webscraping_phase1

July 19, 2023

```python
[10]: #install and import required libraries.
      from bs4 import BeautifulSoup
      import requests
      import pandas as pd
      import csv
      import numpy as np
      from time import sleep
      from random import randint
```

```python
[11]: # information to be in english language
      header={"Accept=language":"en-US,en;q=0.5"}
```

```python
[12]: # extracting information of pages
      pages=np.arange(1,1749,50)
```

```python
[13]: # Create empty lists for the data
      Sno = []
      Movie_Name = []
      Director_Name = []
      Director1 = []
      Director2 = []
      Director3 = []
      Duration = []
      Year = []
      Ratings = []
      Metascore = []
```

```python
[14]: #creating a for loop for required pages
      for page in pages:
          #getting required url
          page=requests.get("https://www.imdb.com/search/title/?
      →title_type=feature&num_votes=25000,&genres=action&sort=user_rating,desc")
          soup = BeautifulSoup(page.text, 'html.parser')
          # Find movie data on the current page
          movie_data = soup.find_all('div', attrs={'class': 'lister-item-content'})
          sleep(randint(2,8))
          # Iterate over each movie on the current page
```

```python
for movie in movie_data:
    # Extract serial number
    Sno.append(len(Sno) + 1)

    # Extract movie name
    Movie_Name.append(movie.h3.a.text)

    # Extract director names
    directors = movie.find_all('p')[2].find_all('a')
    director_names = [director.text for director in directors][:-4]
    Director_Name.append(director_names)
     # Assign values to Director1, Director2, and Director3
    director_count = len(director_names)
    Director1.append(director_names[0] if director_count > 0 else "-")
    Director2.append(director_names[1] if director_count > 1 else "-")
    Director3.append(director_names[2] if director_count > 2 else "-")

    # Extract duration
    Duration.append(movie.find('span', attrs={'class': 'runtime'}).text)

    # Extract year
    Year.append(movie.find('span', attrs={'class': 'lister-item-year'}).
 ↪text.strip('()'))

    # Extract rating
    Ratings.append(movie.strong.text)

    # Extract metascore
    metascore = movie.find('span', attrs={'class': 'metascore'})
    Metascore.append(metascore.text if metascore else "-")
```

```python
[15]: df = pd.DataFrame({
     'Sno': Sno,
     'Movie Name': Movie_Name,
     'Director1': Director1,
     'Director2': Director2,
     'Director3': Director3,
     'Duration': Duration,
     'Year': Year,
     'Ratings': Ratings,
     'Metascore': Metascore
})
```

```python
[16]: df
```

```
[16]:    Sno                         Movie Name  \
     0      1                   The Dark Knight
```

```
1        2           The Lord of the Rings: The Return of the King
2        3                        Spider-Man: Across the Spider-Verse
3        4                                                  Inception
4        5     The Lord of the Rings: The Fellowship of the Ring
…        …                                                        …
1745   1746                      Demon Slayer the Movie: Mugen Train
1746   1747                                     Gangs of Wasseypur
1747   1748                           Baahubali 2: The Conclusion
1748   1749                                                  Yojimbo
1749   1750                              Uri: The Surgical Strike

                 Director1       Director2              Director3  Duration  Year  \
0       Christopher Nolan              -                      -    152 min  2008
1          Peter Jackson              -                      -    201 min  2003
2     Joaquim Dos Santos   Kemp Powers   Justin K. Thompson  140 min  2023
3       Christopher Nolan              -                      -    148 min  2010
4          Peter Jackson              -                      -    178 min  2001
…                     …             …                     …       …     …
1745       Haruo Sotozaki              -                      -    117 min  2020
1746       Anurag Kashyap              -                      -    321 min  2012
1747        S.S. Rajamouli              -                      -    167 min  2017
1748        Akira Kurosawa              -                      -    110 min  1961
1749          Aditya Dhar              -                      -    138 min  2019

        Ratings   Metascore
0          9.0   84
1          9.0   94
2          8.9   86
3          8.8   74
4          8.8   92
…           …          …
1745       8.2   72
1746       8.2   89
1747       8.2          -
1748       8.2   93
1749       8.2          -

[1750 rows x 9 columns]
```

```
[30]: # saving file to csv
      df.to_csv('IMBD_Data1.csv', index=False)
```

