In [1]:
```
import pandas as pd
```

In [42]:
```
IMDB = 'IMDB'
REVIEWS = 'Reviews'
PLOT_DESCRIPTIONS = 'Plot_Descriptions'
```

In [43]:
```
NAME_BASICS = IMDB + '/name_basics.tsv'
TITLE_AKAS = IMDB + '/title_akas.tsv'
TITLE_BASICS = IMDB + '/title_basics.tsv'
TITLE_CREW = IMDB + '/title_crew.tsv'
TITLE_EPISODES = IMDB + '/title_episodes.tsv'
TITLE_PRINCIPALS = IMDB + '/title_principals.tsv'
TITLE_RATINGS = IMDB + '/title_ratings.tsv'
TITLE_REVIEWS = REVIEWS + '/title_reviews.csv'
PLOTS = PLOT_DESCRIPTIONS + '/IMDB_movie_details.json'
```

In [11]:
```
df = pd.read_csv(NAME_BASICS, sep='\t')
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10742756 entries, 0 to 10742755
Data columns (total 6 columns):
 #   Column             Dtype 
---  ------             ----- 
 0   nconst             object
 1   primaryName        object
 2   birthYear          object
 3   deathYear          object
 4   primaryProfession  object
 5   knownForTitles     object
dtypes: object(6)
memory usage: 491.8+ MB
```

In [12]:
```
df.head
```

Out[12]:
```
<bound method NDFrame.head of              nconst         primaryName birthYear
deathYear  \
0         nm0000001        Fred Astaire      1899      1987
1         nm0000002       Lauren Bacall      1924      2014
2         nm0000003     Brigitte Bardot      1934        \N
3         nm0000004        John Belushi      1949      1982
4         nm0000005      Ingmar Bergman      1918      2007
...             ...                 ...       ...       ...
10742751  nm9993714   Romeo del Rosario        \N        \N
10742752  nm9993716       Essias Loberg        \N        \N
10742753  nm9993717  Harikrishnan Rajan        \N        \N
10742754  nm9993718         Aayush Nair        \N        \N
10742755  nm9993719          Andre Hill        \N        \N

                            primaryProfession  \
0             soundtrack,actor,miscellaneous   
1                         actress,soundtrack   
2         actress,soundtrack,music_department   
3                    actor,soundtrack,writer   
4                      writer,director,actor   
...                                       ...   
10742751  animation_department,art_department   
10742752                                  NaN   
10742753                       cinematographer   
10742754                       cinematographer   
10742755                                  NaN   
```

```
                                   knownForTitles
0         tt0053137,tt0031983,tt0072308,tt0050419
1         tt0037382,tt0038355,tt0075213,tt0117057
2         tt0049189,tt0056404,tt0054452,tt0057345
3         tt0078723,tt0072562,tt0077975,tt0080455
4         tt0050986,tt0050976,tt0069467,tt0060827
...                                           ...
10742751                                tt2455546
10742752                                       \N
10742753                                tt8736744
10742754                                       \N
10742755                                       \N

[10742756 rows x 6 columns]>
```

In [13]:
```
df = pd.read_csv(TITLE_AKAS, sep='\t')
df.info()
```

```
/Users/jon/opt/anaconda3/lib/python3.8/site-packages/IPython/core/interactiveshe
ll.py:3146: DtypeWarning: Columns (7) have mixed types.Specify dtype option on i
mport or set low_memory=False.
  has_raised = await self.run_ast_nodes(code_ast.body, cell_name,
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25262036 entries, 0 to 25262035
Data columns (total 8 columns):
 #   Column           Dtype 
---  ------           ----- 
 0   titleId          object
 1   ordering         int64 
 2   title            object
 3   region           object
 4   language         object
 5   types            object
 6   attributes       object
 7   isOriginalTitle  object
dtypes: int64(1), object(7)
memory usage: 1.5+ GB
```

