Webscraping Disney movies' dataset (wikipedia)

In [1]:
```
import requests
from bs4 import BeautifulSoup as bs
import pandas as pd
```

In [2]:
```
url = "https://en.wikipedia.org/wiki/Toy_Story_3"

r = requests.get(url)
web = bs(r.content)
```

Dumping Toy Story 3 Dataset in dictionary

In [3]:
```
# 1 way:
d = dict()
table = web.select("table.infobox.vevent")[0]
# print(table.prettify())
labels = table.select("tbody tr th.infobox-label")
data = table.select("tbody tr td.infobox-data")
name = table.select("tbody tr th.infobox-above.summary")[0].get_text()
keys = [key.get_text().strip(' \t\n\r').replace('\xa0', ' ') for key in labels]
values = [value.get_text().strip(' \t\n\r').replace('\xa0', ' ') for value in data]
d['Name'] = name
for i in range(len(keys)):
    d[keys[i]] = values[i]

# for key, value in d.items():
#     print(key + ": " + value + '\n')
```

In [4]:
```
# 2 way
inf = web.find(class_="infobox vevent")
inf_rows = inf.find_all("tr")

d = {}


def content_value(row):
    if row.find("li"):
        return [li.get_text(' ', strip=True).replace('\xa0', ' ') for li in row.find_all('
    else:
        return row.get_text(' ', strip=True).replace('\xa0', ' ')

for index, row in enumerate(inf_rows):
    if index == 0:
        d['Name'] = row.find("th").get_text(' ', strip=True)
    elif index == 1:
        continue
    else:
        key = row.find("th").get_text(' ', strip=True)
        value = content_value(row.find("td"))
        d[key] = value

d
```

Out[4]:
```
{'Name': 'Toy Story 3',
 'Directed by': 'Lee Unkrich',
 'Screenplay by': 'Michael Arndt',
```

```
'Story by': ['John Lasseter', 'Andrew Stanton', 'Lee Unkrich'],
'Produced by': 'Darla K. Anderson',
'Starring': ['Tom Hanks',
 'Tim Allen',
 'Joan Cusack',
 'Don Rickles',
 'Wallace Shawn',
 'John Ratzenberger',
 'Estelle Harris',
 'Ned Beatty',
 'Michael Keaton',
 'Jodi Benson',
 'John Morris'],
'Cinematography': ['Jeremy Lasky', 'Kim White'],
'Edited by': 'Ken Schretzmann',
'Music by': 'Randy Newman',
'Production companies': ['Walt Disney Pictures', 'Pixar Animation Studios'],
'Distributed by': 'Walt Disney Studios Motion Pictures',
'Release date': ['June 12, 2010 ( 2010-06-12 ) ( Taormina Film Fest )',
 'June 18, 2010 ( 2010-06-18 ) (United States)'],
'Running time': '103 minutes [1]',
'Country': 'United States',
'Language': 'English',
'Budget': '$200 million [1]',
'Box office': '$1.067 billion [1]'}
```

Infobox from all movie links

In [5]:

```
url = "https://en.wikipedia.org/wiki/List_of_Walt_Disney_Pictures_films"

r = requests.get(url)
web = bs(r.content)
```

In [6]:

```
# 1 way
tables = web.select('table.wikitable.sortable')
l = []

def content_value(row):
        if row:
            if row.find("li"):
                return [li.get_text(' ', strip=True).replace('\xa0', ' ') for li in row.fi
            elif row.find("br"):
                return [t for t in row.stripped_strings]
            else:
                return row.get_text(' ', strip=True).replace('\xa0', ' ')

def clean_tags(web):
    for tag in web.find_all(['sup', 'span']):
        tag.decompose()


for table in tables:
    url = table.select("tbody tr td i a")
    for i in url:
            d = {}
            link = "https://en.wikipedia.org" + i['href']
            r = requests.get(link)
            web = bs(r.content)

            clean_tags(web)

            inf = web.find(class_="infobox vevent")
            if inf:
```

```
                inf_rows = inf.find_all("tr")
                for index, row in enumerate(inf_rows):
                    if row.find("th"):
                        if index == 0:
                            d['Name'] = row.find("th").get_text(' ', strip=True)
                        elif index == 1:
                            continue
                        else:
                            key = row.find("th").get_text(' ', strip=True)
                            value = content_value(row.find("td"))
                            d[key] = value
            l.append(d)

print(len(l))
l[:5]
```