process Table 2

```
[31]: import csv
      import requests
      from bs4 import BeautifulSoup
```

```python
# Send a GET request to the IMDb URL
url = "https://www.imdb.com/search/title/?
 ↪genres=action&sort=user_rating,desc&title_type=feature&num_votes=25000,&pf_rd_m=A2FGELUUNOQ
response = requests.get(url)

# Create a BeautifulSoup object
soup = BeautifulSoup(response.text, "html.parser")

# Extract the data
data = []
movie_elements = soup.find_all("div", class_="lister-item-content")
for movie_element in movie_elements:
    movie_name = movie_element.find("h3").find("a").text
    stars_element = movie_element.find("p", class_="").find_all("a")
    stars = [star.text for star in stars_element]

    votes_element = movie_element.find("p", class_="sort-num_votes-visible").
 ↪find("span", attrs={"name": "nv"})
    votes = votes_element.text if votes_element else None

    genre_element = movie_element.find("span", class_="genre")
    genre = [g.strip() for g in genre_element.text.strip().split(",")] if␣
 ↪genre_element else []

    gross_collection_element = movie_element.find("p",␣
 ↪class_="sort-num_votes-visible").find_next_sibling("span")
    gross_collection = gross_collection_element.text.strip() if␣
 ↪gross_collection_element else None

    popularity_element = movie_element.find("span", class_="lister-item-rank")
    popularity = popularity_element.text.strip(".") if popularity_element else␣
 ↪None

    certification_element = movie_element.find("span", class_="certificate")
    certification = certification_element.text.strip() if certification_element␣
 ↪else None

    # Bifurcate stars into subfields star1, star2, star3, star4
    star1 = stars[0] if len(stars) >= 1 else None
    star2 = stars[1] if len(stars) >= 2 else None
    star3 = stars[2] if len(stars) >= 3 else None
    star4 = stars[3] if len(stars) >= 4 else None

    # Bifurcate genre into subfields genre1, genre2, genre3
    genre1 = genre[0] if len(genre) >= 1 else None
```

```python
        genre2 = genre[1] if len(genre) >= 2 else None
        genre3 = genre[2] if len(genre) >= 3 else None

        data.append([movie_name, star1, star2, star3, star4, votes, genre1, genre2,
  →genre3, gross_collection, popularity, certification])

    # Store the data in the second CSV file
    filename = "IMDB_Data_2.csv"
    with open(filename, "w", newline="", encoding="utf-8") as csvfile:
        writer = csv.writer(csvfile)
        writer.writerow(["Movie Name", "Star1", "Star2", "Star3", "Star4", "Votes",
  →"Genre1", "Genre2", "Genre3", "Gross Collection", "Popularity",
  →"Certification"])
        writer.writerows(data)

    print("Data for the second CSV file stored in", filename)
```

Data for the second CSV file stored in IMDB_Data_2.csv

```python
[32]: filename = "IMDB_Data_2.csv"
      data = []

      with open(filename, "r", newline="", encoding="utf-8") as csvfile:
          reader = csv.reader(csvfile)
          header = next(reader)  # Skip the header row

          for row in reader:
              data.append(row)

      # Print the data
      for row in data:
          print(row)
```

```
['The Dark Knight', 'Christopher Nolan', 'Christian Bale', 'Heath Ledger',
'Aaron Eckhart', '2,741,127', 'Action', 'Crime', 'Drama', '', '', 'PG-13']
['The Lord of the Rings: The Return of the King', 'Peter Jackson', 'Elijah
Wood', 'Viggo Mortensen', 'Ian McKellen', '1,899,013', 'Action', 'Adventure',
'Drama', '', '', 'PG-13']
['Spider-Man: Across the Spider-Verse', 'Joaquim Dos Santos', 'Kemp Powers',
'Justin K. Thompson', 'Shameik Moore', '185,491', 'Animation', 'Action',
'Adventure', '', '', 'PG']
['Inception', 'Christopher Nolan', 'Leonardo DiCaprio', 'Joseph Gordon-Levitt',
'Elliot Page', '2,432,436', 'Action', 'Adventure', 'Sci-Fi', '', '', 'PG-13']
['The Lord of the Rings: The Fellowship of the Ring', 'Peter Jackson', 'Elijah
Wood', 'Ian McKellen', 'Orlando Bloom', '1,927,425', 'Action', 'Adventure',
'Drama', '', '', 'PG-13']
['The Lord of the Rings: The Two Towers', 'Peter Jackson', 'Elijah Wood', 'Ian
```

