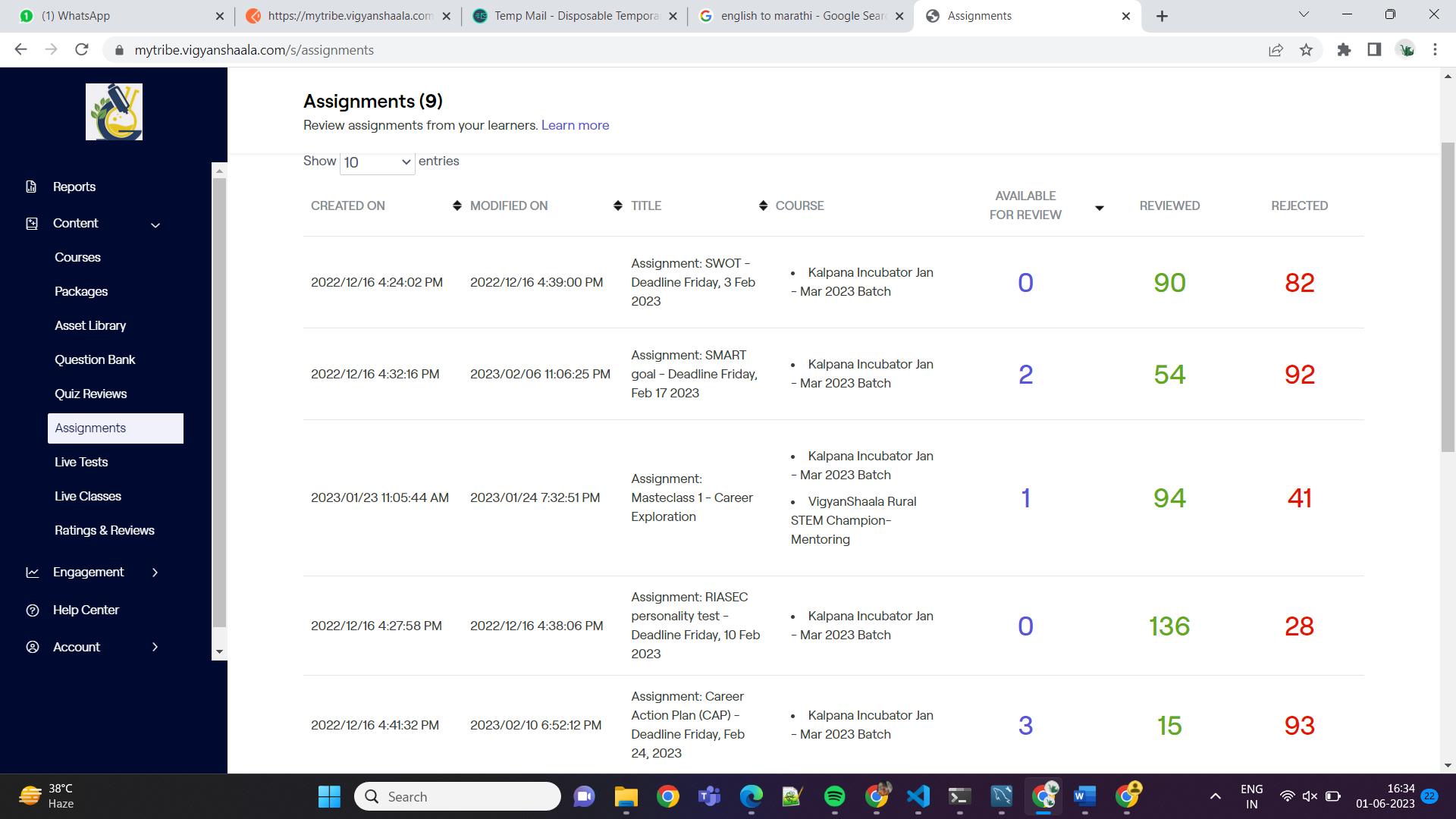
Assignment: Masterclass 1 - Career Exploration

Assignment: SWOT



By adopting a standardized naming format, such as starting with "Assignment" and excluding specific details like deadlines, you can help while data cleaning process.







**Common Mistakes:**

* Make sure to copy the data using the "Copy as cURL (bash)" option in Step 9, not "Copy as all cURL (bash)."
* Double-check the URL and password to ensure they are entered correctly.
* Pay attention to the network tab in the Developer Tools panel to select the appropriate file for extraction.
* When modifying the length in Postman, be mindful of the number of rows to include, ensuring it covers all the required data.
* Ensure the correct JSON data is copied and pasted into the JSON to CSV conversion tool.
* Verify the column names in the CSV file are accurately reflecting the desired information for analysis.

**Why we Used:**

1. Postman:

Postman is used in this step because it provides a convenient platform for working with APIs and handling HTTP requests. By creating an account and accessing the Postman dashboard, you can leverage its features for the following purposes:

* Converting cURL Command: The cURL command copied in Step 10 needs to be converted into a format that can be easily managed and modified. Postman allows you to import the cURL command and automatically converts it into a more user-friendly representation. This makes it easier to adjust the parameters, such as the length attribute, to control the data extraction process.
* Modifying Length Attribute (Pagination): In Step 12, the length attribute is modified to specify the number of rows to retrieve in each request. Postman provides a straightforward way to edit and customize this attribute, allowing you to adjust it according to your requirements.
* Sending Requests: Once the cURL command is imported and the length attribute is modified, Postman enables you to send the modified request to the server. This initiates the data retrieval process and allows you to obtain the JSON response containing the desired data.

1. Convertcsv

Link: <https://www.convertcsv.com/json-to-csv.htm>

* Simple to understand and perform.
* We can give name to file as we wanted before downloading.
* However, there are multiple alternatives available for performing this conversion. There is various other website to conversion.

**Summary:**

The steps provided above outline how to perform web scraping on a specific website, specifically using the Postman tool to interact with the website's API and convert JSON data to CSV format. The process involves signing up to the website, inspecting the network activity, copying the API request as a cURL command, and using Postman to modify the request and retrieve the desired data. The JSON data obtained is then converted to CSV format using an online tool.

Here's a summarized list of the steps:

1. Sign up to the website.
2. Go to the page with the data you want to extract.
3. Use the browser's Developer Tools to inspect the network activity.
4. Copy the API request as a cURL command.
5. Open Postman and import the cURL command.
6. Modify the request by adjusting the length attribute for pagination.
7. Send the request in Postman to retrieve the JSON data.
8. Copy the JSON data obtained from Postman.
9. Use an online tool like "convertcsv.com" to convert the JSON data to CSV format.
10. Customize the column names in the CSV output as desired.
11. Download the converted CSV file and save it in your desired folder.

The reason for using Postman is that it provides a user-friendly interface for working with APIs and handling HTTP requests, making it easier to modify and send requests. The "convertcsv.com" tool is chosen because it allows you to give custom names to columns, which can be helpful for organizing and cleaning the data.

It's important to be careful with the steps and avoid common mistakes to ensure the accuracy of the extracted data. Double-checking the URL, password, and network activity is crucial to avoid errors. Following a consistent naming format for the files can also help in data cleaning and analysis later.