Out[6]:

```
509
[{'Name': 'Academy Award Review of',
  'Production company': 'Walt Disney Productions',
  'Distributed by': 'RKO Radio Pictures',
  'Release date': ['May 19, 1937'],
  'Running time': '41 minutes (74 minutes 1966 release)',
  'Country': 'United States',
  'Language': 'English',
  'Box office': '$45.472'},
 {'Name': 'Snow White and the Seven Dwarfs',
  'Directed by': ['David Hand',
   'William Cottrell',
   'Wilfred Jackson',
   'Larry Morey',
   'Perce Pearce',
   'Ben Sharpsteen'],
  'Written by': ['Ted Sears',
   'Richard Creedon',
   'Otto Englander',
   'Dick Rickard',
   'Earl Hurd',
   'Merrill De Maris',
   'Dorothy Ann Blank',
   'Webb Smith'],
  'Based on': ['Snow White', 'by The', 'Brothers Grimm'],
  'Produced by': 'Walt Disney',
  'Starring': ['Adriana Caselotti',
   'Lucille La Verne',
   'Harry Stockwell',
   'Roy Atwell',
   'Pinto Colvig',
   'Otis Harlan',
   'Scotty Mattraw',
   'Billy Gilbert',
   'Eddie Collins',
   'Moroni Olsen',
   'Stuart Buchanan'],
  'Music by': ['Frank Churchill', 'Paul Smith', 'Leigh Harline'],
  'Production company': 'Walt Disney Productions',
  'Distributed by': 'RKO Radio Pictures',
  'Release date': ['December 21, 1937 ( Carthay Circle Theatre )'],
  'Running time': '83 minutes',
  'Country': 'United States',
  'Language': 'English',
  'Budget': '$1.49 million',
  'Box office': '$418 million'},
 {'Name': 'Pinocchio',
  'Directed by': ['Ben Sharpsteen',
```

```
  'Hamilton Luske',
  'Bill Roberts',
  'Norman Ferguson',
  'Jack Kinney',
  'Wilfred Jackson',
  'T. Hee'],
 'Story by': ['Ted Sears',
  'Otto Englander',
  'Webb Smith',
  'William Cottrell',
  'Joseph Sabo',
  'Erdman Penner',
  'Aurelius Battaglia'],
 'Based on': ['The Adventures of Pinocchio', 'by', 'Carlo Collodi'],
 'Produced by': 'Walt Disney',
 'Starring': ['Cliff Edwards',
  'Dickie Jones',
  'Christian Rub',
  'Walter Catlett',
  'Charles Judels',
  'Evelyn Venable',
  'Frankie Darro'],
 'Music by': ['Leigh Harline', 'Paul J. Smith'],
 'Production company': 'Walt Disney Productions',
 'Distributed by': 'RKO Radio Pictures',
 'Release date': ['February 7, 1940 ( Center Theatre )',
  'February 23, 1940 (United States)'],
 'Running time': '88 minutes',
 'Country': 'United States',
 'Language': 'English',
 'Budget': '$2.6 million',
 'Box office': '$164 million'},
{'Name': 'Fantasia',
 'Directed by': ['Samuel Armstrong',
  'James Algar',
  'Bill Roberts',
  'Paul Satterfield',
  'Ben Sharpsteen',
  'David D. Hand',
  'Hamilton Luske',
  'Jim Handley',
  'Ford Beebe',
  'T. Hee',
  'Norman Ferguson',
  'Wilfred Jackson'],
 'Story by': ['Joe Grant', 'Dick Huemer'],
 'Produced by': ['Walt Disney', 'Ben Sharpsteen'],
 'Starring': ['Leopold Stokowski', 'Deems Taylor'],
 'Narrated by': 'Deems Taylor',
 'Cinematography': 'James Wong Howe',
 'Music by': 'See program',
 'Production company': 'Walt Disney Productions',
 'Distributed by': 'RKO Radio Pictures',
 'Release date': ['November 13, 1940'],
 'Running time': '126 minutes',
 'Country': 'United States',
 'Language': 'English',
 'Budget': '$2.28 million',
 'Box office': '$76.4–$83.3 million (United States and Canada)'},
{'Name': 'The Reluctant Dragon',
 'Directed by': ['Alfred Werker',
  '(live action)',
  'Hamilton Luske',
  '(animation)',
  'Jack Cutting',
  ','],
```

```
    'Ub Iwerks',
    ',',
    'Jack Kinney',
    '(sequence directors)'],
   'Written by': ['Live-action:',
    'Ted Sears',
    'Al Perkins',
    'Larry Clemmons',
    'Bill Cottrell',
    'Harry Clork',
    'Robert Benchley',
    'The Reluctant Dragon',
    'segment:',
    'Kenneth Grahame',
    '(original book)',
    'Erdman Penner',
    'T. Hee',
    'Baby Weems',
    'segment:',
    'Joe Grant',
    'Dick Huemer',
    'John Miller'],
   'Produced by': 'Walt Disney',
   'Starring': ['Robert Benchley',
    'Frances Gifford',
    'Buddy Pepper',
    'Nana Bryant'],
   'Cinematography': 'Bert Glennon',
   'Edited by': 'Paul Weatherwax',
   'Music by': ['Frank Churchill', 'Larry Morey'],
   'Production company': 'Walt Disney Productions',
   'Distributed by': 'RKO Radio Pictures',
   'Release date': ['June 27, 1941'],
   'Running time': '74 minutes',
   'Country': 'United States',
   'Language': 'English',
   'Budget': '$600,000',
   'Box office': '$960,000 (worldwide rentals)'}]
```