McKellen', 'Viggo Mortensen', '1,713,772', 'Action', 'Adventure', 'Drama', '',
'', 'PG-13']
['The Matrix', 'Lana Wachowski', 'Lilly Wachowski', 'Keanu Reeves', 'Laurence
Fishburne', '1,971,133', 'Action', 'Sci-Fi', '', '', '', 'R']
['Star Wars: Episode V - The Empire Strikes Back', 'Irvin Kershner', 'Mark
Hamill', 'Harrison Ford', 'Carrie Fisher', '1,329,834', 'Action', 'Adventure',
'Fantasy', '', '', 'PG']
['Soorarai Pottru', 'Sudha Kongara', 'Suriya', 'Paresh Rawal', 'Aparna
Balamurali', '120,033', 'Action', 'Drama', '', '', '', 'TV-MA']
['Terminator 2: Judgment Day', 'James Cameron', 'Arnold Schwarzenegger', 'Linda
Hamilton', 'Edward Furlong', '1,131,068', 'Action', 'Sci-Fi', '', '', '', 'R']
['Star Wars: Episode IV - A New Hope', 'George Lucas', 'Mark Hamill', 'Harrison
Ford', 'Carrie Fisher', '1,402,082', 'Action', 'Adventure', 'Fantasy', '', '',
'PG']
['Seven Samurai', 'Akira Kurosawa', 'Toshirô Mifune', 'Takashi Shimura', 'Keiko
Tsushima', '354,942', 'Action', 'Drama', '', '', '', 'Not Rated']
['Harakiri', 'Masaki Kobayashi', 'Tatsuya Nakadai', 'Akira Ishihama', 'Shima
Iwashita', '62,143', 'Action', 'Drama', 'Mystery', '', '', 'Not Rated']
['Sita Ramam', 'Hanu Raghavapudi', 'Dulquer Salmaan', 'Mrunal Thakur', 'Rashmika
Mandanna', '61,432', 'Action', 'Drama', 'Mystery', '', '', '']
['Gladiator', 'Ridley Scott', 'Russell Crowe', 'Joaquin Phoenix', 'Connie
Nielsen', '1,546,170', 'Action', 'Adventure', 'Drama', '', '', 'R']
['Léon: The Professional', 'Luc Besson', 'Jean Reno', 'Gary Oldman', 'Natalie
Portman', '1,197,398', 'Action', 'Crime', 'Drama', '', '', 'R']
['Raiders of the Lost Ark', 'Steven Spielberg', 'Harrison Ford', 'Karen Allen',
'Paul Freeman', '1,003,306', 'Action', 'Adventure', '', '', '', 'PG']
['Spider-Man: Into the Spider-Verse', 'Bob Persichetti', 'Peter Ramsey', 'Rodney
Rothman', 'Shameik Moore', '609,082', 'Animation', 'Action', 'Adventure', '',
'', 'PG']
['Avengers: Endgame', 'Anthony Russo', 'Joe Russo', 'Robert Downey Jr.', 'Chris
Evans', '1,194,734', 'Action', 'Adventure', 'Drama', '', '', 'PG-13']
['Aliens', 'James Cameron', 'Sigourney Weaver', 'Michael Biehn', 'Carrie Henn',
'737,021', 'Action', 'Adventure', 'Sci-Fi', '', '', 'R']
['Oldboy', 'Park Chan-wook', 'Choi Min-sik', 'Yoo Ji-tae', 'Kang Hye-jeong',
'601,507', 'Action', 'Drama', 'Mystery', '', '', 'R']
['The Dark Knight Rises', 'Christopher Nolan', 'Christian Bale', 'Tom Hardy',
'Anne Hathaway', '1,754,277', 'Action', 'Drama', 'Thriller', '', '', 'PG-13']
['Avengers: Infinity War', 'Anthony Russo', 'Joe Russo', 'Robert Downey Jr.',
'Chris Hemsworth', '1,137,081', 'Action', 'Adventure', 'Sci-Fi', '', '',
'PG-13']
['Kaithi', 'Lokesh Kanagaraj', 'Karthi', 'Narain', 'Arjun Das', '36,785',
'Action', 'Adventure', 'Crime', '', '', 'Not Rated']
['Asuran', 'Vetrimaaran', 'Dhanush', 'Manju Warrier', 'Prakash Raj', '31,537',
'Action', 'Drama', '', '', '', 'Not Rated']
['Top Gun: Maverick', 'Joseph Kosinski', 'Tom Cruise', 'Jennifer Connelly',
'Miles Teller', '609,552', 'Action', 'Drama', '', '', '', 'PG-13']
['Heat', 'Michael Mann', 'Al Pacino', 'Robert De Niro', 'Val Kilmer', '683,149',
'Action', 'Crime', 'Drama', '', '', 'R']

