23-04-22

**TODAYS AGENDA:--**

* What is web scrapping
* Web scrapping using python selenium
* Class work
* **TASK :- to save the information in csv or excel file using python**

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Allied modulus :- to work with python + selenium

Web scrapping = selenium + beautifulsoup

The process of gathering information from internet is called web scrapping

Company don’t give us the info about their web page

using some certain module of python to web scrapping

Information can be in either in static or dynamic

Website scrapping --- some website allows and some don’t i.e some allow u to copy from that web page but some don’t --- using special tools we can do web scrapping

Why web scrapping ? ---- many company rely on data from internet so company hire some people to collect data i.e copy and paste but if the data is dynamic and data keeps on changing, that’s the problem

Eg:- hotel reviews—google business page --- but like booking.com charge for reviews for API use || but some people do it for free using web scrapping || web scrapping --- part time job also --- eg 30 dollars using freelancers

Open flipkart --- search eg:- laptops --- there are some startups who look for trends using web scrapping --- how things are trending --- which laptop is showing 1st --- so some company always check like laptop manufacturing company they want to know the market trends --- i.e what customers are looking for , all about the product || so these all can be done using web scrapping

**pip install requests -------** in cmd

this is used to have connection with the outside world

i.e http request on the internet from local host

**pip install beautifulsoup4** --------- in cmd

this is python library --- used to pull data out of html page

also shortcut name bs4

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

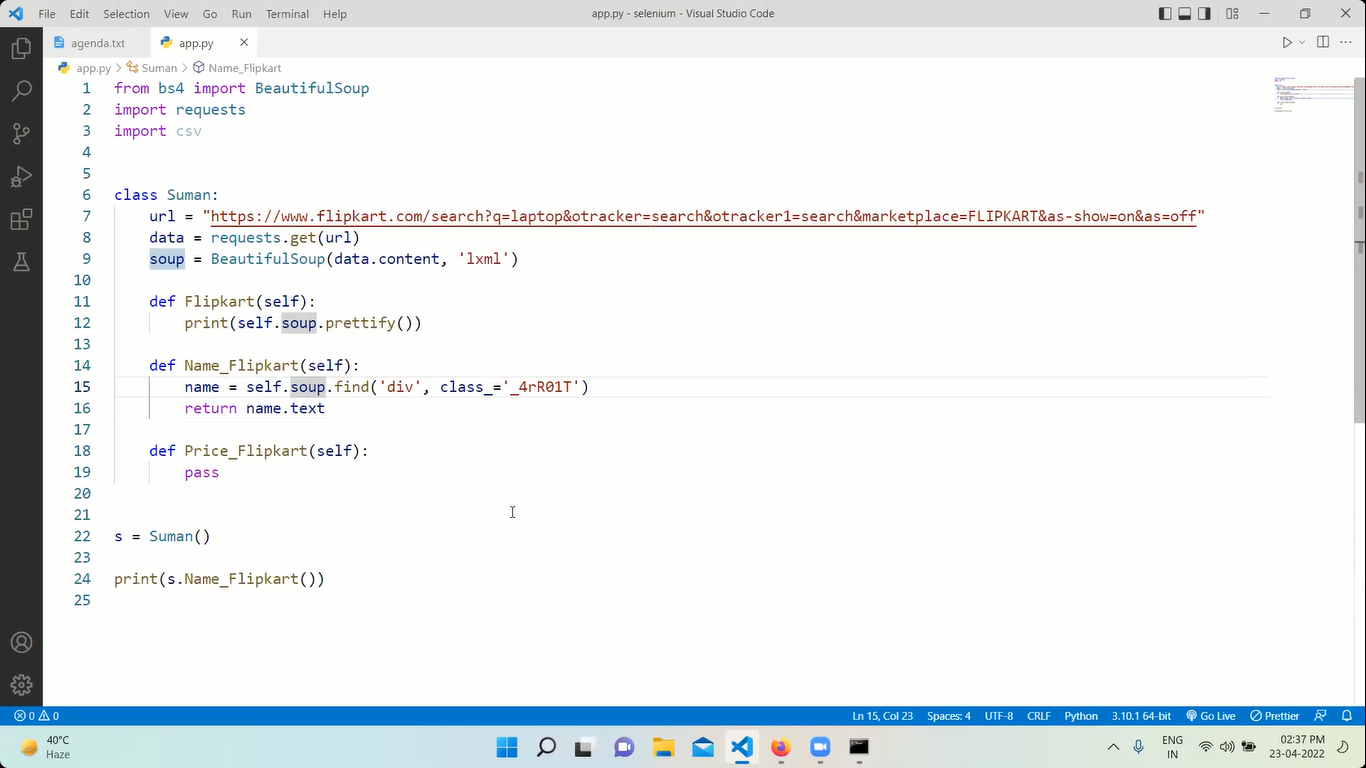
**lxml** ----- is very fast

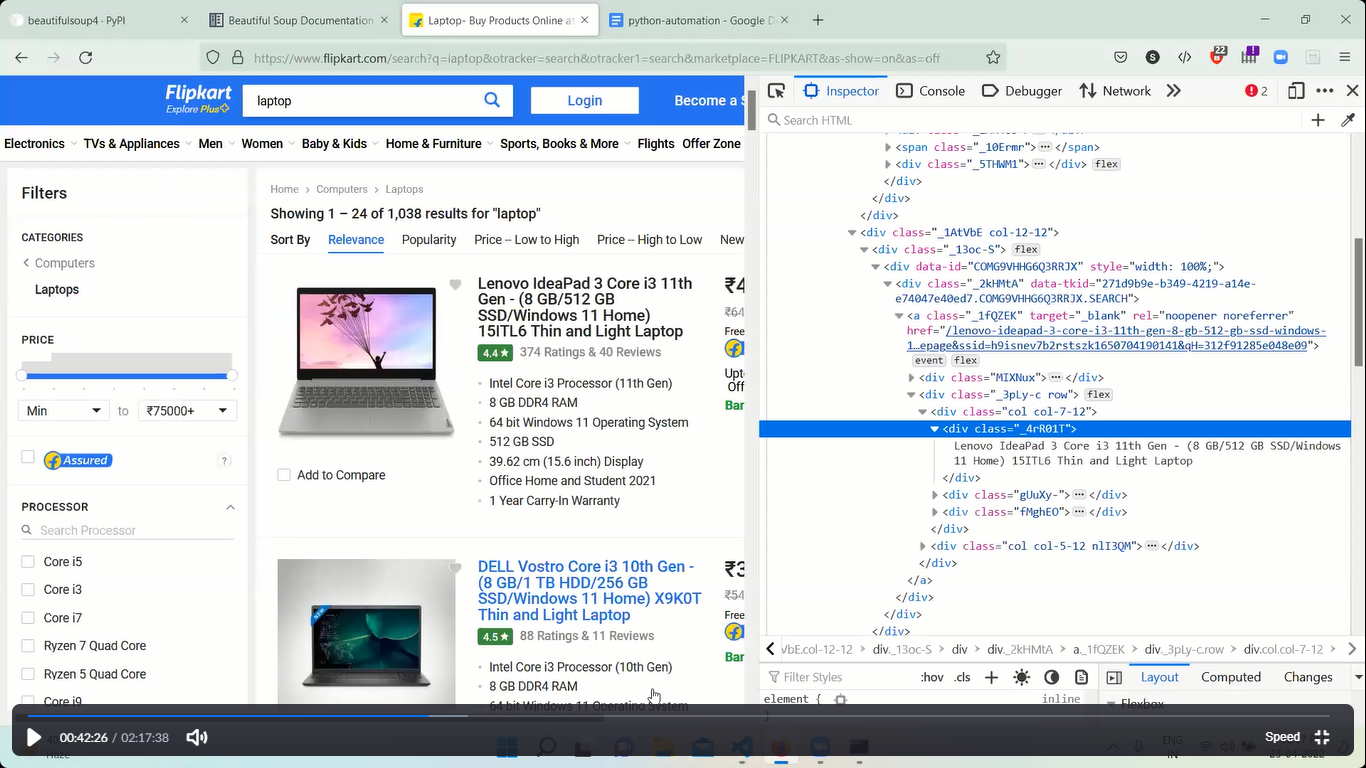
**html.parser**

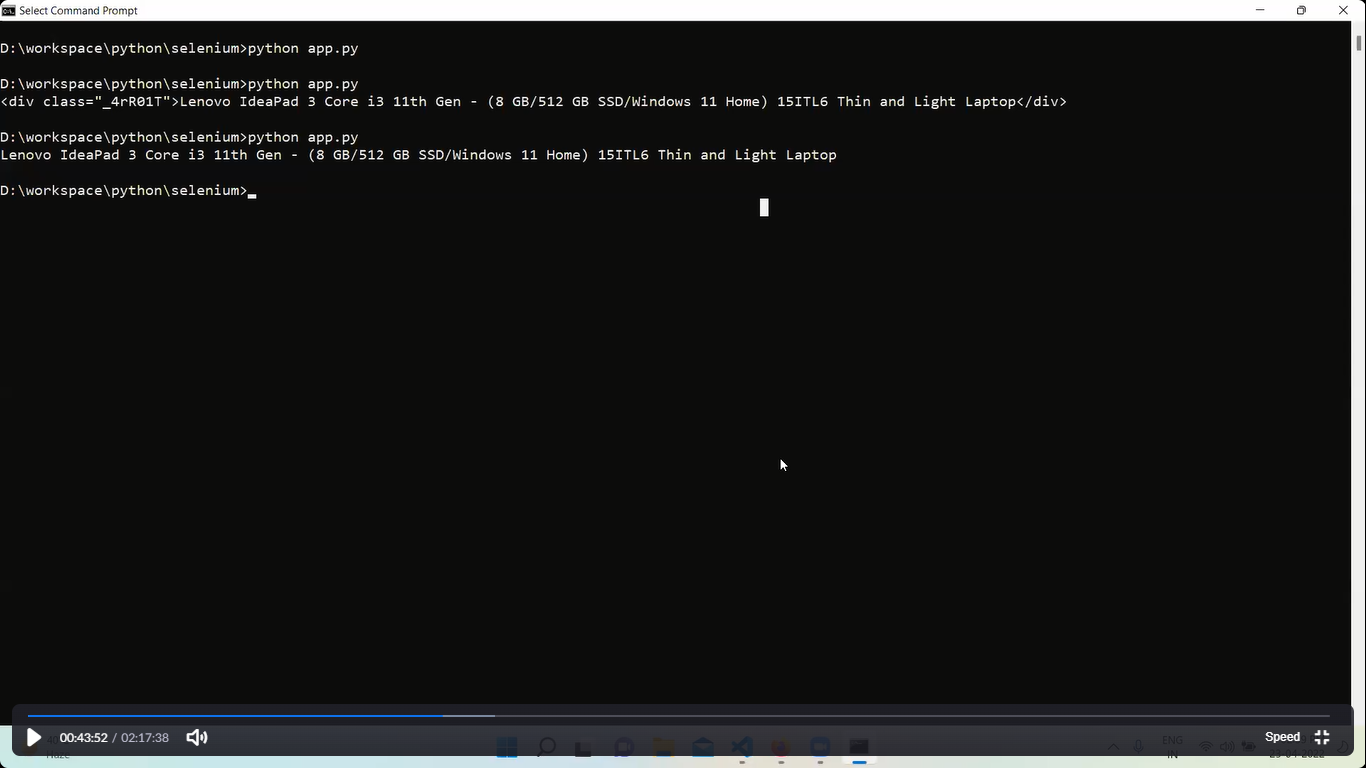
**html5lib**

to parse a markup language page we need a parser engine ---- the above are the 3 parse engines ---

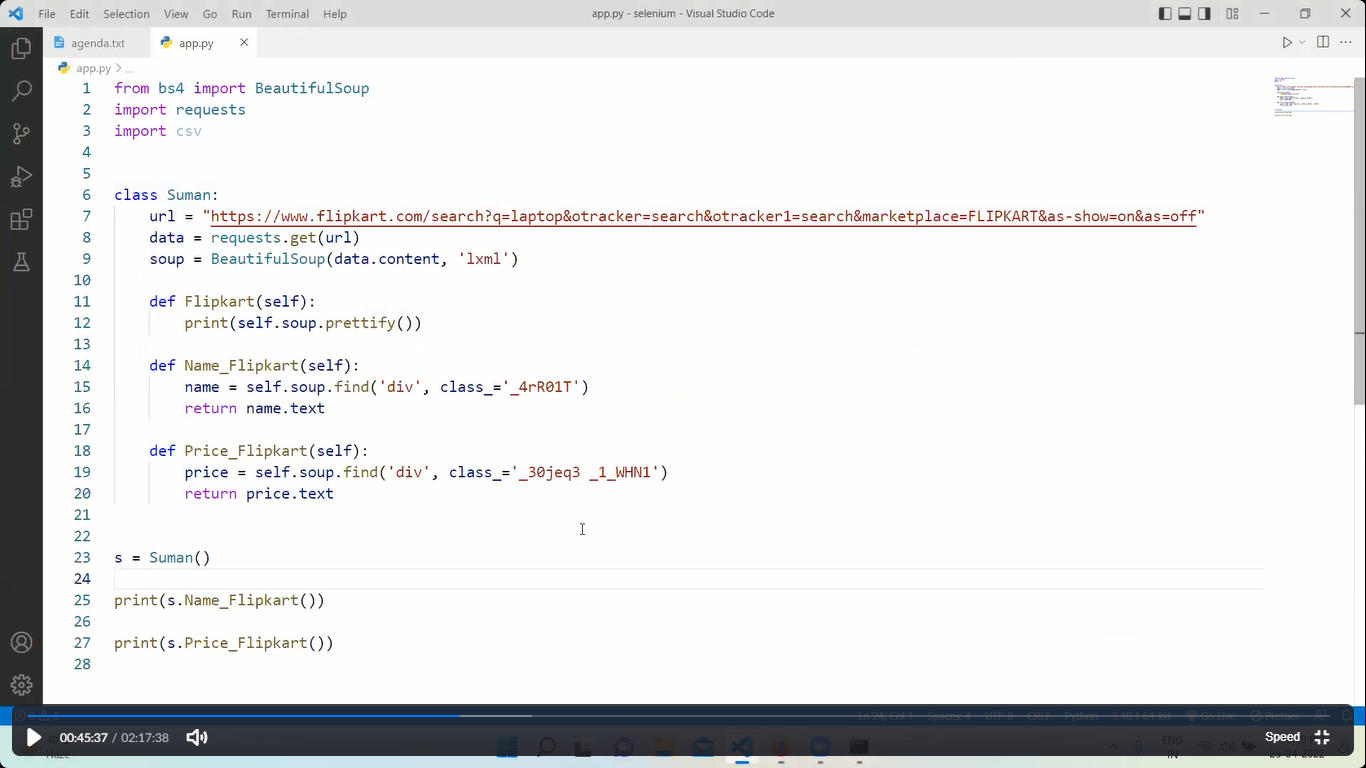
\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

