**DOCUMENTATION FOR THE TASK**

In the assessment I used the link <https://sam.gov/data-services> and <https://eprocure.gov.in/eprocure/app> for data extraction. In link1 it was easy to retrieve the data as it was already provided in .csv form whereas in link2, I have to make an autonomous code to retrieve the data from site by using selenium and beautifulsoup respectively.

CHALLENGES FACED

* In code to find the class on the website. Although it’s done by using the inspect feature on the website.
* In link2, the data is not provided easily so I have to research for so long to find genuine data.
* In link2, after finding the genuine data, it is provided in the table and we have to click on the next button to see more entries and each entry can be added manually, so the autonomous code is required in this case.

WORK DONE

* Downloaded 3 .csv files from link1 to inspect the data.
* Cleaned the data using library fuzzywuzzy for fixing the spelling as required in documentation.
* Data taken for cleaning is 150 entries. From the Link1
* Made an autonomous code for data retrieving from link2. (which is challenging part in assignment).
* Made the Scalability and Production Readiness document for future development.

My understanding towards the assignment is that you need a complete pipeline for this.

For example: - A URL can be uploaded to the code and the data is collected then, another automated code for its trim and cleanliness.

I think, due to the time crunch I couldn’t give my 100% but it can be improved much better. I would love to work on this project if we move forward.

THANK YOU.