Saving data as JSON

In [7]:
```
import json

def save_data(title, data):
    with open(title, 'w', encoding="utf-8") as f:
        json.dump(data, f, ensure_ascii=False, indent = 2)

def load_data(title):
    with open(title, encoding="utf- 8") as f:
        return json.load(f)
```

In [8]:
```
save_data("disney_data.json", l)
```

In [9]:
```
# 2 way
# a bit inefficient duw to corner cases
'''
movies = web.select('.wikitable.sortable i a')

l = []
for index, movie in enumerate(movies):
    try:
```

```
        rel_path = movie['href']
        title = movie['title']
        l.append(get_info_box("https://en.wikipedia.org/" + rel_path))
    except Exception as e:
        print(movie.get_text())
        print(e)

def content_value(row):
    if row.find("li"):
        return [li.get_text(' ', strip=True).replace('\xa0', ' ') for li in row.find_all('
    else:
        return row.get_text(' ', strip=True).replace('\xa0', ' ')

def get_info_box(url):
    r = requests.get(url)
    web = bs(r.content)

    inf = web.find(class_="infobox vevent")
    inf_rows = inf.find_all("tr")

    d = {}
    for index, row in enumerate(inf_rows):
        if index == 0:
            d['Name'] = row.find("th").get_text(' ', strip=True)
        elif index == 1:
            continue
        else:
            key = row.find("th").get_text(' ', strip=True)
            value = content_value(row.find("td"))
            d[key] = value
    return d
'''
```

Out[9]:

```
'\nmovies = web.select(\'.wikitable.sortable i a\')\n\nl = []\nfor index, movie in enumera
te(movies):\n    try:\n        rel_path = movie[\'href\']\n        title = movie[\'title
\']\n        l.append(get_info_box("https://en.wikipedia.org/" + rel_path))\n    except Ex
ception as e:\n        print(movie.get_text())\n        print(e)\n\ndef content_value(ro
w):\n    if row.find("li"):\n        return [li.get_text(\' \', strip=True).replace(\'\xa0
\', \' \') for li in row.find_all("li")]\n    else:\n        return row.get_text(\' \', st
rip=True).replace(\'\xa0\', \' \')\n\ndef get_info_box(url):\n    r = requests.get(url)\n
   web = bs(r.content)\n    \n    inf = web.find(class_="infobox vevent")\n    inf_rows =
inf.find_all("tr")\n    \n    d = {}\n    for index, row in enumerate(inf_rows):\n
if index == 0:\n            d[\'Name\'] = row.find("th").get_text(\' \', strip=True)\n
    elif index == 1:\n            continue\n        else:\n            key = row.find("t
h").get_text(\' \', strip=True)\n            value = content_value(row.find("td"))\n
      d[key] = value \n    return d\n'
```

Cleaning the collected data

Cleaning references

In [10]:

```
# 1 way
movie_info_list = load_data("disney_data.json")
dummy_list = movie_info_list
```

