Joachim Ndhokero (448004)

Ayoola Timothy Ayetigbo (444421)

Xolani Keith Mpala (444463)

Scraping the First 100 Cryptocurrencies in terms of Market Capitalization.

According to Wikipaedia, a cryptocurrency is a digital currency designed to work as a medium of exchange through a computer network that is not reliant on any central authority, such as a government or bank, to uphold or maintain it. It is tradable digital asset built on the block-chain technology that enables securing online transactions through encryption and authentication. Cryptocurrencies use encryption to authenticate and protect transactions, hence their name. There are currently over a thousand different cryptocurrencies in the world.

Over the last few years, cryptocurrency prices have risen and then fallen. Crypto marketplaces do not guarantee that an investor is completing a purchase or trade at the optimal price. As a result, many investors take advantage of this by using arbitrage to find the difference in price across several markets.

The first decentralized cryptocurrency was Bitcoin, which was first released as open-source software in 2009. Since the release of Bitcoin, many other cryptocurrencies have been created as more people become interested in using cryptocurrencies for transactions.

Our project is majorly based on performing web-scrapping operations on the website(https://www.coingecko.com) through the use of appropriate tools/frameworks. The tools to be utilized are:

- Beautiful Soup
- Scrapy
- Selenium

About coingecko.com:

It is a digital currency price and information data platform that helps its users quantitatively evaluate and rank their coins. It is a Singapore-based company that was founded in April, 2014 by TM Lee and Bobby Ong with the mission to democratize the access of crypto data and empower users with actionable insights.

CoinGecko is the world's largest independent cryptocurrency data aggregator with over 13,000+ different cryptoassets tracked across more than 600+ exchanges worldwide. CoinGecko provides a fundamental analysis of the crypto market, in addition to tracking price, volume and market capitalisation.

This is a website that shows cryptocurrency prices by their marketcap, 24hour volume traded, price chart for different periods. Forex and cryptocurrency traders use this website to find the prices of the cryptocurrencies they are interested in before filling in for a buy or sell. Besides cryptocurrencies it also shows crypto exchanges, NFT information, information and tutorials on learning crypto and other products.

Description that cuts across all the frameworks

Progress bar: element used to visualize the progression of the data scraping operation.

Charging bar: this helps in knowing what percentage of progress the execution of the program has reached.

*Catching error: w*hile making a request, if there is any bug, it prints out the exception. When the program breaks, it shows where the error occurred.

Elapsed time: printed helps in knowing how much time it has taken the program to actually scrap the website.

a) Beautiful soup

We start off by importing the needed libraries.

- pip install request helps in parsing the beautiful soup to fetch details one would want to scrap.
- Import time. This helps the program to make a specific number of requests in the process of scrapping. Therefore, it relaxes the program not to make so many requests such that its IP addresses is not blocked by the server from making extra requests.
- Headers. We make headers so that the browser can easily perform identifications and not mistake us for bots.

At this stage, we start using beautiful soup, so we parse in a response which is a string type of http while using the lxml. It then finds all the table body and under each table body we also find all table rows.

Then containers were created to which we shall append our scrapped data. Slugs are part of unique identify to get the 100 links of the cryptos. This is because the links scrapped are not the real URLs for the official websites of the cryptocurrencies but rather from coingecko.com. For the slugs we get their class and also their ***"href"*** from the inspection. We now go ahead and scrap the data, update our **COINS** dictionary with the scrapped data and append the data in the crypto-coins container earlier created.

We then loop through the slugs earlier created and append them to the base URL (URL for coingecko.com). Create a session to request for the headers and scrap the URLS, find the class for the name and URLS. Proceed to get the real data however the strip helps in removing any white spaces within the scrapped data.

At this point in time, append a tuple with the names and URL bar which helps us to scrap the data next in line according to the code (how the data is arranged on the website). We then create a csv file where we store our data, open the file, write the COINS to the data for the program to print it.

b) Scrapy

A custom spider was created sub-classing the spider in order to inherit the spider class which is from scrappy. In both scripts we create a function that opens the file for reading and writing. Create a new variable crypto body using the css method to get information about the table body and its table rows. Append the dictionary *COINS*. Loop through the table rows under the table to find out where the data actually lies, basically stripping through each element and appending it to a string as a new URL. Append the URLS and Coins to crypto slugs. We then print out the data (COINS) using yield print () so that the information printed out doesn't come out in a broken way.

c) Selenium

We started by importing the necessarily libraries for scraping. We added options so that we can comment out/ignore some errors while allowing the level 3 errors (fatal errors) to show while running the codes. We do this effectively by appending them to the driver.

We added the sleep function so that the program doesn't make a lot of requests in a short time which could lead to blocking of the IP address from scraping data from the website.

Create a function that finds the tag name for each table body (tbody) and under each tbody to retrieve the table rows. Create a slug which finds Xpath of elements and also their hrefs. Append the slug to the crypto URLS.

We go ahead and find the index of elements in question i.e., NAME, PRICE. Dump the json files such as bitcoin.org so we can get the official URLS. The slug is basically for coingecko.com but not for the official sites for the cryptocurrencies.

Inspect using the (Driver.get) to find the class which is under a div. (.text) to mean that we need the data to be scrapped to be in text format. for the URLS. We also had to get the tag.name for the names.