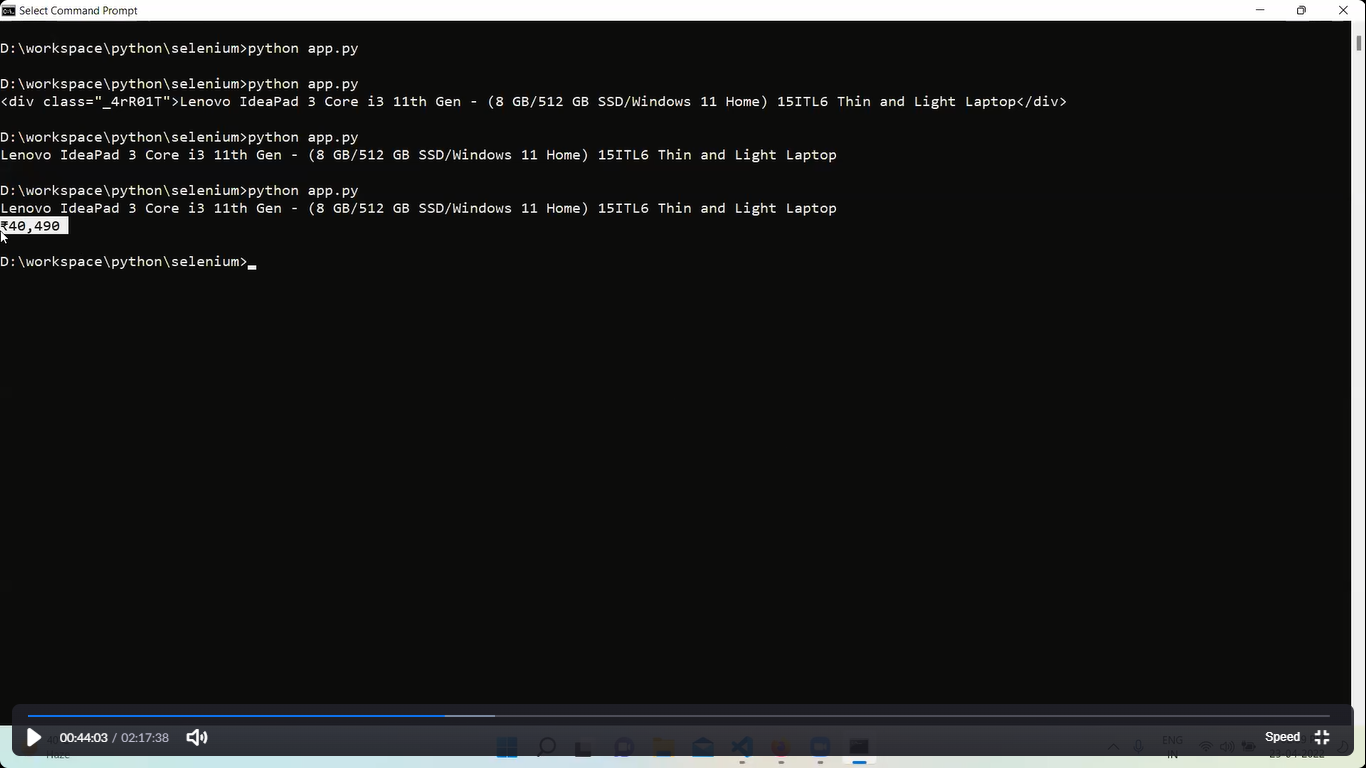






\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

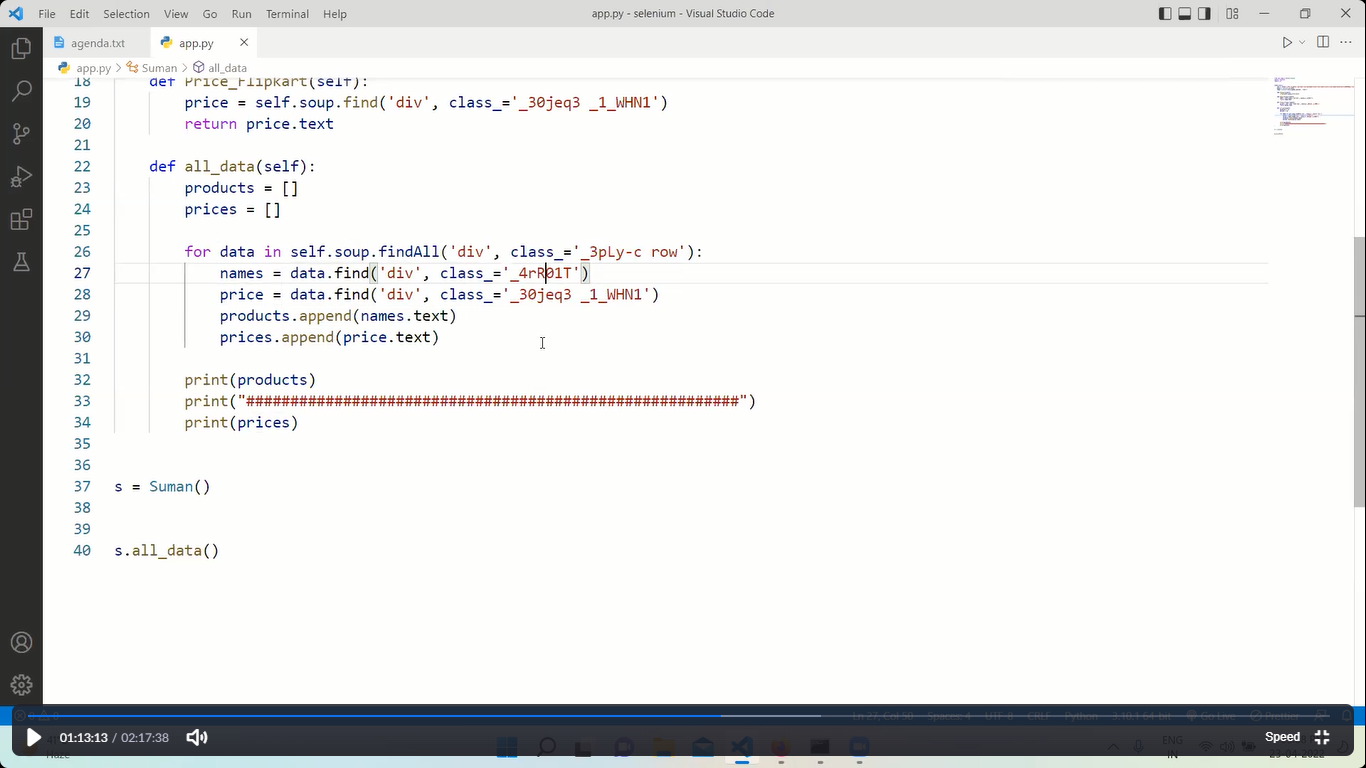




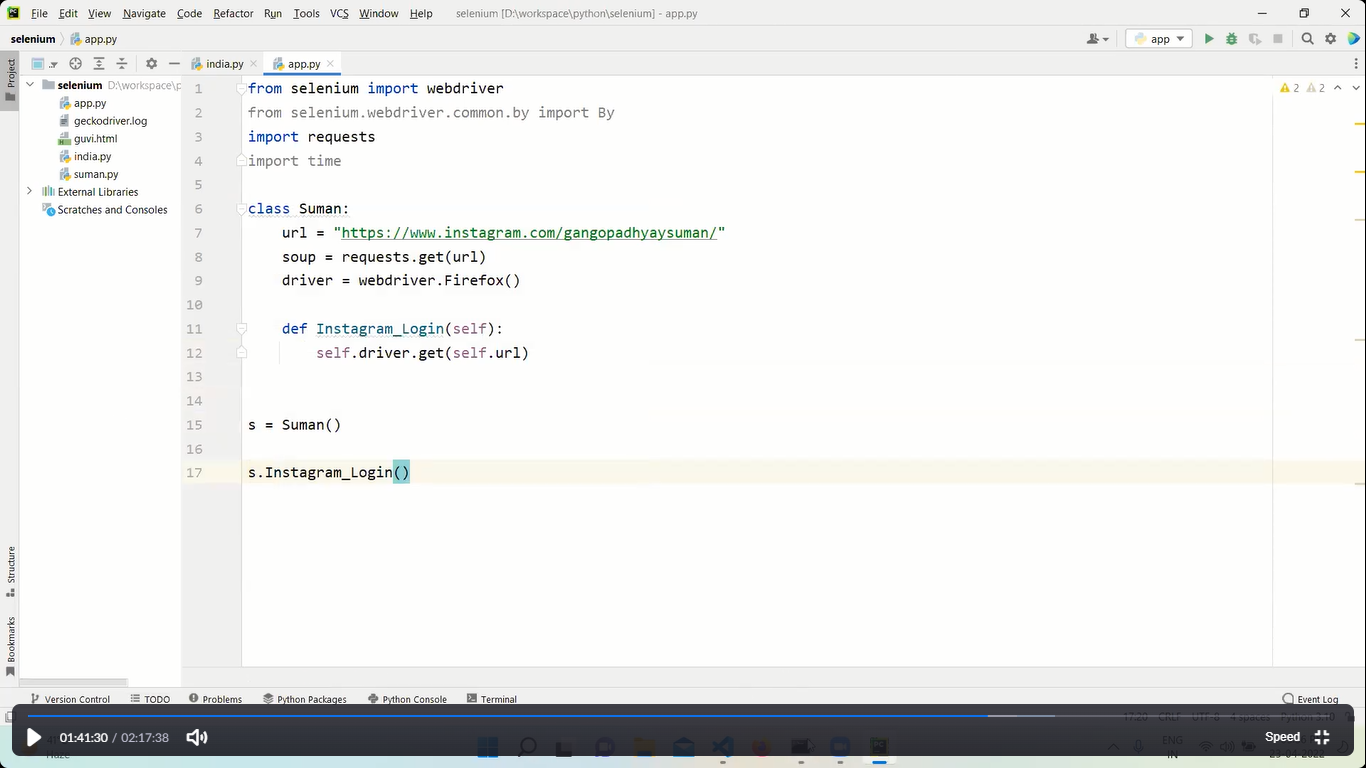
\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

FLIPKART----- to get all items names and their price

* + It’s a hit and trail method
  + All ways inspect the exact html element where the items info is located
  + Check for the generalized class i.e a specific class name which is being called again and again.
  + Create a method where we search the entire html using div class
  + Always check the class names or xpath correctly bcoz most of the webpages are dynamic

