The assignment could be approached from multiple ways:

Way 1 - We can directly use LLM APIs to prompt them into visiting the website and giving us the required data. This is the easiest and the major disadvantage is that the LLM frequently provides wrong data (by visiting some other source etc.)

Way 2 – We can scrape the website for the data convert it into Pandas Dataframes and prompt LLM with this as context to gather the required data. This could be hard to implement but will definitely provide more accurate data.

I went ahead with the second approach. The five selected companies are in “companies.xlsx” and the final output is “company\_data.xlsx”:

1. Scrape the screener.in website to get PnL and Balance Sheet data and parse them into dataframes, using BeautifulSoup and Pandas
2. Pass the dataframes as context with the prompt, to get the desired data to an LLM, using Groq and LangChain. Experimented with many models, but “llama-3.3-70b-versatile” performed the best.
3. Convert the result into a dataframe and generate the Excel out of it.

This worked, but there are some limitations:

1. If the LLM hallucinates, then parsing into Excel fails and nothing can be done.