['Star Wars: Episode VI - Return of the Jedi', 'Richard Marquand', 'Mark Hamill', 'Harrison Ford', 'Carrie Fisher', '1,084,286', 'Action', 'Adventure', 'Fantasy', '', '', 'PG']
['Princess Mononoke', 'Hayao Miyazaki', 'Yôji Matsuda', 'Yuriko Ishida', 'Yûko Tanaka', '410,355', 'Animation', 'Action', 'Adventure', '', '', 'PG-13']
['North by Northwest', 'Alfred Hitchcock', 'Cary Grant', 'Eva Marie Saint', 'James Mason', '336,420', 'Action', 'Adventure', 'Mystery', '', '', 'Approved']
['Dangal', 'Nitesh Tiwari', 'Aamir Khan', 'Sakshi Tanwar', 'Fatima Sana Shaikh', '199,079', 'Action', 'Biography', 'Drama', '', '', 'Not Rated']
['Kantara', 'Rishab Shetty', 'Rishab Shetty', 'Kishore Kumar G.', 'Achyuth Kumar', '96,853', 'Action', 'Adventure', 'Drama', '', '', 'Not Rated']
['K.G.F: Chapter 2', 'Prashanth Neel', 'Yash', 'Sanjay Dutt', 'Raveena Tandon', '140,977', 'Action', 'Crime', 'Drama', '', '', 'Not Rated']
['Vikram', 'Lokesh Kanagaraj', 'Kamal Haasan', 'Vijay Sethupathi', 'Fahadh Faasil', '64,776', 'Action', 'Crime', 'Thriller', '', '', '']
['Shershaah', 'Vishnuvardhan', 'Sidharth Malhotra', 'Kiara Advani', 'Shiv Panditt', '125,557', 'Action', 'Biography', 'Drama', '', '', 'TV-14']
['Raatchasan', 'Ram Kumar', 'Vishnu Vishal', 'Amala Paul', 'Radha Ravi', '47,479', 'Action', 'Crime', 'Drama', '', '', 'Not Rated']
['Indiana Jones and the Last Crusade', 'Steven Spielberg', 'Harrison Ford', 'Sean Connery', 'Alison Doody', '783,759', 'Action', 'Adventure', '', '', '', 'PG-13']
['Spider-Man: No Way Home', 'Jon Watts', 'Tom Holland', 'Zendaya', 'Benedict Cumberbatch', '813,538', 'Action', 'Adventure', 'Fantasy', '', '', 'PG-13']
['Jurassic Park', 'Steven Spielberg', 'Sam Neill', 'Laura Dern', 'Jeff Goldblum', '1,022,937', 'Action', 'Adventure', 'Sci-Fi', '', '', 'PG-13']
['Batman Begins', 'Christopher Nolan', 'Christian Bale', 'Michael Caine', 'Ken Watanabe', '1,514,502', 'Action', 'Crime', 'Drama', '', '', 'PG-13']
['1917', 'Sam Mendes', 'Dean-Charles Chapman', 'George MacKay', 'Daniel Mays', '629,307', 'Action', 'Drama', 'War', '', '', 'R']
['Die Hard', 'John McTiernan', 'Bruce Willis', 'Alan Rickman', 'Bonnie Bedelia', '906,120', 'Action', 'Thriller', '', '', '', 'R']
['Kill Bill: Vol. 1', 'Quentin Tarantino', 'Uma Thurman', 'David Carradine', 'Daryl Hannah', '1,147,660', 'Action', 'Crime', 'Thriller', '', '', 'R']
['V for Vendetta', 'James McTeigue', 'Hugo Weaving', 'Natalie Portman', 'Rupert Graves', '1,145,964', 'Action', 'Drama', 'Sci-Fi', '', '', 'R']
['Ran', 'Akira Kurosawa', 'Tatsuya Nakadai', 'Akira Terao', 'Jinpachi Nezu', '130,308', 'Action', 'Drama', 'War', '', '', 'R']
['Demon Slayer the Movie: Mugen Train', 'Haruo Sotozaki', 'Natsuki Hanae', 'Akari Kitô', 'Yoshitsugu Matsuoka', '66,095', 'Animation', 'Action', 'Adventure', '', '', 'R']
['Gangs of Wasseypur', 'Anurag Kashyap', 'Manoj Bajpayee', 'Nawazuddin Siddiqui', 'Tigmanshu Dhulia', '99,904', 'Action', 'Comedy', 'Crime', '', '', 'Not Rated']
['Baahubali 2: The Conclusion', 'S.S. Rajamouli', 'Prabhas', 'Rana Daggubati', 'Anushka Shetty', '108,086', 'Action', 'Drama', '', '', '', 'Not Rated']
['Yojimbo', 'Akira Kurosawa', 'Toshirô Mifune', 'Eijirô Tôno', 'Tatsuya Nakadai', '126,724', 'Action', 'Drama', 'Thriller', '', '', 'Not Rated']

```
['Uri: The Surgical Strike', 'Aditya Dhar', 'Vicky Kaushal', 'Paresh Rawal',
'Mohit Raina', '68,308', 'Action', 'Drama', 'History', '', '', 'Not Rated']
```

[33]:
```python
df2 = pd.DataFrame(data, columns=header)
```

[34]:
```python
df2
```

[34]:

	Movie Name	Star1 \
0	The Dark Knight	Christopher Nolan
1	The Lord of the Rings: The Return of the King	Peter Jackson
2	Spider-Man: Across the Spider-Verse	Joaquim Dos Santos
3	Inception	Christopher Nolan
4	The Lord of the Rings: The Fellowship of the Ring	Peter Jackson
5	The Lord of the Rings: The Two Towers	Peter Jackson
6	The Matrix	Lana Wachowski
7	Star Wars: Episode V - The Empire Strikes Back	Irvin Kershner
8	Soorarai Pottru	Sudha Kongara
9	Terminator 2: Judgment Day	James Cameron
10	Star Wars: Episode IV - A New Hope	George Lucas
11	Seven Samurai	Akira Kurosawa
12	Harakiri	Masaki Kobayashi
13	Sita Ramam	Hanu Raghavapudi
14	Gladiator	Ridley Scott
15	Léon: The Professional	Luc Besson
16	Raiders of the Lost Ark	Steven Spielberg
17	Spider-Man: Into the Spider-Verse	Bob Persichetti
18	Avengers: Endgame	Anthony Russo
19	Aliens	James Cameron
20	Oldboy	Park Chan-wook
21	The Dark Knight Rises	Christopher Nolan
22	Avengers: Infinity War	Anthony Russo
23	Kaithi	Lokesh Kanagaraj
24	Asuran	Vetrimaaran
25	Top Gun: Maverick	Joseph Kosinski
26	Heat	Michael Mann
27	Star Wars: Episode VI - Return of the Jedi	Richard Marquand
28	Princess Mononoke	Hayao Miyazaki
29	North by Northwest	Alfred Hitchcock
30	Dangal	Nitesh Tiwari
31	Kantara	Rishab Shetty
32	K.G.F: Chapter 2	Prashanth Neel
33	Vikram	Lokesh Kanagaraj
34	Shershaah	Vishnuvardhan
35	Raatchasan	Ram Kumar
36	Indiana Jones and the Last Crusade	Steven Spielberg
37	Spider-Man: No Way Home	Jon Watts
38	Jurassic Park	Steven Spielberg