In [14]:
```
df.head
```

Out[14]:
```
<bound method NDFrame.head of             titleId  ordering
title region language  \
0         tt0000001         1                 Карменсіта     UA       \N
1         tt0000001         2                 Carmencita     DE       \N
2         tt0000001         3  Carmencita - spanyol tánc     HU       \N
3         tt0000001         4                 Καρμενσίτα     GR       \N
4         tt0000001         5                 Карменсита     RU       \N
...             ...       ...                        ...    ...      ...
25262031  tt9916852         4                エピソード #3.20     JP       ja
25262032  tt9916852         5             Episódio #3.20     PT       pt
25262033  tt9916852         6             Episodio #3.20     IT       it
25262034  tt9916852         7               एपिसोड #3.20     IN       hi
25262035  tt9916856         1                   The Wind     DE       \N

                types     attributes isOriginalTitle
0         imdbDisplay             \N               0
1                  \N  literal title               0
2         imdbDisplay             \N               0
3         imdbDisplay             \N               0
4         imdbDisplay             \N               0
...               ...            ...             ...
25262031           \N             \N               0
25262032           \N             \N               0
```

```
25262033           \N             \N                0
25262034           \N             \N                0
25262035           \N             \N                0

[25262036 rows x 8 columns]>
```

In [16]:

```
df = pd.read_csv(TITLE_BASICS, sep='\t')
df.info()
```

```
/Users/jon/opt/anaconda3/lib/python3.8/site-packages/IPython/core/interactiveshe
ll.py:3146: DtypeWarning: Columns (4,5) have mixed types.Specify dtype option on
import or set low_memory=False.
  has_raised = await self.run_ast_nodes(code_ast.body, cell_name,
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7650595 entries, 0 to 7650594
Data columns (total 9 columns):
 #   Column          Dtype 
---  ------          ----- 
 0   tconst          object
 1   titleType       object
 2   primaryTitle    object
 3   originalTitle   object
 4   isAdult         object
 5   startYear       object
 6   endYear         object
 7   runtimeMinutes  object
 8   genres          object
dtypes: object(9)
memory usage: 525.3+ MB
```

In [17]:

```
df.head
```

Out[17]:

```
<bound method NDFrame.head of             tconst  titleType               primar
yTitle  \
0        tt0000001      short                Carmencita
1        tt0000002      short     Le clown et ses chiens
2        tt0000003      short             Pauvre Pierrot
3        tt0000004      short                Un bon bock
4        tt0000005      short           Blacksmith Scene
...            ...        ...                        ...
7650590  tt9916848  tvEpisode              Episode #3.17
7650591  tt9916850  tvEpisode              Episode #3.19
7650592  tt9916852  tvEpisode              Episode #3.20
7650593  tt9916856      short                   The Wind
7650594  tt9916880  tvEpisode  Horrid Henry Knows It All

                     originalTitle isAdult startYear endYear runtimeMinutes  \
0                       Carmencita       0      1894      \N              1
1           Le clown et ses chiens       0      1892      \N              5
2                   Pauvre Pierrot       0      1892      \N              4
3                      Un bon bock       0      1892      \N             12
4                 Blacksmith Scene       0      1893      \N              1
...                            ...     ...       ...     ...            ...
7650590              Episode #3.17       0      2010      \N             \N
7650591              Episode #3.19       0      2010      \N             \N
7650592              Episode #3.20       0      2010      \N             \N
7650593                   The Wind       0      2015      \N             27
7650594  Horrid Henry Knows It All       0      2014      \N             10

                           genres
0               Documentary,Short
1                 Animation,Short
2        Animation,Comedy,Romance
3                 Animation,Short
```

```
4                     Comedy,Short
...                            ...
7650590       Action,Drama,Family
7650591       Action,Drama,Family
7650592       Action,Drama,Family
7650593                      Short
7650594   Animation,Comedy,Family

[7650595 rows x 9 columns]>
```

In [18]:

```
df = pd.read_csv(TITLE_CREW, sep='\t')
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7650595 entries, 0 to 7650594
Data columns (total 3 columns):
 #   Column     Dtype 
---  ------     ----- 
 0   tconst     object
 1   directors  object
 2   writers    object
dtypes: object(3)
memory usage: 175.1+ MB
```