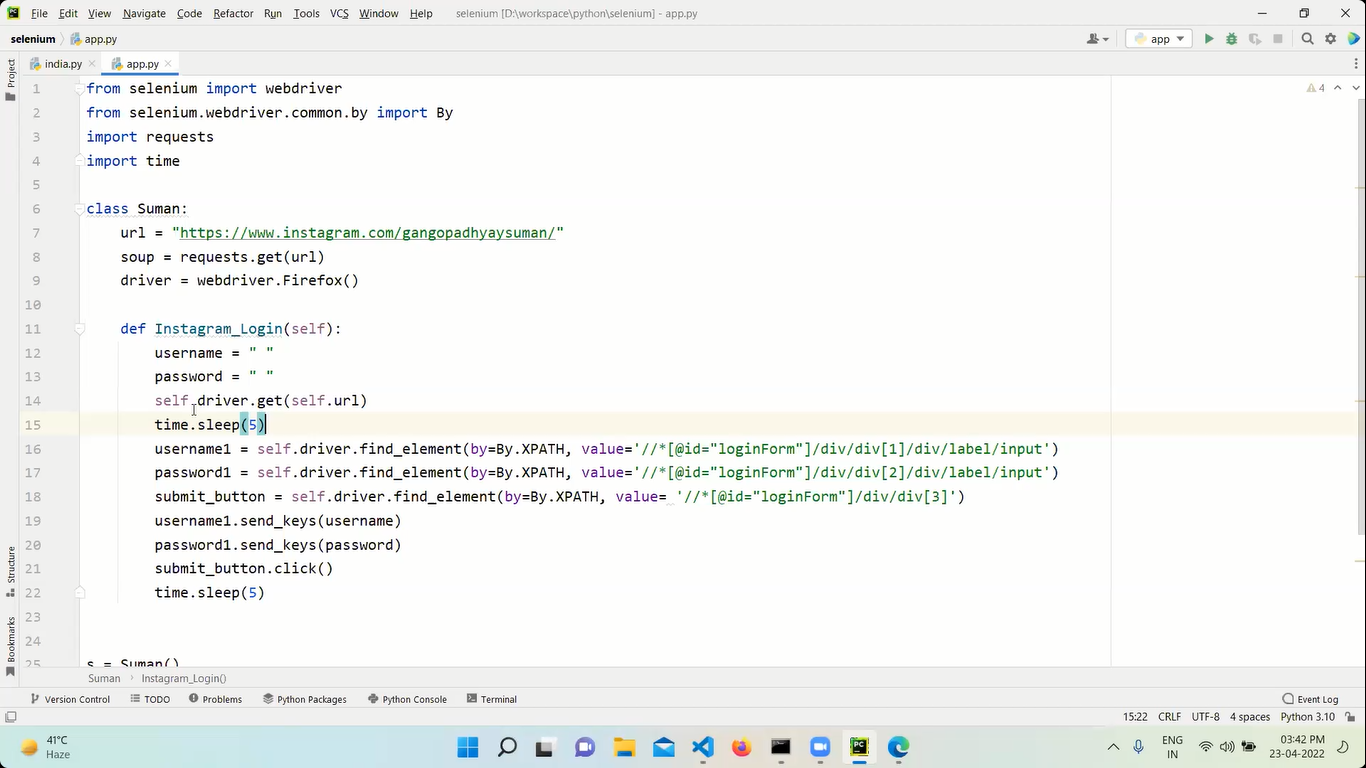


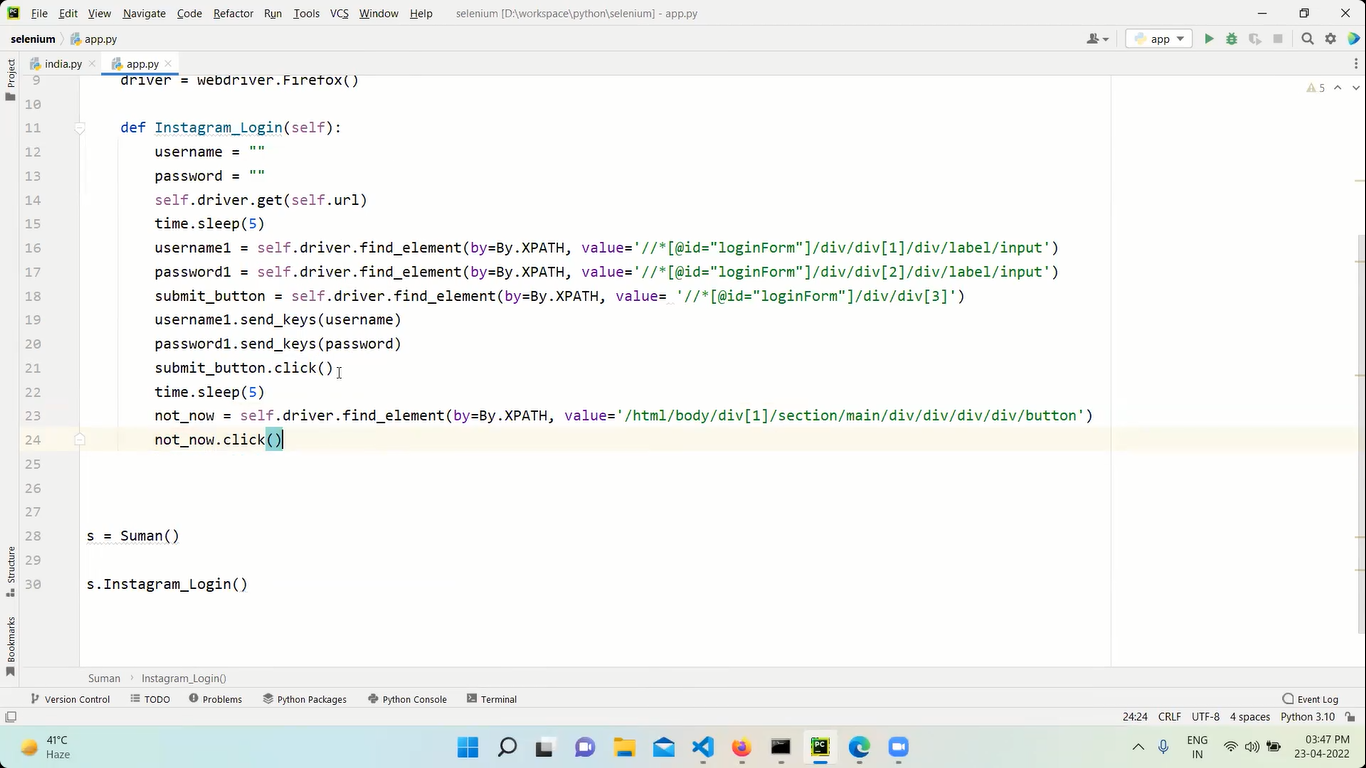
\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

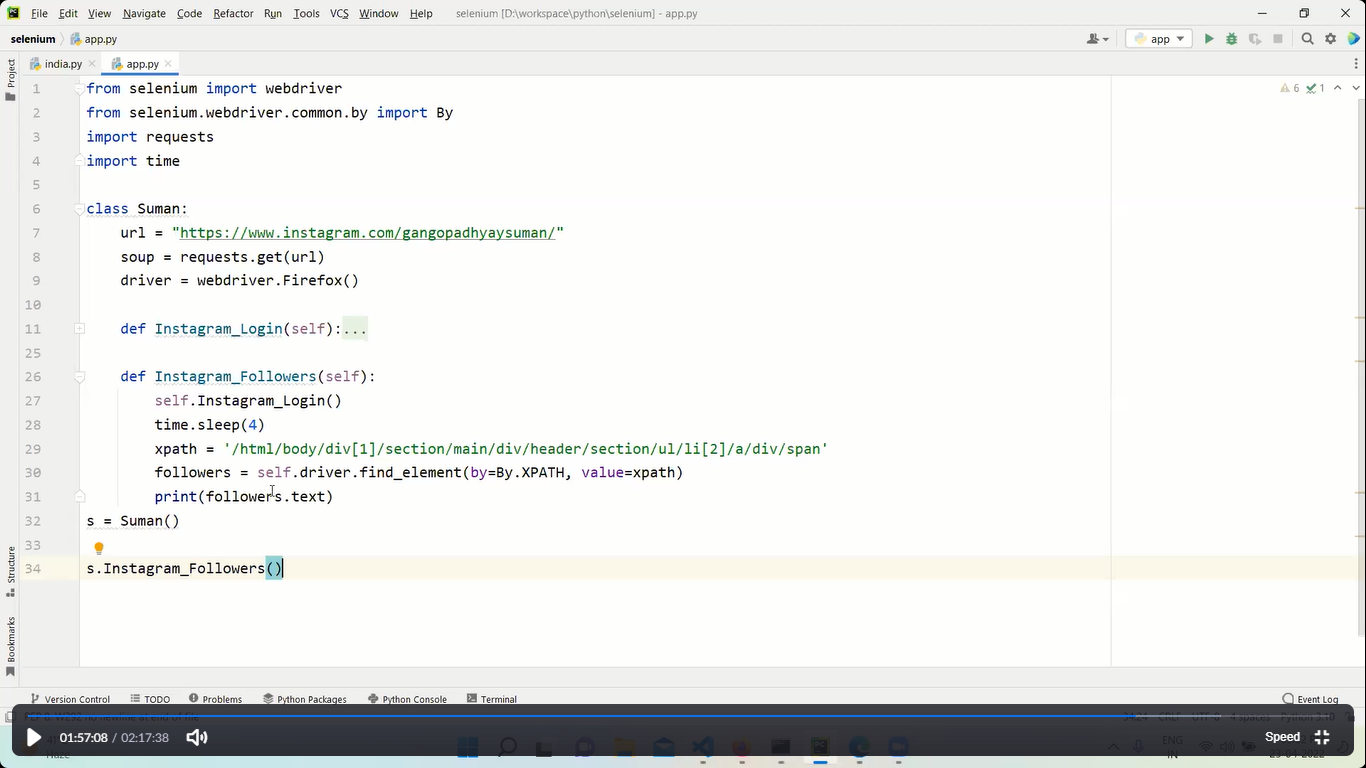


Instagram know that we are using automation in above program so it ask every time the login credientials

So to avoid it







The above program is to show followers count from instagram