Exploring the website "https://www.cgv.id (https://www.cgv.id)" it is observed that the website has common pattern:

1. Each cinema is labeled with number: for example: Grand Indonesia: https://www.cgv.id/en/schedule/cinema/002 (https://www.cgv.id/en/schedule/cinema/002), Pacific Place: https://www.cgv.id/en/schedule/cinema/003 (https://www.cgv.id/en/schedule/cinema/003)
2. The date in cinema is labeled directly after the link: for example: 2018-02-21: https://www.cgv.id/en/schedule/cinema/003/2018-02-21 (https://www.cgv.id/en/schedule/cinema/003/2018-02-21)
3. Label in movie: Dilan: https://www.cgv.id/en/movies/info/17009800 (https://www.cgv.id/en/movies/info/17009800), The Greatest Showman: https://www.cgv.id/en/movies/info/17009700 (https://www.cgv.id/en/movies/info/17009700).

Now, we will loop through multiple web pages to extract information from it.

- Website with the root: https://www.cgv.id/en/schedule/cinema (https://www.cgv.id/en/schedule/cinema) will be used to extract: Location Data (City, Theater) and Movies Data (Date, Theater, Time)
- Website with the root: https://www.cgv.id/en/movies/info (https://www.cgv.id/en/movies/info) will be used to extract: Movies Data (Title, Genre, Image URL).

In [1]:

```
# load library
%matplotlib inline
import matplotlib.pyplot as plt
import urllib.request
import pandas as pd
from bs4 import BeautifulSoup
from subprocess import check_output
```

1. Exploration

In [2]:

```
url = "https://www.cgv.id/en/schedule/cinema/"
request = urllib.request.Request(url)
page = urllib.request.urlopen(request)
soup = BeautifulSoup(page,"lxml")

# city
soup.find('div', class_="sect-city").find_all('a')
```