In [11]:

```
# ref = [str([i]) for i in range(10)]

# for j in dummy_list:
#     for key, value in j.items():
#         if type(value) is list:
#             for i in range(len(value)):
#                 if value[i].split(" ")[-1] in ref:
```

```
#                     j[key][i] = value[i].split(" ")[:-2]
#         else:
#             if str(value).split(" ")[-1] in ref:
#                 j[key] = str(value).split(" ")[:-2]


# dummy_list
```

In [12]:

```
# 2 way (to optimise our data scraping itself, removing superscript tag)
# [t for t in row.stripped_strings] is separating starring long string of names (hit n tri
```

Coverting running time to integer

In [13]:

```
def min_to_int(value):
    if type(value) == list:
            t = int(value[0].split(" ")[0])
    else:
        try:
            t = int(value.split(" ")[0])
        except ValueError as e:
            t = int(value.split(" ")[0].split("-")[0])
    return t

for j in dummy_list:
    time = 0
    for key, value in j.items():
        if key == "Running time":
            time = min_to_int(j["Running time"])
    j["Running time (int)"] = time

dummy_list[:5]
```

Out[13]:

```
[{'Name': 'Academy Award Review of',
  'Production company': 'Walt Disney Productions',
  'Distributed by': 'RKO Radio Pictures',
  'Release date': ['May 19, 1937'],
  'Running time': '41 minutes (74 minutes 1966 release)',
  'Country': 'United States',
  'Language': 'English',
  'Box office': '$45.472',
  'Running time (int)': 41},
 {'Name': 'Snow White and the Seven Dwarfs',
  'Directed by': ['David Hand',
   'William Cottrell',
   'Wilfred Jackson',
   'Larry Morey',
   'Perce Pearce',
   'Ben Sharpsteen'],
  'Written by': ['Ted Sears',
   'Richard Creedon',
   'Otto Englander',
   'Dick Rickard',
   'Earl Hurd',
   'Merrill De Maris',
   'Dorothy Ann Blank',
   'Webb Smith'],
  'Based on': ['Snow White', 'by The', 'Brothers Grimm'],
  'Produced by': 'Walt Disney',
  'Starring': ['Adriana Caselotti',
   'Lucille La Verne',
   'Harry Stockwell',
   'Roy Atwell',
```

```
   'Pinto Colvig',
   'Otis Harlan',
   'Scotty Mattraw',
   'Billy Gilbert',
   'Eddie Collins',
   'Moroni Olsen',
   'Stuart Buchanan'],
  'Music by': ['Frank Churchill', 'Paul Smith', 'Leigh Harline'],
  'Production company': 'Walt Disney Productions',
  'Distributed by': 'RKO Radio Pictures',
  'Release date': ['December 21, 1937 ( Carthay Circle Theatre )'],
  'Running time': '83 minutes',
  'Country': 'United States',
  'Language': 'English',
  'Budget': '$1.49 million',
  'Box office': '$418 million',
  'Running time (int)': 83},
 {'Name': 'Pinocchio',
  'Directed by': ['Ben Sharpsteen',
   'Hamilton Luske',
   'Bill Roberts',
   'Norman Ferguson',
   'Jack Kinney',
   'Wilfred Jackson',
   'T. Hee'],
  'Story by': ['Ted Sears',
   'Otto Englander',
   'Webb Smith',
   'William Cottrell',
   'Joseph Sabo',
   'Erdman Penner',
   'Aurelius Battaglia'],
  'Based on': ['The Adventures of Pinocchio', 'by', 'Carlo Collodi'],
  'Produced by': 'Walt Disney',
  'Starring': ['Cliff Edwards',
   'Dickie Jones',
   'Christian Rub',
   'Walter Catlett',
   'Charles Judels',
   'Evelyn Venable',
   'Frankie Darro'],
  'Music by': ['Leigh Harline', 'Paul J. Smith'],
  'Production company': 'Walt Disney Productions',
  'Distributed by': 'RKO Radio Pictures',
  'Release date': ['February 7, 1940 ( Center Theatre )',
   'February 23, 1940 (United States)'],
  'Running time': '88 minutes',
  'Country': 'United States',
  'Language': 'English',
  'Budget': '$2.6 million',
  'Box office': '$164 million',
  'Running time (int)': 88},
 {'Name': 'Fantasia',
  'Directed by': ['Samuel Armstrong',
   'James Algar',
   'Bill Roberts',
   'Paul Satterfield',
   'Ben Sharpsteen',
   'David D. Hand',
   'Hamilton Luske',
   'Jim Handley',
   'Ford Beebe',
   'T. Hee',
   'Norman Ferguson',
   'Wilfred Jackson'],
  'Story by': ['Joe Grant', 'Dick Huemer'],
```

```
 'Produced by': ['Walt Disney', 'Ben Sharpsteen'],
 'Starring': ['Leopold Stokowski', 'Deems Taylor'],
 'Narrated by': 'Deems Taylor',
 'Cinematography': 'James Wong Howe',
 'Music by': 'See program',
 'Production company': 'Walt Disney Productions',
 'Distributed by': 'RKO Radio Pictures',
 'Release date': ['November 13, 1940'],
 'Running time': '126 minutes',
 'Country': 'United States',
 'Language': 'English',
 'Budget': '$2.28 million',
 'Box office': '$76.4–$83.3 million (United States and Canada)',
 'Running time (int)': 126},
{'Name': 'The Reluctant Dragon',
 'Directed by': ['Alfred Werker',
  '(live action)',
  'Hamilton Luske',
  '(animation)',
  'Jack Cutting',
  ',',
  'Ub Iwerks',
  ',',
  'Jack Kinney',
  '(sequence directors)'],
 'Written by': ['Live-action:',
  'Ted Sears',
  'Al Perkins',
  'Larry Clemmons',
  'Bill Cottrell',
  'Harry Clork',
  'Robert Benchley',
  'The Reluctant Dragon',
  'segment:',
  'Kenneth Grahame',
  '(original book)',
  'Erdman Penner',
  'T. Hee',
  'Baby Weems',
  'segment:',
  'Joe Grant',
  'Dick Huemer',
  'John Miller'],
 'Produced by': 'Walt Disney',
 'Starring': ['Robert Benchley',
  'Frances Gifford',
  'Buddy Pepper',
  'Nana Bryant'],
 'Cinematography': 'Bert Glennon',
 'Edited by': 'Paul Weatherwax',
 'Music by': ['Frank Churchill', 'Larry Morey'],
 'Production company': 'Walt Disney Productions',
 'Distributed by': 'RKO Radio Pictures',
 'Release date': ['June 27, 1941'],
 'Running time': '74 minutes',
 'Country': 'United States',
 'Language': 'English',
 'Budget': '$600,000',
 'Box office': '$960,000 (worldwide rentals)',
 'Running time (int)': 74}]
```