39	Batman Begins	Christopher Nolan
40	1917	Sam Mendes
41	Die Hard	John McTiernan
42	Kill Bill: Vol. 1	Quentin Tarantino
43	V for Vendetta	James McTeigue
44	Ran	Akira Kurosawa
45	Demon Slayer the Movie: Mugen Train	Haruo Sotozaki
46	Gangs of Wasseypur	Anurag Kashyap
47	Baahubali 2: The Conclusion	S.S. Rajamouli
48	Yojimbo	Akira Kurosawa
49	Uri: The Surgical Strike	Aditya Dhar

	Star2	Star3	Star4 \\
0	Christian Bale	Heath Ledger	Aaron Eckhart
1	Elijah Wood	Viggo Mortensen	Ian McKellen
2	Kemp Powers	Justin K. Thompson	Shameik Moore
3	Leonardo DiCaprio	Joseph Gordon-Levitt	Elliot Page
4	Elijah Wood	Ian McKellen	Orlando Bloom
5	Elijah Wood	Ian McKellen	Viggo Mortensen
6	Lilly Wachowski	Keanu Reeves	Laurence Fishburne
7	Mark Hamill	Harrison Ford	Carrie Fisher
8	Suriya	Paresh Rawal	Aparna Balamurali
9	Arnold Schwarzenegger	Linda Hamilton	Edward Furlong
10	Mark Hamill	Harrison Ford	Carrie Fisher
11	Toshirô Mifune	Takashi Shimura	Keiko Tsushima
12	Tatsuya Nakadai	Akira Ishihama	Shima Iwashita
13	Dulquer Salmaan	Mrunal Thakur	Rashmika Mandanna
14	Russell Crowe	Joaquin Phoenix	Connie Nielsen
15	Jean Reno	Gary Oldman	Natalie Portman
16	Harrison Ford	Karen Allen	Paul Freeman
17	Peter Ramsey	Rodney Rothman	Shameik Moore
18	Joe Russo	Robert Downey Jr.	Chris Evans
19	Sigourney Weaver	Michael Biehn	Carrie Henn
20	Choi Min-sik	Yoo Ji-tae	Kang Hye-jeong
21	Christian Bale	Tom Hardy	Anne Hathaway
22	Joe Russo	Robert Downey Jr.	Chris Hemsworth
23	Karthi	Narain	Arjun Das
24	Dhanush	Manju Warrier	Prakash Raj
25	Tom Cruise	Jennifer Connelly	Miles Teller
26	Al Pacino	Robert De Niro	Val Kilmer
27	Mark Hamill	Harrison Ford	Carrie Fisher
28	Yôji Matsuda	Yuriko Ishida	Yûko Tanaka
29	Cary Grant	Eva Marie Saint	James Mason
30	Aamir Khan	Sakshi Tanwar	Fatima Sana Shaikh
31	Rishab Shetty	Kishore Kumar G.	Achyuth Kumar
32	Yash	Sanjay Dutt	Raveena Tandon
33	Kamal Haasan	Vijay Sethupathi	Fahadh Faasil

```
34      Sidharth Malhotra         Kiara Advani          Shiv Panditt
35         Vishnu Vishal           Amala Paul            Radha Ravi
36        Harrison Ford          Sean Connery          Alison Doody
37          Tom Holland               Zendaya  Benedict Cumberbatch
38            Sam Neill            Laura Dern          Jeff Goldblum
39        Christian Bale         Michael Caine          Ken Watanabe
40  Dean-Charles Chapman        George MacKay           Daniel Mays
41          Bruce Willis          Alan Rickman         Bonnie Bedelia
42          Uma Thurman        David Carradine          Daryl Hannah
43          Hugo Weaving        Natalie Portman         Rupert Graves
44       Tatsuya Nakadai          Akira Terao         Jinpachi Nezu
45         Natsuki Hanae           Akari Kitô  Yoshitsugu Matsuoka
46        Manoj Bajpayee  Nawazuddin Siddiqui   Tigmanshu Dhulia
47              Prabhas       Rana Daggubati       Anushka Shetty
48       Toshirô Mifune          Eijirô Tôno       Tatsuya Nakadai
49         Vicky Kaushal          Paresh Rawal           Mohit Raina


        Votes      Genre1    Genre2      Genre3 Gross Collection Popularity  \
0   2,741,127      Action     Crime       Drama
1   1,899,013      Action  Adventure      Drama
2     185,491   Animation     Action   Adventure
3   2,432,436      Action  Adventure     Sci-Fi
4   1,927,425      Action  Adventure      Drama
5   1,713,772      Action  Adventure      Drama
6   1,971,133      Action     Sci-Fi
7   1,329,834      Action  Adventure     Fantasy
8     120,033      Action      Drama
9   1,131,068      Action     Sci-Fi
10  1,402,082      Action  Adventure     Fantasy
11    354,942      Action      Drama
12     62,143      Action      Drama     Mystery
13     61,432      Action      Drama     Mystery
14  1,546,170      Action  Adventure      Drama
15  1,197,398      Action     Crime       Drama
16  1,003,306      Action  Adventure
17    609,082   Animation     Action   Adventure
18  1,194,734      Action  Adventure      Drama
19    737,021      Action  Adventure     Sci-Fi
20    601,507      Action      Drama     Mystery
21  1,754,277      Action      Drama    Thriller
22  1,137,081      Action  Adventure     Sci-Fi
23     36,785      Action  Adventure      Crime
24     31,537      Action      Drama
25    609,552      Action      Drama
26    683,149      Action     Crime       Drama
27  1,084,286      Action  Adventure     Fantasy
28    410,355   Animation     Action   Adventure
```

```
29    336,420     Action  Adventure    Mystery
30    199,079     Action  Biography      Drama
31     96,853     Action  Adventure      Drama
32    140,977     Action      Crime      Drama
33     64,776     Action      Crime   Thriller
34    125,557     Action  Biography      Drama
35     47,479     Action      Crime      Drama
36    783,759     Action  Adventure
37    813,538     Action  Adventure    Fantasy
38  1,022,937     Action  Adventure     Sci-Fi
39  1,514,502     Action      Crime      Drama
40    629,307     Action      Drama        War
41    906,120     Action   Thriller
42  1,147,660     Action      Crime   Thriller
43  1,145,964     Action      Drama     Sci-Fi
44    130,308     Action      Drama        War
45     66,095  Animation     Action  Adventure
46     99,904     Action     Comedy      Crime
47    108,086     Action      Drama
48    126,724     Action      Drama   Thriller
49     68,308     Action      Drama    History

    Certification
0           PG-13
1           PG-13
2              PG
3           PG-13
4           PG-13
5           PG-13
6               R
7              PG
8           TV-MA
9               R
10             PG
11      Not Rated
12      Not Rated
13
14              R
15              R
16             PG
17             PG
18          PG-13
19              R
20              R
21          PG-13
22          PG-13
23      Not Rated
```

```
24      Not Rated
25          PG-13
26              R
27             PG
28          PG-13
29       Approved
30      Not Rated
31      Not Rated
32      Not Rated
33
34          TV-14
35      Not Rated
36          PG-13
37          PG-13
38          PG-13
39          PG-13
40              R
41              R
42              R
43              R
44              R
45              R
46      Not Rated
47      Not Rated
48      Not Rated
49      Not Rated
```