Out[2]:

```
[<a href="javascript:void(0);">Jakarta</a>,
 <a class="cinema_fav" href="/en/schedule/cinema/002" id="002" title="Gran
d Indonesia">Grand Indonesia</a>,
 <a class="cinema_fav" href="/en/schedule/cinema/003" id="003" title="Paci
fic Place">Pacific Place</a>,
 <a class="cinema_fav" href="/en/schedule/cinema/004" id="004" title="Mall
of Indonesia">Mall of Indonesia</a>,
 <a class="cinema_fav" href="/en/schedule/cinema/006" id="006" title="Cent
ral Park">Central Park</a>,
 <a class="cinema_fav" href="/en/schedule/cinema/020" id="020" title="Slip
i Jaya">Slipi Jaya</a>,
 <a class="cinema_fav" href="/en/schedule/cinema/025" id="025" title="Gree
n Pramuka Mall">Green Pramuka Mall</a>,
 <a class="cinema_fav" href="/en/schedule/cinema/028" id="028" title="Bell
a Terra Lifestyle Center">Bella Terra Lifestyle Center</a>,
 <a class="cinema_fav" href="/en/schedule/cinema/035" id="035" title="Tran
smart Cempaka Putih">Transmart Cempaka Putih</a>,
 <a class="cinema_fav" href="/en/schedule/cinema/037" id="037" title="Aeon
Mall">Aeon Mall</a>,
 <a href="javascript:void(0);">Bandung</a>,
 <a class="cinema_fav" href="/en/schedule/cinema/001" id="001" title="Pari
s Van Java">Paris Van Java</a>,
 <a class="cinema_fav" href="/en/schedule/cinema/011" id="011" title="Miko
Mall">Miko Mall</a>,
 <a class="cinema_fav" href="/en/schedule/cinema/014" id="014" title="BEC
Mall">BEC Mall</a>,
 <a class="cinema_fav" href="/en/schedule/cinema/029" id="029" title="23 P
askal Shopping Center">23 Paskal Shopping Center</a>,
 <a class="cinema_fav" href="/en/schedule/cinema/038" id="038" title="Metr
o Indah Mall">Metro Indah Mall</a>,
 <a href="javascript:void(0);" id="0">Balikpapan</a>,
 <a class="cinema_fav" href="/en/schedule/cinema/008" id="008" title="Plaz
a Balikpapan">Plaza Balikpapan</a>,
 <a href="javascript:void(0);" id="1">Batam</a>,
 <a class="cinema_fav" href="/en/schedule/cinema/009" id="009" title="Kepr
i Mall">Kepri Mall</a>,
 <a class="cinema_fav" href="/en/schedule/cinema/010" id="010" title="Harb
our Bay">Harbour Bay</a>,
 <a href="javascript:void(0);" id="2">Bekasi</a>,
 <a class="cinema_fav" href="/en/schedule/cinema/007" id="007" title="Beka
si Cyber Park">Bekasi Cyber Park</a>,
 <a class="cinema_fav" href="/en/schedule/cinema/036" id="036" title="Beka
si Trade Center">Bekasi Trade Center</a>,
 <a class="cinema_fav" href="/en/schedule/cinema/040" id="040" title="Lago
on Avenue Bekasi">Lagoon Avenue Bekasi</a>,
 <a href="javascript:void(0);" id="3">Cirebon</a>,
 <a class="cinema_fav" href="/en/schedule/cinema/016" id="016" title="Grag
e City Mall">Grage City Mall</a>,
 <a class="cinema_fav" href="/en/schedule/cinema/042" id="042" title="Tran
smart Cirebon">Transmart Cirebon</a>,
smart Cirebon">Transmart Cirebon</a>,
 <a href="javascript:void(0);" id="4">Depok</a>,
 <a class="cinema_fav" href="/en/schedule/cinema/030" id="030" title="Depo
k Mall">Depok Mall</a>,
 <a href="javascript:void(0);" id="5">Karawang</a>,
 <a class="cinema_fav" href="/en/schedule/cinema/019" id="019" title="Fest
ive Walk">Festive Walk</a>,
 <a class="cinema_fav" href="/en/schedule/cinema/046" id="046" title="Tech
nomart">Technomart</a>,
 <a href="javascript:void(0);" id="6">Lampung</a>,
 <a class="cinema_fav" href="/en/schedule/cinema/043" id="043" title="Tran
smart Lampung">Transmart Lampung</a>,
 <a href="javascript:void(0);" id="7">Makassar</a>,
 <a class="cinema_fav" href="/en/schedule/cinema/044" id="044" title="Daya
Grand Square">Daya Grand Square</a>,
 <a href="javascript:void(0);" id="8">Manado</a>,
 <a class="cinema_fav" href="/en/schedule/cinema/021" id="021" title="Gran
d Kawanua City">Grand Kawanua City</a>,
 <a href="javascript:void(0);" id="9">Mataram</a>,
 <a class="cinema_fav" href="/en/schedule/cinema/034" id="034" title="Tran
smart Mataram">Transmart Mataram</a>,
 <a href="javascript:void(0);" id="10">Medan</a>,
 <a class="cinema_fav" href="/en/schedule/cinema/024" id="024" title="Foca
l Point">Focal Point</a>,
 <a href="javascript:void(0);" id="11">Mojokerto</a>,
 <a class="cinema_fav" href="/en/schedule/cinema/023" id="023" title="Sunr
ise Mall">Sunrise Mall</a>,
 <a href="javascript:void(0);" id="12">Other</a>,
 <a class="cinema_fav" href="/en/schedule/cinema/047" id="047" title="[E]T
ransmart Sidoarjo">[E]Transmart Sidoarjo</a>,
 <a class="cinema_fav" href="/en/schedule/cinema/051" id="051" title="[E]I
con Mall Gresik">[E]Icon Mall Gresik</a>,
 <a href="javascript:void(0);" id="13">Palembang</a>,
 <a class="cinema_fav" href="/en/schedule/cinema/027" id="027" title="Soci
al Market">Social Market</a>,
 <a class="cinema_fav" href="/en/schedule/cinema/039" id="039" title="Tran
smart Palembang">Transmart Palembang</a>,
 <a href="javascript:void(0);" id="14">Pekanbaru</a>,
 <a class="cinema_fav" href="/en/schedule/cinema/033" id="033" title="Tran
smart Pekanbaru">Transmart Pekanbaru</a>,
 <a href="javascript:void(0);" id="15">Purwokerto</a>,
 <a class="cinema_fav" href="/en/schedule/cinema/026" id="026" title="Rita
Supermall">Rita Supermall</a>,
 <a href="javascript:void(0);" id="16">Solo</a>,
 <a class="cinema_fav" href="/en/schedule/cinema/041" id="041" title="Tran
smart Solo">Transmart Solo</a>,
 <a href="javascript:void(0);" id="17">Surabaya</a>,
 <a class="cinema_fav" href="/en/schedule/cinema/018" id="018" title="Marv
ell City">Marvell City</a>,
 <a class="cinema_fav" href="/en/schedule/cinema/048" id="048" title="[E]B
G Junction">[E]BG Junction</a>,
 <a href="javascript:void(0);" id="18">Tangerang</a>,
 <a class="cinema_fav" href="/en/schedule/cinema/005" id="005" title="Tera
s Kota">Teras Kota</a>,
 <a class="cinema_fav" href="/en/schedule/cinema/015" id="015" title="Band
ara City Mall">Bandara City Mall</a>,
 <a class="cinema_fav" href="/en/schedule/cinema/022" id="022" title="Ecop
laza Citraraya Cikupa">Ecoplaza Citraraya Cikupa</a>,
 <a href="javascript:void(0);" id="19">Tangerang Selatan</a>,
 <a class="cinema_fav" href="/en/schedule/cinema/045" id="045" title="Tran
smart Bintaro">Transmart Bintaro</a>,
 <a href="javascript:void(0);" id="20">Tegal</a>,
 <a class="cinema_fav" href="/en/schedule/cinema/032" id="032" title="Tran
smart Tegal">Transmart Tegal</a>,
 <a href="javascript:void(0);" id="21">Yogyakarta</a>,

 <a class="cinema_fav" href="/en/schedule/cinema/013" id="013" title="Jwal
k Mall">Jwalk Mall</a>,
 <a class="cinema_fav" href="/en/schedule/cinema/017" id="017" title="Hart
ono Mall">Hartono Mall</a>,
 <a class="cinema_fav" href="/en/schedule/cinema/031" id="031" title="Tran
smart Maguwo">Transmart Maguwo</a>]
```

In [3]:

```
# cinema meta-data
soup.find_all('div', class_='schedule-container')[0].select('.active')
```

Out[3]:

```
[<a attr-fmt="SCREENX 2D" attr-mov="SCREENX 2D BLACK PANTHER" class="active"
href="javascript:void(0);" id="225555">19:30</a>,
 <a attr-fmt="SCREENX 2D" attr-mov="SCREENX 2D BLACK PANTHER" class="active"
href="javascript:void(0);" id="225556">22:30</a>,
 <a attr-fmt="2D" attr-mov=" BLACK PANTHER" class="active" href="javascript:
void(0);" id="225509">20:00</a>,
 <a attr-fmt="2D" attr-mov=" BLACK PANTHER" class="active" href="javascript:
void(0);" id="225510">20:00</a>,
 <a attr-fmt="2D" attr-mov=" BLACK PANTHER" class="active" href="javascript:
void(0);" id="225545">22:50</a>,
 <a attr-fmt="2D" attr-mov=" BLACK PANTHER" class="active" href="javascript:
void(0);" id="225550">18:30</a>,
 <a attr-fmt="2D" attr-mov=" BLACK PANTHER" class="active" href="javascript:
void(0);" id="225551">21:30</a>,
 <a attr-fmt="2D" attr-mov=" BLACK PANTHER" class="active" href="javascript:
void(0);" id="225515">19:00</a>,
 <a attr-fmt="2D" attr-mov=" BLACK PANTHER" class="active" href="javascript:
void(0);" id="225516">22:00</a>,
 <a attr-fmt="4DX3D" attr-mov="4DX3D BLACK PANTHER" class="active" href="jav
ascript:void(0);" id="225530">20:40</a>,
 <a attr-fmt="2D" attr-mov=" THE POST" class="active" href="javascript:void
(0);" id="225541">18:45</a>,
 <a attr-fmt="2D" attr-mov=" THE POST" class="active" href="javascript:void
(0);" id="225542">21:10</a>,
 <a attr-fmt="2D" attr-mov=" THE POST" class="active" href="javascript:void
(0);" id="225544">20:10</a>,
 <a attr-fmt="2D" attr-mov=" SAMSON" class="active" href="javascript:void
(0);" id="225519">19:15</a>,
 <a attr-fmt="2D" attr-mov=" SAMSON" class="active" href="javascript:void
(0);" id="231169">21:40</a>,
 <a attr-fmt="2D" attr-mov=" EIFFEL I`M IN LOVE 2" class="active" href="java
script:void(0);" id="225536">19:15</a>,
 <a attr-fmt="2D" attr-mov=" EIFFEL I`M IN LOVE 2" class="active" href="java
script:void(0);" id="225537">21:45</a>,
 <a attr-fmt="2D" attr-mov=" MONSTER HUNT 2" class="active" href="javascrip
t:void(0);" id="225532">19:40</a>,
 <a attr-fmt="2D" attr-mov=" DILAN 1990" class="active" href="javascript:voi
d(0);" id="225523">18:15</a>,
 <a attr-fmt="2D" attr-mov=" DILAN 1990" class="active" href="javascript:voi
d(0);" id="233792">22:00</a>,
 <a attr-fmt="2D" attr-mov=" AIYAARY" class="active" href="javascript:void
(0);" id="225526">20:45</a>]
```

In [4]:

```
soup.find_all('div', class_='schedule-container')[0].select('.disabled')
```

Out[4]:

```
[<a attr-fmt="SCREENX 2D" attr-mov="SCREENX 2D BLACK PANTHER" class="disab
led" href="javascript:void(0);" id="225552">10:30</a>,
 <a attr-fmt="SCREENX 2D" attr-mov="SCREENX 2D BLACK PANTHER" class="disab
led" href="javascript:void(0);" id="225553">13:30</a>,
 <a attr-fmt="SCREENX 2D" attr-mov="SCREENX 2D BLACK PANTHER" class="disab
led" href="javascript:void(0);" id="225554">16:30</a>,
 <a attr-fmt="2D" attr-mov=" BLACK PANTHER" class="disabled" href="javascr
ipt:void(0);" id="225503">11:00</a>,
 <a attr-fmt="2D" attr-mov=" BLACK PANTHER" class="disabled" href="javascr
ipt:void(0);" id="225505">14:00</a>,
 <a attr-fmt="2D" attr-mov=" BLACK PANTHER" class="disabled" href="javascr
ipt:void(0);" id="225507">17:00</a>,
 <a attr-fmt="2D" attr-mov=" BLACK PANTHER" class="disabled" href="javascr
ipt:void(0);" id="225504">11:00</a>,
 <a attr-fmt="2D" attr-mov=" BLACK PANTHER" class="disabled" href="javascr
ipt:void(0);" id="225506">14:00</a>,
 <a attr-fmt="2D" attr-mov=" BLACK PANTHER" class="disabled" href="javascr
ipt:void(0);" id="225508">17:00</a>,
 <a attr-fmt="2D" attr-mov=" BLACK PANTHER" class="disabled" href="javascr
ipt:void(0);" id="225547">11:30</a>,
 <a attr-fmt="2D" attr-mov=" BLACK PANTHER" class="disabled" href="javascr
ipt:void(0);" id="225546">14:30</a>,
 <a attr-fmt="2D" attr-mov=" BLACK PANTHER" class="disabled" href="javascr
ipt:void(0);" id="225548">12:30</a>,
 <a attr-fmt="2D" attr-mov=" BLACK PANTHER" class="disabled" href="javascr
ipt:void(0);" id="225549">15:30</a>,
 <a attr-fmt="2D" attr-mov=" BLACK PANTHER" class="disabled" href="javascr
ipt:void(0);" id="225513">13:00</a>,
 <a attr-fmt="2D" attr-mov=" BLACK PANTHER" class="disabled" href="javascr
ipt:void(0);" id="225514">16:00</a>,
 <a attr-fmt="4DX2D" attr-mov="4DX2D BLACK PANTHER" class="disabled" href
="javascript:void(0);" id="225527">12:00</a>,
 <a attr-fmt="4DX2D" attr-mov="4DX2D BLACK PANTHER" class="disabled" href
="javascript:void(0);" id="225529">17:45</a>,
 <a attr-fmt="4DX3D" attr-mov="4DX3D BLACK PANTHER" class="disabled" href
="javascript:void(0);" id="225528">14:50</a>,
 <a attr-fmt="2D" attr-mov=" THE POST" class="disabled" href="javascript:v
oid(0);" id="225538">11:15</a>,
 <a attr-fmt="2D" attr-mov=" THE POST" class="disabled" href="javascript:v
oid(0);" id="228532">13:45</a>,
 <a attr-fmt="2D" attr-mov=" THE POST" class="disabled" href="javascript:v
oid(0);" id="225540">16:15</a>,
 <a attr-fmt="2D" attr-mov=" THE POST" class="disabled" href="javascript:v
oid(0);" id="225543">17:30</a>,
 <a attr-fmt="2D" attr-mov=" SAMSON" class="disabled" href="javascript:voi
d(0);" id="225521">11:15</a>,
 <a attr-fmt="2D" attr-mov=" SAMSON" class="disabled" href="javascript:voi
d(0);" id="225520">16:50</a>,
 <a attr-fmt="2D" attr-mov=" EIFFEL I`M IN LOVE 2" class="disabled" href
="javascript:void(0);" id="225535">11:45</a>,
 <a attr-fmt="2D" attr-mov=" EIFFEL I`M IN LOVE 2" class="disabled" href
="javascript:void(0);" id="232426">14:15</a>,
 <a attr-fmt="2D" attr-mov=" EIFFEL I`M IN LOVE 2" class="disabled" href
="javascript:void(0);" id="225522">16:45</a>,
 <a attr-fmt="2D" attr-mov=" MONSTER HUNT 2" class="disabled" href="javasc
ript:void(0);" id="225534">12:35</a>,
 <a attr-fmt="2D" attr-mov=" MONSTER HUNT 2" class="disabled" href="javasc
ript:void(0);" id="233791">15:00</a>,
 <a attr-fmt="2D" attr-mov=" MONSTER HUNT 2" class="disabled" href="javasc
ript:void(0);" id="225531">17:20</a>,
 <a attr-fmt="2D" attr-mov=" DILAN 1990" class="disabled" href="javascrip
t:void(0);" id="225525">13:15</a>,
 <a attr-fmt="2D" attr-mov=" DILAN 1990" class="disabled" href="javascrip
t:void(0);" id="228416">15:40</a>,
 <a attr-fmt="2D" attr-mov=" AIYAARY" class="disabled" href="javascript:vo
id(0);" id="225517">13:40</a>]
```

Note: Some of the disabled movie has the attr-mov removed. Hence, we could not extract the movie title. This is the source of missing value in Movie feature. Example:

In [5]:

```
url = 'https://www.cgv.id/en/schedule/cinema/041/2018-02-22'
request = urllib.request.Request(url)
page = urllib.request.urlopen(request)
soup2 = BeautifulSoup(page,"lxml")
soup2.find_all('div', class_='schedule-container')[0].select('.disabled')
```

Out[5]:

```
[<a attr-fmt="2D" attr-mov=" DILAN 1990" class="disabled" href="javascript:v
oid(0);" id="213708">10:45</a>,
 <a attr-fmt="2D" attr-mov=" DILAN 1990" class="disabled" href="javascript:v
oid(0);" id="233147">15:35</a>,
 <a attr-fmt="2D" attr-mov=" YOWIS BEN" class="disabled" href="javascript:vo
id(0);" id="213819">11:00</a>,
 <a attr-fmt="2D" attr-mov=" YOWIS BEN" class="disabled" href="javascript:vo
id(0);" id="213820">13:10</a>,
 <a attr-fmt="2D" attr-mov=" YOWIS BEN" class="disabled" href="javascript:vo
id(0);" id="213821">15:20</a>,
 <a attr-fmt="2D" attr-mov=" YOWIS BEN" class="disabled" href="javascript:vo
id(0);" id="213822">17:30</a>,
 <a attr-fmt="2D" attr-mov=" MEET ME AFTER SUNSET" class="disabled" href="ja
vascript:void(0);" id="213825">11:15</a>,
 <a attr-fmt="2D" attr-mov=" MEET ME AFTER SUNSET" class="disabled" href="ja
vascript:void(0);" id="213826">13:25</a>,
 <a attr-fmt="2D" attr-mov=" BLACK PANTHER" class="disabled" href="javascrip
t:void(0);" id="213714">11:30</a>,
 <a attr-fmt="2D" attr-mov=" BLACK PANTHER" class="disabled" href="javascrip
t:void(0);" id="213715">14:15</a>,
 <a attr-fmt="2D" attr-mov=" BLACK PANTHER" class="disabled" href="javascrip
t:void(0);" id="213716">17:00</a>,
 <a attr-fmt="2D" attr-mov=" THE POST" class="disabled" href="javascript:voi
d(0);" id="213829">12:00</a>,
 <a attr-fmt="2D" attr-mov=" THE POST" class="disabled" href="javascript:voi
d(0);" id="213722">16:25</a>,
 <a attr-fmt="2D" attr-mov=" EIFFEL I`M IN LOVE 2" class="disabled" href="ja
vascript:void(0);" id="213709">13:05</a>,
 <a attr-fmt="2D" attr-mov=" EIFFEL I`M IN LOVE 2" class="disabled" href="ja
vascript:void(0);" id="213711">17:30</a>,
 <a attr-fmt="2D" attr-mov=" BAYI GAIB: BAYI TUMBAL BAYI MATI" class="disabl
ed" href="javascript:void(0);" id="213831">14:25</a>,
 <a attr-fmt="2D" attr-mov=" LONDON LOVE STORY 3" class="disabled" href="jav
ascript:void(0);" id="213710">15:30</a>]
```