Converting dates to datetimes

In [14]:

```
from datetime import datetime
```

```
def date_conversion(date):
    if isinstance(date, list):
        date = date[0].strip()
    if date == 'N/A':
        return None
    else:
        date_str = date.split("(")[0].strip()
        order = "%B %d, %Y" #{local month name, day, year(without century)}
        try:
            return datetime.strptime(date_str, order)
        except ValueError as e:
            try:
                return datetime.strptime(date_str, "%d %B %Y")
            except:
                pass
        return None

# dates = [j.get("Release date", "N/A") for j in dummy_list]

# for d in date:
#     print(date_conversion(d))

for j in dummy_list:
    j["Release Date (datetime)"] = date_conversion(j.get("Release date", "N/A"))
```

Saving the data using Pickle

In [15]:
```
# JSON is not able to handle datetime objects
# Downside is pickle data is not human readable unless loaded via code
import pickle

def save_data_pickle(name, data):
    with open(name, 'wb') as f:
        pickle.dump(data, f)

def load_data_pickle(name):
    with open(name, 'rb') as f:
        return pickle.load(f)
```

In [16]:
```
save_data_pickle("disney_movie_data_cleaned.pickle", dummy_list)
```

In [17]:
```
movie_info_list = load_data_pickle("disney_movie_data_cleaned.pickle")
movie_info_list[:5]
```