[35]:
```python
# saving file to csv
df.to_csv('IMBD_Data1.csv', index=False)
df2.to_csv("IMDB_Data_2.csv", index=False)
```

[36]:
```python
output_path = "C:\shiietal\Downloads simplilearn\sonali project scrapping"
```

[37]:
```python
output_path2 = "C:\shiietal\Downloads simplilearn\sonali project scrapping"
```

[38]:
```python
output_file = "C:\shiietal\Downloads simplilearn\sonali project scrapping" +␣
 ↪"IMBD_Data1.csv"
```

[39]:
```python
output_file2 = "C:\shiietal\Downloads simplilearn\sonali project scrapping" +␣
 ↪"IMDB_Data_2.csv"
```

[40]:
```python
df
```

[40]:
```
   Sno                                      Movie Name  \
0    1                                 The Dark Knight
1    2    The Lord of the Rings: The Return of the King
2    3                  Spider-Man: Across the Spider-Verse
```

```
3        4                                                  Inception
4        5   The Lord of the Rings: The Fellowship of the Ring
…       …                                                         …
1745  1746                    Demon Slayer the Movie: Mugen Train
1746  1747                              Gangs of Wasseypur
1747  1748                       Baahubali 2: The Conclusion
1748  1749                                            Yojimbo
1749  1750                           Uri: The Surgical Strike


              Director1       Director2           Director3 Duration  Year  \
0        Christopher Nolan            -                   -   152 min  2008
1           Peter Jackson            -                   -   201 min  2003
2     Joaquim Dos Santos  Kemp Powers  Justin K. Thompson  140 min  2023
3        Christopher Nolan            -                   -   148 min  2010
4           Peter Jackson            -                   -   178 min  2001
…                   …             …                   …        …     …
1745      Haruo Sotozaki            -                   -   117 min  2020
1746      Anurag Kashyap            -                   -   321 min  2012
1747      S.S. Rajamouli            -                   -   167 min  2017
1748      Akira Kurosawa            -                   -   110 min  1961
1749        Aditya Dhar            -                   -   138 min  2019


      Ratings   Metascore
0         9.0  84
1         9.0  94
2         8.9  86
3         8.8  74
4         8.8  92
…         …         …
1745      8.2  72
1746      8.2  89
1747      8.2         -
1748      8.2  93
1749      8.2         -

[1750 rows x 9 columns]
```

