Thiago Schuck

Mrs. Binowski

CMP262

12 December 2023

Final Project Summary

At first, I wanted to manually scrape Amazon since it was the first website that came to mind, but I found it very difficult to navigate through the HTML structure so I decided an API would be easier. That wasn’t much easier, there wasn’t a website that provided what I was looking for so I went back to manually scraping the data. I was able to find the table with all the products and their information through the HTML structure, but when I started testing I would get error code 500 and error code 429. The first error code was an easy fix, all I needed to do was prove to Amazon that I wasn’t a bot which could be done by adding headers to the request. However, the second error code, 429, was because I was requesting from Amazon too many times. I tried refreshing the browser and the IDE, clearing my browsing cache, and changing the header but none of it worked; eventually, I would always get the error code 429. So I gave up Amazon and started looking for data on Walmart. The HTML structure was even more complicated but luckily there was an API that would give me a JSON of all the product information on the page. I set up a loop to get a JSON of all the sub-categories in the best-sellers page of Walmart, then I extracted the information I wanted to use and put it into a data frame using pandas. Then before moving on, I created a CSV file for the data so I wouldn’t have to send a request every time. I decided it would be better to have a separate notebook for analyzing the data. In that notebook, I read the CSV file into a data frame, then I went through it programmatically to answer the questions I had asked at the beginning of the project. Finally, I printed the answers to those questions and created chart to visualize the data.