Out[17]:
```
[{'Name': 'Academy Award Review of',
  'Production company': 'Walt Disney Productions',
  'Distributed by': 'RKO Radio Pictures',
  'Release date': ['May 19, 1937'],
  'Running time': '41 minutes (74 minutes 1966 release)',
  'Country': 'United States',
  'Language': 'English',
  'Box office': '$45.472',
  'Running time (int)': 41,
  'Release Date (datetime)': datetime.datetime(1937, 5, 19, 0, 0)},
 {'Name': 'Snow White and the Seven Dwarfs',
  'Directed by': ['David Hand',
   'William Cottrell',
   'Wilfred Jackson',
   'Larry Morey',
```

```
 'Perce Pearce',
 'Ben Sharpsteen'],
'Written by': ['Ted Sears',
 'Richard Creedon',
 'Otto Englander',
 'Dick Rickard',
 'Earl Hurd',
 'Merrill De Maris',
 'Dorothy Ann Blank',
 'Webb Smith'],
'Based on': ['Snow White', 'by The', 'Brothers Grimm'],
'Produced by': 'Walt Disney',
'Starring': ['Adriana Caselotti',
 'Lucille La Verne',
 'Harry Stockwell',
 'Roy Atwell',
 'Pinto Colvig',
 'Otis Harlan',
 'Scotty Mattraw',
 'Billy Gilbert',
 'Eddie Collins',
 'Moroni Olsen',
 'Stuart Buchanan'],
'Music by': ['Frank Churchill', 'Paul Smith', 'Leigh Harline'],
'Production company': 'Walt Disney Productions',
'Distributed by': 'RKO Radio Pictures',
'Release date': ['December 21, 1937 ( Carthay Circle Theatre )'],
'Running time': '83 minutes',
'Country': 'United States',
'Language': 'English',
'Budget': '$1.49 million',
'Box office': '$418 million',
'Running time (int)': 83,
'Release Date (datetime)': datetime.datetime(1937, 12, 21, 0, 0)},
{'Name': 'Pinocchio',
 'Directed by': ['Ben Sharpsteen',
  'Hamilton Luske',
  'Bill Roberts',
  'Norman Ferguson',
  'Jack Kinney',
  'Wilfred Jackson',
  'T. Hee'],
 'Story by': ['Ted Sears',
  'Otto Englander',
  'Webb Smith',
  'William Cottrell',
  'Joseph Sabo',
  'Erdman Penner',
  'Aurelius Battaglia'],
 'Based on': ['The Adventures of Pinocchio', 'by', 'Carlo Collodi'],
 'Produced by': 'Walt Disney',
 'Starring': ['Cliff Edwards',
  'Dickie Jones',
  'Christian Rub',
  'Walter Catlett',
  'Charles Judels',
  'Evelyn Venable',
  'Frankie Darro'],
 'Music by': ['Leigh Harline', 'Paul J. Smith'],
 'Production company': 'Walt Disney Productions',
 'Distributed by': 'RKO Radio Pictures',
 'Release date': ['February 7, 1940 ( Center Theatre )',
  'February 23, 1940 (United States)'],
 'Running time': '88 minutes',
 'Country': 'United States',
 'Language': 'English',
```

```
 'Budget': '$2.6 million',
 'Box office': '$164 million',
 'Running time (int)': 88,
 'Release Date (datetime)': datetime.datetime(1940, 2, 7, 0, 0)},
{'Name': 'Fantasia',
 'Directed by': ['Samuel Armstrong',
  'James Algar',
  'Bill Roberts',
  'Paul Satterfield',
  'Ben Sharpsteen',
  'David D. Hand',
  'Hamilton Luske',
  'Jim Handley',
  'Ford Beebe',
  'T. Hee',
  'Norman Ferguson',
  'Wilfred Jackson'],
 'Story by': ['Joe Grant', 'Dick Huemer'],
 'Produced by': ['Walt Disney', 'Ben Sharpsteen'],
 'Starring': ['Leopold Stokowski', 'Deems Taylor'],
 'Narrated by': 'Deems Taylor',
 'Cinematography': 'James Wong Howe',
 'Music by': 'See program',
 'Production company': 'Walt Disney Productions',
 'Distributed by': 'RKO Radio Pictures',
 'Release date': ['November 13, 1940'],
 'Running time': '126 minutes',
 'Country': 'United States',
 'Language': 'English',
 'Budget': '$2.28 million',
 'Box office': '$76.4–$83.3 million (United States and Canada)',
 'Running time (int)': 126,
 'Release Date (datetime)': datetime.datetime(1940, 11, 13, 0, 0)},
{'Name': 'The Reluctant Dragon',
 'Directed by': ['Alfred Werker',
  '(live action)',
  'Hamilton Luske',
  '(animation)',
  'Jack Cutting',
  ',',
  'Ub Iwerks',
  ',',
  'Jack Kinney',
  '(sequence directors)'],
 'Written by': ['Live-action:',
  'Ted Sears',
  'Al Perkins',
  'Larry Clemmons',
  'Bill Cottrell',
  'Harry Clork',
  'Robert Benchley',
  'The Reluctant Dragon',
  'segment:',
  'Kenneth Grahame',
  '(original book)',
  'Erdman Penner',
  'T. Hee',
  'Baby Weems',
  'segment:',
  'Joe Grant',
  'Dick Huemer',
  'John Miller'],
 'Produced by': 'Walt Disney',
 'Starring': ['Robert Benchley',
  'Frances Gifford',
  'Buddy Pepper',
```

```
 'Nana Bryant'],
'Cinematography': 'Bert Glennon',
'Edited by': 'Paul Weatherwax',
'Music by': ['Frank Churchill', 'Larry Morey'],
'Production company': 'Walt Disney Productions',
'Distributed by': 'RKO Radio Pictures',
'Release date': ['June 27, 1941'],
'Running time': '74 minutes',
'Country': 'United States',
'Language': 'English',
'Budget': '$600,000',
'Box office': '$960,000 (worldwide rentals)',
'Running time (int)': 74,
'Release Date (datetime)': datetime.datetime(1941, 6, 27, 0, 0)}]
```