[41]: `df2`

[41]:
```
                                        Movie Name                Star1  \
0                                   The Dark Knight    Christopher Nolan
1       The Lord of the Rings: The Return of the King         Peter Jackson
2             Spider-Man: Across the Spider-Verse  Joaquim Dos Santos
3                                         Inception    Christopher Nolan
4   The Lord of the Rings: The Fellowship of the Ring         Peter Jackson
5               The Lord of the Rings: The Two Towers         Peter Jackson
6                                         The Matrix     Lana Wachowski
```

7	Star Wars: Episode V - The Empire Strikes Back	Irvin Kershner
8	Soorarai Pottru	Sudha Kongara
9	Terminator 2: Judgment Day	James Cameron
10	Star Wars: Episode IV - A New Hope	George Lucas
11	Seven Samurai	Akira Kurosawa
12	Harakiri	Masaki Kobayashi
13	Sita Ramam	Hanu Raghavapudi
14	Gladiator	Ridley Scott
15	Léon: The Professional	Luc Besson
16	Raiders of the Lost Ark	Steven Spielberg
17	Spider-Man: Into the Spider-Verse	Bob Persichetti
18	Avengers: Endgame	Anthony Russo
19	Aliens	James Cameron
20	Oldboy	Park Chan-wook
21	The Dark Knight Rises	Christopher Nolan
22	Avengers: Infinity War	Anthony Russo
23	Kaithi	Lokesh Kanagaraj
24	Asuran	Vetrimaaran
25	Top Gun: Maverick	Joseph Kosinski
26	Heat	Michael Mann
27	Star Wars: Episode VI - Return of the Jedi	Richard Marquand
28	Princess Mononoke	Hayao Miyazaki
29	North by Northwest	Alfred Hitchcock
30	Dangal	Nitesh Tiwari
31	Kantara	Rishab Shetty
32	K.G.F: Chapter 2	Prashanth Neel
33	Vikram	Lokesh Kanagaraj
34	Shershaah	Vishnuvardhan
35	Raatchasan	Ram Kumar
36	Indiana Jones and the Last Crusade	Steven Spielberg
37	Spider-Man: No Way Home	Jon Watts
38	Jurassic Park	Steven Spielberg
39	Batman Begins	Christopher Nolan
40	1917	Sam Mendes
41	Die Hard	John McTiernan
42	Kill Bill: Vol. 1	Quentin Tarantino
43	V for Vendetta	James McTeigue
44	Ran	Akira Kurosawa
45	Demon Slayer the Movie: Mugen Train	Haruo Sotozaki
46	Gangs of Wasseypur	Anurag Kashyap
47	Baahubali 2: The Conclusion	S.S. Rajamouli
48	Yojimbo	Akira Kurosawa
49	Uri: The Surgical Strike	Aditya Dhar

	Star2	Star3	Star4 \
0	Christian Bale	Heath Ledger	Aaron Eckhart
1	Elijah Wood	Viggo Mortensen	Ian McKellen