In [6]:

```
x = soup2.find_all('div', class_='schedule-container')[0].select('.disabled')
str(type(x[0].get('attr-mov'))) == "<class 'NoneType'>"
```

Out[6]:

```
False
```

In [7]:

```
str(type(x[0].get('attr-mov')))
```

Out[7]:

```
"<class 'str'>"
```

In [8]:

```
url = "https://www.cgv.id/en/movies/info/17009800"
request = urllib.request.Request(url)
page = urllib.request.urlopen(request)
soup = BeautifulSoup(page,"lxml")

# movie title
soup.find('div', class_='movie-info-title').string
```

Out[8]:

```
'\r\n\t\t\t\t\tDILAN 1990\t\t\t\t'
```

In [9]:

```
# movie link
soup.find('div', class_="poster-section left").find('img')['src']
```

Out[9]:

```
'/uploads/movie/compressed/17009800.jpg'
```

In [36]:

```
# movie meta-data
soup.find('div', class_='movie-add-info left').find_all('li')
```

Out[36]:

```
[<li>STARRING : Sushar Manaying, Azman Hassan, Teddy Chin</li>,
 <li>DIRECTOR : Ryon Lee</li>,
 <li>CENSOR RATING : 17+</li>,
 <li>GENRE : HORROR </li>,
 <li>LANGUAGE : Other</li>,
 <li>SUBSTITLE : BAHASA INDONESIA</li>,
 <li>DURATION : 92 Minutes</li>]
```

2. Scrapping

In [11]:

```
def city_cinema_df(soup):
    """
    This function create city and cinema dataframe.
    Input: soup
    Output: dataframe
    """
    # Create city, cinema dataframe
    x = soup.find('div', class_="sect-city").find_all('a')
    cin = ''
    city = []
    cinema = []
    for i in range(len(x)):
        if soup.find('div', class_="sect-city").find_all('a')[i].get('href') == 'javascript
            cin = soup.find('div', class_="sect-city").find_all('a')[i].string
            continue
        city.append(cin)
        cinema.append(soup.find('div', class_="sect-city").find_all('a')[i].string)
        df1 = pd.DataFrame(cinema, columns=['Cinema'])
        df1['City'] = city
    return df1

def cinema_df(soup, day):
    """
    This function create cinema and its metadata
    Input: soup, date of the movie
    Output: dataframe
    """
    cinema = []
    date = []
    movie = []
    time = []

    active = soup.find_all('div', class_='schedule-container')[0].select('.active')
    disable = soup.find_all('div', class_='schedule-container')[0].select('.disabled')

    for i in range(len(active)):
        cinema.append(soup.find('div', class_='cinema-info-body').h4.string)
        date.append(day)
        movie.append(active[i].get('attr-mov'))
        time.append(active[i].string)

    for i in range(len(disable)):
        cinema.append(soup.find('div', class_='cinema-info-body').h4.string)
        date.append(day)
        movie.append(disable[i].get('attr-mov'))
        time.append(disable[i].string)

    df2 = pd.DataFrame(cinema, columns=['Cinema'])
    df2['Date'] = date
    df2['Title'] = movie # There is a space before the title name
    df2['Time'] = time
    return df2
```

In [12]:

```
# Get Cinema ID
url = "https://www.cgv.id/en/schedule/cinema/"
request = urllib.request.Request(url)
page = urllib.request.urlopen(request)
soup = BeautifulSoup(page,"lxml")

ID = []
x = soup.find('div', class_="sect-city").find_all('a')
for i in range(len(x)):
    ID.append(x[i].get('id'))
print(ID)
```

```
[None, '002', '003', '004', '006', '020', '025', '028', '035', '037', None,
'001', '011', '014', '029', '038', '0', '008', '1', '009', '010', '2', '00
7', '036', '040', '3', '016', '042', '4', '030', '5', '019', '046', '6', '04
3', '7', '044', '8', '021', '9', '034', '10', '024', '11', '023', '12', '04
7', '051', '13', '027', '039', '14', '033', '15', '026', '16', '041', '17',
'018', '048', '18', '005', '015', '022', '19', '045', '20', '032', '21', '01
3', '017', '031']
```

In [13]:

```
# Create city (df1) and cinema (df2) dataframe
# Here, we cheat a little, remove None and non 3 digit numbers
cin_no = ['002', '003', '004', '006', '020', '025', '028', '035', '037', '001', '011', '014
          '009', '010', '007', '036', '040', '016', '042', '030', '019', '046', '043', '044
          '023', '047', '051', '027', '039', '033', '026', '041', '018', '048', '005', '015
          '21', '013', '017', '031']
date = ['2018-02-19', '2018-02-20', '2018-02-21', '2018-02-22', '2018-02-23']
df1 = city_cinema_df(soup)
key = []
dict_df2 = {}
for no in cin_no:
    for day in date:
        url = 'https://www.cgv.id/en/schedule/cinema/'+no+'/'+day
        request = urllib.request.Request(url)
        page = urllib.request.urlopen(request)
        soup = BeautifulSoup(page,"lxml")
        name = no+' '+day
        key.append(name)
        dict_df2[name] = cinema_df(soup, day)

df2 = dict_df2[key[0]]
for i in range(1, len(key)):
    df2 = df2.append(dict_df2[key[i]], ignore_index=True)
```

In [14]:

```
# Missing value due to disabled value
df2.isnull().sum()
```

Out[14]:

```
Cinema     0
Date       0
Title     70
Time       0
dtype: int64
```

Note on 22-Feb-2018, 2 movie detail is deleted as they are 2 new movie included.

- Padmaaval https://www.cgv.id/en/movies/info/18002800 (https://www.cgv.id/en/movies/info/18002800)
- Padman https://www.cgv.id/en/movies/info/18003700 (https://www.cgv.id/en/movies/info/18003700)

In [37]:

```
# Create movie (df3) dataframe
# movie link
movie_no = ['17009800', '18000800', '17007400', '18003600', '18004500',
            '18004600','18002100','18004700', '18001400', '17009700',
            '18001600', '18002000', '18001500', '18000700', '18002600',
            '18000600', '18004300', '18003400']

def extract(string):
    # pick only the necessary part of string
    no = string.find(' : ')
    return string[no+3:]

title = []
link = []
star = []
director = []
censor = []
genre = []
language = []
subtitle = []
duration = []

for no in movie_no:
    url = "https://www.cgv.id/en/movies/info/"+no
    request = urllib.request.Request(url)
    page = urllib.request.urlopen(request)
    soup = BeautifulSoup(page,"lxml")

    x = soup.find('div', class_='movie-add-info left').find_all('li')
    if len(x)==6:
        """
        Note this if is created due to some movies does not have director mentioned:
        Example: Bunda (https://www.cgv.id/en/movies/info/18002600) and
        London Love Story (https://www.cgv.id/en/movies/info/18000600)
        """
        title.append(soup.find('div', class_='movie-info-title').string.replace('\r','').re
        link.append('https://www.cgv.id'+soup.find('div', class_="poster-section left").fir
        star.append(extract(x[0].string))
        director.append('N.A')
        censor.append(extract(x[1].string))
        genre.append(extract(x[2].string))
        language.append(extract(x[3].string))
        subtitle.append(extract(x[4].string))
        duration.append(extract(x[5].string))
        continue

    # image URL
    title.append(soup.find('div', class_='movie-info-title').string.replace('\r','').replac
    link.append('https://www.cgv.id'+soup.find('div', class_="poster-section left").find('i
    star.append(extract(x[0].string))
    director.append(extract(x[1].string))
    censor.append(extract(x[2].string))
    genre.append(extract(x[3].string))
    language.append(extract(x[4].string))
    subtitle.append(extract(x[5].string))
    duration.append(extract(x[6].string))

df3 = pd.DataFrame(title, columns=['Title'])
df3['Image_Link'] = link
```

```
df3['Starring'] = star

df3['Director'] = director
df3['Censor'] = censor
df3['Genre'] = genre
df3['Language'] = language
df3['Subtitle'] = subtitle
df3['Duration'] = duration
```