Saving data as JSON and CSV

In [18]:
```
dummy_list = [m.copy() for m in movie_info_list]
for m in dummy_list:
    current_date = m["Release Date (datetime)"]
    if current_date:
        m["Release Date (datetime)"] = current_date.strftime("%B %d, %Y")
    else:
        m["Release Date (datetime)"] = None
```

In [19]:
```
save_data("Disney_data_final.json", dummy_list)
```

In [20]:
```
import pandas as pd

df = pd.DataFrame(dummy_list)
df.head()
```

Out[20]:

	Name	Production company	Distributed by	Release date	Running time	Country	Language	Box office	Running time (int)	Release Date (datetime)	...
0	Academy Award Review of	Walt Disney Productions	RKO Radio Pictures	[May 19, 1937]	41 minutes (74 minutes 1966 release)	United States	English	$45.472	41	May 19, 1937	...
1	Snow White and the Seven Dwarfs	Walt Disney Productions	RKO Radio Pictures	[December 21, 1937 (Carthay Circle Theatre)]	83 minutes	United States	English	$418 million	83	December 21, 1937	...
2	Pinocchio	Walt Disney Productions	RKO Radio Pictures	[February 7, 1940 (Center Theatre), February...	88 minutes	United States	English	$164 million	88	February 07, 1940	...
3	Fantasia	Walt Disney Productions	RKO Radio Pictures	[November 13, 1940]	126 minutes	United States	English	76.4–83.3 million (United States and Canada)	126	November 13, 1940	...

	Name	Production company	Distributed by	Release date	Running time	Country	Language	Box office	Running time (int)	Release Date (datetime)	...
4	The Reluctant Dragon	Walt Disney Productions	RKO Radio Pictures	[June 27, 1941]	74 minutes	United States	English	$960,000 (worldwide rentals)	74	June 27, 1941	...

5 rows × 72 columns

In [21]:
```
df.to_csv("Disney_data_final.csv")
```

Analysis trial

longest runtime movies

In [22]:
```
run_time = df.sort_values('Running time (int)', ascending = False)
run_time.head()
```

Out[22]:

	Name	Production company	Distributed by	Release date	Running time	Country	Language	Box office	Running time (int)	Relea Da (datetim
330	Pirates of the Caribbean: At World's End	NaN	Buena Vista Pictures	[May 19, 2007 (Disneyland Resort), May 25, 2...	167 minutes	United States	English	$960.9 million	167	May 20
88	The Happiest Millionaire	Walt Disney Productions	Buena Vista Distribution	[June 23, 1967, November 30, 1967]	[164 minutes, (, Los Angeles, premiere), 144 m...	United States	English	$5 million (U.S./Canada rentals)	164	June 19
443	Jagga Jasoos	NaN	UTV Motion Pictures	[14 July 2017]	162 minutes	India	Hindi	83 crore	162	July 20
436	Dangal	NaN	UTV Motion Pictures	[21 December 2016 (United States), 23 December...	161 minutes	India	Hindi	(US$270 million)	161	Decem 21, 20
468	Hamilton	NaN	Walt Disney Studios Motion Pictures	[July 3, 2020]	160 minutes	United States	English	NaN	160	July 20

5 rows × 72 columns

Project by Anmol Sharma