Sunil Sah

In this project, I began by importing the required libraries, including requests for HTTP requests, BeautifulSoup for HTML parsing, and pandas for data manipulation. Next, I fetched the IMDb Top Rated Movies page using its URL and added a user-agent header to simulate a real browser request. After receiving the webpage content, I parsed it with BeautifulSoup to locate the structured data containing movie information. Then, I extracted JSON-structured data by locating the script tag with type "application/ld+json" and converting it into a Python dictionary. I initialized lists for movie details like titles, durations, and ratings, and looped through each movie's data, using 'N/A' as a fallback for any missing fields.

The extracted data was stored in a pandas DataFrame, making it easier to manipulate and analyze. To save this data, I exported the DataFrame to a CSV file. I then began data cleaning by identifying NaN and null values and checking for duplicates, which provided a clearer view of potential data inconsistencies. I converted the ISO 8601 time format in the "Duration" column into minutes for easier analysis. Finally, I began data visualization by setting up a box plot to visualize movie duration distributions.