In [19]:

```
df.head
```

Out[19]:

```
<bound method NDFrame.head of              tconst            directors  
writers
0        tt0000001            nm0005690                              \N
1        tt0000002            nm0721526                              \N
2        tt0000003            nm0721526                              \N
3        tt0000004            nm0721526                              \N
4        tt0000005            nm0005690                              \N
...            ...                  ...                             ...
7650590  tt9916848  nm5519454,nm5519375  nm6182221,nm1628284,nm2921377
7650591  tt9916850  nm5519375,nm5519454  nm6182221,nm1628284,nm2921377
7650592  tt9916852  nm5519375,nm5519454  nm6182221,nm1628284,nm2921377
7650593  tt9916856           nm10538645                       nm6951431
7650594  tt9916880            nm0996406             nm1482639,nm2586970

[7650595 rows x 3 columns]>
```

In [20]:

```
df = pd.read_csv(TITLE_EPISODES, sep='\t')
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5552120 entries, 0 to 5552119
Data columns (total 4 columns):
 #   Column         Dtype 
---  ------         ----- 
 0   tconst         object
 1   parentTconst   object
 2   seasonNumber   object
 3   episodeNumber  object
dtypes: object(4)
memory usage: 169.4+ MB
```

In [21]:

```
df.head
```

Out[21]:

```
<bound method NDFrame.head of              tconst parentTconst seasonNumber episo
deNumber
0        tt0041951    tt0041038            1             9
1        tt0042816    tt0989125            1            17
2        tt0042889    tt0989125           \N            \N
```

```
3        tt0043426    tt0040051           3             42
4        tt0043631    tt0989125           2             16
...            ...          ...         ...            ...
5552115  tt9916846    tt1289683           3             18
5552116  tt9916848    tt1289683           3             17
5552117  tt9916850    tt1289683           3             19
5552118  tt9916852    tt1289683           3             20
5552119  tt9916880    tt0985991           4              2

[5552120 rows x 4 columns]>
```

```
In [22]: df = pd.read_csv(TITLE_PRINCIPALS, sep='\t')
         df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 43236547 entries, 0 to 43236546
Data columns (total 6 columns):
 #   Column      Dtype 
---  ------      ----- 
 0   tconst      object
 1   ordering    int64 
 2   nconst      object
 3   category    object
 4   job         object
 5   characters  object
dtypes: int64(1), object(5)
memory usage: 1.9+ GB
```

```
In [23]: df.head
```

```
Out[23]: <bound method NDFrame.head of              tconst  ordering     nconst          c
ategory  \
0         tt0000001         1  nm1588970             self
1         tt0000001         2  nm0005690         director
2         tt0000001         3  nm0374658  cinematographer
3         tt0000002         1  nm0721526         director
4         tt0000002         2  nm1335271         composer
...             ...       ...        ...              ...
43236542  tt9916880         5  nm0996406         director
43236543  tt9916880         6  nm1482639           writer
43236544  tt9916880         7  nm2586970           writer
43236545  tt9916880         8  nm1594058         producer
43236546  tt9916880         9  nm1052583          actress

                               job                characters
0                               \N                  ["Self"]
1                               \N                        \N
2          director of photography                        \N
3                               \N                        \N
4                               \N                        \N
...                            ...                       ...
43236542        principal director                        \N
43236543                        \N                        \N
43236544                     books                        \N
43236545                  producer                        \N
43236546                        \N  ["Mum","Tidy Ted","Fang"]

[43236547 rows x 6 columns]>
```

```
In [24]: df = pd.read_csv(TITLE_RATINGS, sep='\t')
         df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1125659 entries, 0 to 1125658
Data columns (total 3 columns):
 #   Column          Non-Null Count    Dtype  
---  ------          --------------    -----  
 0   tconst          1125659 non-null  object 
 1   averageRating   1125659 non-null  float64
 2   numVotes        1125659 non-null  int64  
dtypes: float64(1), int64(1), object(1)
memory usage: 25.8+ MB
```

In [25]:
```
df.head
```

Out[25]:
```
<bound method NDFrame.head of              tconst  averageRating  numVotes
0        tt0000001            5.7      1686
1        tt0000002            6.0       208
2        tt0000003            6.5      1424
3        tt0000004            6.1       122
4        tt0000005            6.1      2223
...            ...            ...       ...
1125654  tt9916580            7.2         5
1125655  tt9916690            6.6         5
1125656  tt9916720            6.2        72
1125657  tt9916766            6.9        16
1125658  tt9916778            7.4        26

[1125659 rows x 3 columns]>
```

