**Assignment Prompt: Retail Store**

**Location Scraper**

**Problem Statement:**

The task is to scrape the locations of your favourite retail brand in India and extract the following information:

* Store Name
* Address
* Timings
* Coordinates (Latitude/Longitude) (bonus)
* Phone Number

**Approach:**

I have used selenium for scraping the data. The first thing to do while scraping any data is to find a suitable webpage for the scraping of data. The retail store details I wanted to scrape about is MI so I had directly gone to its store locator webpage.

In the jupyter notebook, I imported useful libraries like selenium, numpy etc. After loading the automated window using selenium’s webdriver, we load the URL of the webpage we want to scrape by using webdriver’s function. After that I created a few empty lists to store the data.

The next step was to inspect the webpage and find the elements we want to scrape, I usually use ‘xpaths’ to locate or find an element or elements. After, writing the codes for each and every element we wanted to scrape the next part was to go the next page since a page only consisted 8 stores with details. So, I added a code which would click on the next page button at the end of the code. I used a for loop to repeat this process for next 50 pages. Afterwards, I formed a dataframe for the extracted data and then converted it into csv format and closed the driver.

**Challenges:**

* The first challenge that I came across was that for some elements like store timings, store mobile number the class wasn’t defined. The elements were present with a few tags only. I overcame it by taking the parent classes and then converting it into the element’s xpath.
* Another challenge was to scrape the coordinates for which I tried using geopy a python library for extracting geophysical information. But I couldn’t resolve the issue as the geocoder was returning None type value when feed with the address of the store I wanted to scrape coordinates of. This could be because of the inaccuracy of the scraped data or low functionality of the library.