Create a temporary list of URLS, get the href and split it so that are displayed in separate lines. So, then we create a new array from what has been split. Then we append the temporary lists which is after also appended to the official websites.

Write the URLS to the data frame and create a csv file where it will be saved before we print the official URLS and names of the cryptos following the chronological order in the code.

Finally we scrap all the pages then will quit and close the program.

Technical Description of Output.

RANK: the position of the crypto in terms of market capitalization

SYMBOL: symbol for the cryptocurrency

PRICE: the prevailing price of the cryptocurrency. This normally changes from time to time. Unlike forex, cryptocurrencies are continuously traded on weekends.

MARKET CAPITALISATION: refers to the total value of all a company's shares of stock. The price of the crypto is the share price of the stock of that particular cryptocurrency.

For each scrapper, two csv files were created.

1. It scraped the name of the cryptocurrency, its sign and URL.
2. It scraped the name of the cryptocurrency, its sign, the price and market capitalization

Performance of Scrapers

Below are the recorded running time for the three(3) scrapers used:

- Beautiful Soup: 93.6188 seconds
- Selenium: 276.8974 seconds
- Scrapy: For csv: 0.614118 seconds
  For URL: 7.152398 seconds

From the elapse time recorded for each tools above, we can deduce that scrapy is the fastest, followed by Beautiful soup while Selenium happens to be the slowest among the three.

Work Description

Joachim Ndhokero - Selenium

Ayoola Timothy Ayetigbo – Beautiful Soup

Xolani Keith Mpala – Scrapy

The whole team – Description pdf

The whole program was done standardizing the words used in the codes such as coins, message such as print(‘---successfully created URLS---‘) so that one can easily follow up easily on the codes from selenium, beautiful soup and scrapy.

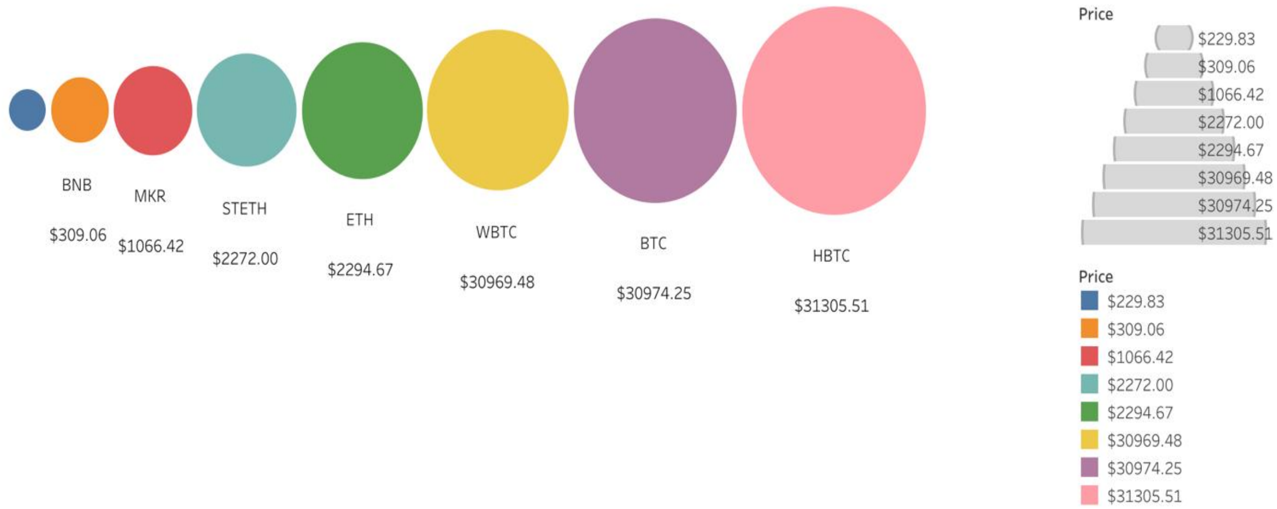
Elementary Data Analysis

Finding the correlation between cryptocurrency prices and cryptocurrency market capitalisation in R;

cor(beautifulsoup_crypto_currencies$Price, beautifulsoup_crypto_currencies$`MKT Cap`)

We deduce that there is a positive correlation between price and market capitalization of 0.5191389. However, it is evident that price doesn’t really affect the Market capitalization as some cryptocurrencies such as “Tether” has a price of $1 and has a bigger market capitalization than BNB which is $308.94.

This analysis should help cryptocurrency traders and those who want to engage in buying cryptocurrencies to mix up those with low prices but high portfolio. A cryptocurrency with a low value but with a high market capitalization means that people have strong feelings that it will appreciate more in the future as more and more people buy the shares. This is because most cryptos start at a very low price in the beginning (Initial Coin Offering).



Visualisation of the most priced cryptocurrencies as of May 2022

From the above visualized data, it is clearly shown that cryptos that are associated with Bitcoin such as HBTC (Huobi BTC), BTC(Bitcoin), WBTC (Wrapped Bitcoin) and BCH (Bitcoin Cash) have high prices. Therefore, it is evident that cryptos that are part of the Bitcoin blockchain have high chances of increasing in prices as a result of Bitcoin increasing in price.

Important Notice: the prices of the cryptocurrencies keep on changing from time to time so the prices displayed in the csv files might be different.