In [34]:
```
df = pd.read_csv(TITLE_REVIEWS, sep=',')
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50000 entries, 0 to 49999
Data columns (total 2 columns):
 #   Column     Non-Null Count  Dtype 
---  ------     --------------  ----- 
 0   review     50000 non-null  object
 1   sentiment  50000 non-null  object
dtypes: object(2)
memory usage: 781.4+ KB
```

In [35]:
```
df.head
```

Out[35]:
```
<bound method NDFrame.head of 
review sentiment
0      One of the other reviewers has mentioned that ...  positive
1      A wonderful little production. <br /><br />The...  positive
2      I thought this was a wonderful way to spend ti...  positive
3      Basically there's a family where a little boy ...  negative
4      Petter Mattei's "Love in the Time of Money" is...  positive
...                                                  ...       ...
49995  I thought this movie did a down right good job...  positive
49996  Bad plot, bad dialogue, bad acting, idiotic di...  negative
49997  I am a Catholic taught in parochial elementary...  negative
49998  I'm going to have to disagree with the previou...  negative
49999  No one expects the Star Trek movies to be high...  negative

[50000 rows x 2 columns]>
```

In [51]:
```
df = pd.read_json(PLOTS,lines=True)
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1572 entries, 0 to 1571
Data columns (total 7 columns):
 #   Column          Non-Null Count  Dtype 
---  ------          --------------  ----- 
```

```
  0   movie_id       1572 non-null   object 
  1   plot_summary   1572 non-null   object 
  2   duration       1572 non-null   object 
  3   genre          1572 non-null   object 
  4   rating         1572 non-null   float64
  5   release_date   1572 non-null   object 
  6   plot_synopsis  1572 non-null   object 
 dtypes: float64(1), object(6)
 memory usage: 86.1+ KB
```

In [52]:

```
df.head
```

Out[52]:

```
<bound method NDFrame.head of          movie_id                                       
plot_summary  duration  \
0      tt0105112  Former CIA analyst, Jack Ryan is in England wi...  1h 57min   
1      tt1204975  Billy (Michael Douglas), Paddy (Robert De Niro...  1h 45min   
2      tt0243655  The setting is Camp Firewood, the year 1981. I...  1h 37min   
3      tt0040897  Fred C. Dobbs and Bob Curtin, both down on the...   2h 6min   
4      tt0126886  Tracy Flick is running unopposed for this year...  1h 43min   
...          ...                                                ...       ...   
1567   tt0289879  Evan Treborn grows up in a small town with his...  1h 53min   
1568   tt1723811  Brandon is a 30-something man living in New Yo...  1h 41min   
1569   tt5013056  Evacuation of Allied soldiers from the British...  1h 46min   
1570  tt0104014/  For a while now, beautiful 24-year-old Diana B...  1h 33min   
1571  tt0114142/  The marriage of David Burgess, a senior execut...  1h 32min   

                              genre  rating release_date  \
0             [Action, Thriller]     6.9   1992-06-05   
1                       [Comedy]     6.6   2013-11-01   
2              [Comedy, Romance]     6.7   2002-04-11   
3     [Adventure, Drama, Western]     8.3   1948-01-24   
4        [Comedy, Drama, Romance]     7.3   1999-05-07   
...                          ...     ...          ...   
1567          [Sci-Fi, Thriller]     7.7   2004-01-23   
1568                     [Drama]     7.2   2012-01-13   
1569     [Action, Drama, History]     8.1   2017-07-21   
1570             [Comedy, Drama]     5.3   1992-02-21   
1571          [Drama, Thriller]     4.0   1999-01-29   

                                          plot_synopsis  
0     Jack Ryan (Ford) is on a "working vacation" in...  
1     Four boys around the age of 10 are friends in ...  
2                                                        
3     Fred Dobbs (Humphrey Bogart) and Bob Curtin (T...  
4     Jim McAllister (Matthew Broderick) is a much-a...  
...                                                 ...  
1567  In the year 1998, Evan Treborn (Ashton Kutcher...  
1568  Brandon (Michael Fassbender) is a successful, ...  
1569  The film alternates between three different pe...  
1570                                                     
1571                                                     

[1572 rows x 7 columns]>
```

In [ ]: