Assignment 1 – Subtask 1

Name: AASTHA A K VERMA

Entry Number: 2022CS11607

Course: COP290

The different file formats used were:

1. CSV
2. JSON
3. HTML
4. LaTeX
5. Excel
6. Feather
7. Parquet

|  |  |
| --- | --- |
| CIPLA | HINDALCO |
| SBIN | TCS |

The trend is quite similar in all the plots. Excel (a binary format) conversion took the longest, while conversion to CSV or JSON (text type formats) files took the least amount of time. Feather and Parquet (binary too) took low amounts of time as well.

Since all of the files store the same information for a particular symbol, they took the similar amount of space. The HTML file took the most space, probably because of the extraneous HTML tags. A similar thing can be said for the LaTeX file because of the extra syntax tags. CSV, Feather and Parquet take the lowest times, which is in correlation to their respective write times. HINDALCO was called for 5 years, and the rest for 2, which is reflected in the relative sizes. However, the trend between file types is constant.

The following method was used to record times and sizes:

start\_time = time.time()

    df\_.to\_csv(symbol + '.csv', index=False)

    times[0] = (float(time.time() - start\_time))

    file\_stats = os.stat(symbol + '.csv')

    sizes[0] = file\_stats.st\_size

This was done for each file type. The times were recorded using Python’s time library. After the file was exported, its size was recorded.

Then matplotlib was used to plot the times and sizes.