Selection Test for RA internship at Ashoka and Krea University

What are we looking for:

- How comfortable are you when coding in python in terms of efficiency in accomplishing a said task with no errors? This includes cleaning data and removing errors.
- How well can you analyze and explain the data extracted in relation to the questions asked?
- How can you accomplish the above as quickly as possible within a time window?

Expected Output:

- Python Script File processing and analyzing the data.
- Answers to the questions in a separate document including analysis (word or pdf).

Sr No (1)	Event Title (2)	Main LCW url (3)	Number of secondary urls within Main LCW url (4)
1.	Farmers Unhappy with Compensation for Canal Project in Andhra Pradesh, Demand More	**https://www.landconflictwatch.org/conflicts/handri-neeva-sujala-water-canal-project#**	7
2.	Landowners Agreed to Give Land for the KGP Expressway	**https://www.landconflictwatch.org/conflicts/kgp-kundli-ghaziabad-palwal-expressway#**	3
3.	Residents Oppose Waste Disposal Plant by Surat Municipal Corporation in Gujarat	**https://www.landconflictwatch.org/conflicts/residents-oppose-waste-disposal-plant-in-bhandut-gujarat#**	**6**

Table 1

Questions:

There are three events described in Table 1 above. Column 2 gives the event title. Column 3 gives the primary URL. Column 4 gives the number of secondary URLs contained within the primary URL. Based on table 1, please answer the following questions

a) Extract/Web Scrape all the information for each event given in the Main LCW url (Table 1, column 3)
b) Using the information extracted in part a) identify, sort and provide the following data. Explain briefly your logic of selection.

   i) Summary of the event using 20 words max
   ii) Start date of the conflict
   iii) Land area affected
   iv) Sector of the conflict event
   v) Reasons/Cause of the conflict
   vi) Legal laws violated
   vii) Demand of the affected community
   viii) Type of Land
   ix) Total Investment (if any)

c) Extract/Web scrape all the information for each event from all the secondary urls (Table 1 col 4)
d) Sort, present and categorize the textual information extracted from each secondary urls in part c). You can use data headings presented in i)-ix) to present the extracted information. This should be a separate output than in part a) and b).
e) Please use your logic on presenting data in part d) in relation to part b)? What is your inference on the information provided for each of the three events in Table 1? Explain briefly.