CEDA Interview Round 2

submitted by: Satpreet Makhija

Task 1: Scraping Task

1. Retail prices for April 2021 for all 135 centers and 22 commodities

   Language used: Python
   Library used: Selenium

   For data scraping, the first step is to understand the DOM structure of the website. This is because once we understand the template, it becomes easier to scrape the required data as every page shares the same structure.

   For fcainfoweb.nic.in, the data gets generated dynamically on the page. I first wrote my scraper script to scrape the relevant data for one day. Once, I was successfully able to do so for one day, it was relatively easy to do so for all the other dates. I ran a for loop changing the date on each iteration and saved the data for each center in a global list. Ultimately, I wrote the data on the list in a CSV file for each day. Therefore, there are 30 .csv files corresponding to each date.

   data files naming convention: `dd/mm/yyyy`

   Data location: `retailPrices/retailPricesData`

2. Data on arrivals of commodities for one week for Delhi's mandis

   Language used: Python
   Library used: BeautifulSoup

   Time period chosen: 10 May 2021 - 16 May 2021

   Again, similar to the first part (1), I spent some time understanding the DOM structure of the website. Firstly, I scraped the commodity list from the homepage and stored it in a list with each element being a tuple (commodity_id, commodity_name). We need the commodity_id because the queries sent takes the commodity_id as input and not its name. Now, we run a for loop with each iteration changing the commodity id and save the tuple (commodity_name, commodity_value) in a global list using BeautifulSoup. Finally, we write the data from the list in CSV format to a .csv file.

   Data Location: `agmarket/agmarketData`