# Extracting **Nykaa** Store data from its web page

Jupyter notebook is used for programming ease.

1. The required python packages are installed into the machine and also imported into the python program. The packages are :

* **pandas** – This package is used to store data and convert it into a CSV file.
* **selenium** – used to pull the web page contents into the program, as done by a browser.
* **bs4** – used to extract the desired data from a html web page.
* **html5lib** – used to parse html.

1. The contents of the desired web page are pulled using the code from the `selenium` package. The desired data in the web page is dynamically loaded by other scripts in the html, thus take time to get populated into the html. Another python built-in module `time` is used to delay the program, to use the fully populated web page.
2. The required data is then, extracted by using `bs4` code to conveniently carry out searching operations on the html to find the name and timings of the store, with regular expressions (`re` package) finding the address and phone number.
3. This data is then stored into a dataframe in their respective columns and the final collected data is exported to a CSV file.
4. To find the geographical coordinates of any address, `geopy` was first used but was unable to find the coordinates for most of the addresses. Google maps API could be easily used but requires a billing account to use it.