3. Our DataFrame

In [38]:

```
df1.head()
```

Out[38]:

	Cinema	City
0	Grand Indonesia	Jakarta
1	Pacific Place	Jakarta
2	Mall of Indonesia	Jakarta
3	Central Park	Jakarta
4	Slipi Jaya	Jakarta

In [39]:

```
df2.head()
```

Out[39]:

	Cinema	Date	Title	Time
0	Grand Indonesia	2018-02-19	SCREENX 2D BLACK PANTHER	10:30
1	Grand Indonesia	2018-02-19	SCREENX 2D BLACK PANTHER	13:30
2	Grand Indonesia	2018-02-19	SCREENX 2D BLACK PANTHER	16:20
3	Grand Indonesia	2018-02-19	SCREENX 2D BLACK PANTHER	19:15
4	Grand Indonesia	2018-02-19	SCREENX 2D BLACK PANTHER	22:10

In [40]:

```
df3.head()
```

Out[40]:

	Title	Image_Link	Starring	Director	Censor	
0	DILAN 1990	https://www.cgv.id/uploads/movie/compressed/17...	Iqbaal Ramadhan , Vanesha Prescilla , Giulio P...	Fajar Bustomi, Pidi Baiq	13+	
1	DOWNSIZING	https://www.cgv.id/uploads/movie/compressed/18...	Matt Damon , Kristen Wiig , Jason Sudeikis	Alexander Payne	13+	C
2	DEN OF THIEVES	https://www.cgv.id/uploads/movie/compressed/17...	Gerard Butler, Pablo Schriber, Curtis '50 Cent...	Christian Gudegast	17+	
3	24 HOURS TO LIVE	https://www.cgv.id/uploads/movie/compressed/18...	Ethan Hawke, Rutger Hauer, Nathalie Boltt	Brian Smrz	17+	
4	MONSTER HUNT 2	https://www.cgv.id/uploads/movie/compressed/18...	Raman Hui	Tony Chiu-Wai Leung, Baihe Bai, Boran Jing	13+	F

The Title in df2 have space in front. We need to remove the space. To do that, first we will fill the missing value. Then loop the whole value.

In [41]:

```
df2['Title'] = df2.Title.fillna(' N.A.')
df2['Title'] = [x[1:] for x in df2.Title.values]
# Use Merge, Join to combine the dataframe
df = pd.merge(df2, df1, on=['Cinema'], how='left')
df = pd.merge(df, df3, on=['Title'], how='left')
df.head()
```

Out[41]:

	Cinema	Date	Title	Time	City	Image_Link	Starring	Director	Censor	Genre	I
0	Grand Indonesia	2018-02-19	CREENX 2D BLACK PANTHER	10:30	Jakarta	NaN	NaN	NaN	NaN	NaN	
1	Grand Indonesia	2018-02-19	CREENX 2D BLACK PANTHER	13:30	Jakarta	NaN	NaN	NaN	NaN	NaN	
2	Grand Indonesia	2018-02-19	CREENX 2D BLACK PANTHER	16:20	Jakarta	NaN	NaN	NaN	NaN	NaN	
3	Grand Indonesia	2018-02-19	CREENX 2D BLACK PANTHER	19:15	Jakarta	NaN	NaN	NaN	NaN	NaN	
4	Grand Indonesia	2018-02-19	CREENX 2D BLACK PANTHER	22:10	Jakarta	NaN	NaN	NaN	NaN	NaN	

In [42]:

```
# export the dataframe
df.to_csv(r'C:\Users\LW130003\cgv.csv', index=False)
```