2	Kemp Powers	Justin K. Thompson	Shameik Moore
3	Leonardo DiCaprio	Joseph Gordon-Levitt	Elliot Page
4	Elijah Wood	Ian McKellen	Orlando Bloom
5	Elijah Wood	Ian McKellen	Viggo Mortensen
6	Lilly Wachowski	Keanu Reeves	Laurence Fishburne
7	Mark Hamill	Harrison Ford	Carrie Fisher
8	Suriya	Paresh Rawal	Aparna Balamurali
9	Arnold Schwarzenegger	Linda Hamilton	Edward Furlong
10	Mark Hamill	Harrison Ford	Carrie Fisher
11	Toshirô Mifune	Takashi Shimura	Keiko Tsushima
12	Tatsuya Nakadai	Akira Ishihama	Shima Iwashita
13	Dulquer Salmaan	Mrunal Thakur	Rashmika Mandanna
14	Russell Crowe	Joaquin Phoenix	Connie Nielsen
15	Jean Reno	Gary Oldman	Natalie Portman
16	Harrison Ford	Karen Allen	Paul Freeman
17	Peter Ramsey	Rodney Rothman	Shameik Moore
18	Joe Russo	Robert Downey Jr.	Chris Evans
19	Sigourney Weaver	Michael Biehn	Carrie Henn
20	Choi Min-sik	Yoo Ji-tae	Kang Hye-jeong
21	Christian Bale	Tom Hardy	Anne Hathaway
22	Joe Russo	Robert Downey Jr.	Chris Hemsworth
23	Karthi	Narain	Arjun Das
24	Dhanush	Manju Warrier	Prakash Raj
25	Tom Cruise	Jennifer Connelly	Miles Teller
26	Al Pacino	Robert De Niro	Val Kilmer
27	Mark Hamill	Harrison Ford	Carrie Fisher
28	Yôji Matsuda	Yuriko Ishida	Yûko Tanaka
29	Cary Grant	Eva Marie Saint	James Mason
30	Aamir Khan	Sakshi Tanwar	Fatima Sana Shaikh
31	Rishab Shetty	Kishore Kumar G.	Achyuth Kumar
32	Yash	Sanjay Dutt	Raveena Tandon
33	Kamal Haasan	Vijay Sethupathi	Fahadh Faasil
34	Sidharth Malhotra	Kiara Advani	Shiv Panditt
35	Vishnu Vishal	Amala Paul	Radha Ravi
36	Harrison Ford	Sean Connery	Alison Doody
37	Tom Holland	Zendaya	Benedict Cumberbatch
38	Sam Neill	Laura Dern	Jeff Goldblum
39	Christian Bale	Michael Caine	Ken Watanabe
40	Dean-Charles Chapman	George MacKay	Daniel Mays
41	Bruce Willis	Alan Rickman	Bonnie Bedelia
42	Uma Thurman	David Carradine	Daryl Hannah
43	Hugo Weaving	Natalie Portman	Rupert Graves
44	Tatsuya Nakadai	Akira Terao	Jinpachi Nezu
45	Natsuki Hanae	Akari Kitô	Yoshitsugu Matsuoka
46	Manoj Bajpayee	Nawazuddin Siddiqui	Tigmanshu Dhulia
47	Prabhas	Rana Daggubati	Anushka Shetty
48	Toshirô Mifune	Eijirô Tôno	Tatsuya Nakadai

	Votes	Genre1	Genre2	Genre3	Gross Collection	Popularity	\
0	2,741,127	Action	Crime	Drama			
1	1,899,013	Action	Adventure	Drama			
2	185,491	Animation	Action	Adventure			
3	2,432,436	Action	Adventure	Sci-Fi			
4	1,927,425	Action	Adventure	Drama			
5	1,713,772	Action	Adventure	Drama			
6	1,971,133	Action	Sci-Fi				
7	1,329,834	Action	Adventure	Fantasy			
8	120,033	Action	Drama				
9	1,131,068	Action	Sci-Fi				
10	1,402,082	Action	Adventure	Fantasy			
11	354,942	Action	Drama				
12	62,143	Action	Drama	Mystery			
13	61,432	Action	Drama	Mystery			
14	1,546,170	Action	Adventure	Drama			
15	1,197,398	Action	Crime	Drama			
16	1,003,306	Action	Adventure				
17	609,082	Animation	Action	Adventure			
18	1,194,734	Action	Adventure	Drama			
19	737,021	Action	Adventure	Sci-Fi			
20	601,507	Action	Drama	Mystery			
21	1,754,277	Action	Drama	Thriller			
22	1,137,081	Action	Adventure	Sci-Fi			
23	36,785	Action	Adventure	Crime			
24	31,537	Action	Drama				
25	609,552	Action	Drama				
26	683,149	Action	Crime	Drama			
27	1,084,286	Action	Adventure	Fantasy			
28	410,355	Animation	Action	Adventure			
29	336,420	Action	Adventure	Mystery			
30	199,079	Action	Biography	Drama			
31	96,853	Action	Adventure	Drama			
32	140,977	Action	Crime	Drama			
33	64,776	Action	Crime	Thriller			
34	125,557	Action	Biography	Drama			
35	47,479	Action	Crime	Drama			
36	783,759	Action	Adventure				
37	813,538	Action	Adventure	Fantasy			
38	1,022,937	Action	Adventure	Sci-Fi			
39	1,514,502	Action	Crime	Drama			
40	629,307	Action	Drama	War			
41	906,120	Action	Thriller				
42	1,147,660	Action	Crime	Thriller			
43	1,145,964	Action	Drama	Sci-Fi			

```
44   130,308     Action      Drama        War
45    66,095   Animation     Action   Adventure
46    99,904     Action      Comedy       Crime
47   108,086     Action      Drama
48   126,724     Action      Drama     Thriller
49    68,308     Action      Drama      History


    Certification
0          PG-13
1          PG-13
2             PG
3          PG-13
4          PG-13
5          PG-13
6              R
7             PG
8          TV-MA
9              R
10            PG
11     Not Rated
12     Not Rated
13
14             R
15             R
16            PG
17            PG
18         PG-13
19             R
20             R
21         PG-13
22         PG-13
23     Not Rated
24     Not Rated
25         PG-13
26             R
27            PG
28         PG-13
29      Approved
30     Not Rated
31     Not Rated
32     Not Rated
33
34          TV-14
35     Not Rated
36         PG-13
37         PG-13
38         PG-13
```

```
39         PG-13
40             R
41             R
42             R
43             R
44             R
45             R
46     Not Rated
47     Not Rated
48     Not Rated
49     Not Rated
```

[ ]: