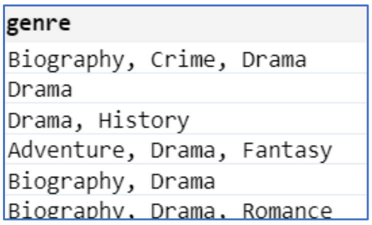
WS2: Cleaning problems, just some examples , *lets get dirty hands*

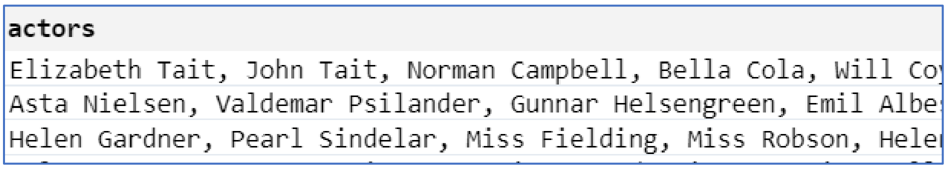
After reading in the IMDb movies.csv

Genre: unknown number but limited



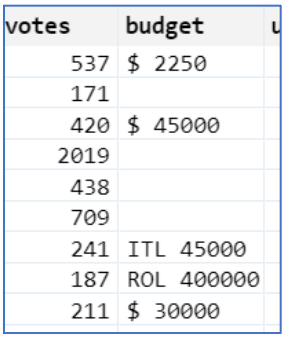
genre
Biography, Crime, Drama
Drama
Drama, History
Adventure, Drama, Fantasy
Biography, Drama
Biography, Drama, Romance

Actors: unknown number, no limits



actors
Elizabeth Tait, John Tait, Norman Campbell, Bella Cola, Will Co
Asta Nielsen, Valdemar Psilander, Gunnar Helsengreen, Emil Albe
Helen Gardner, Pearl Sindelar, Miss Fielding, Miss Robson, Hele

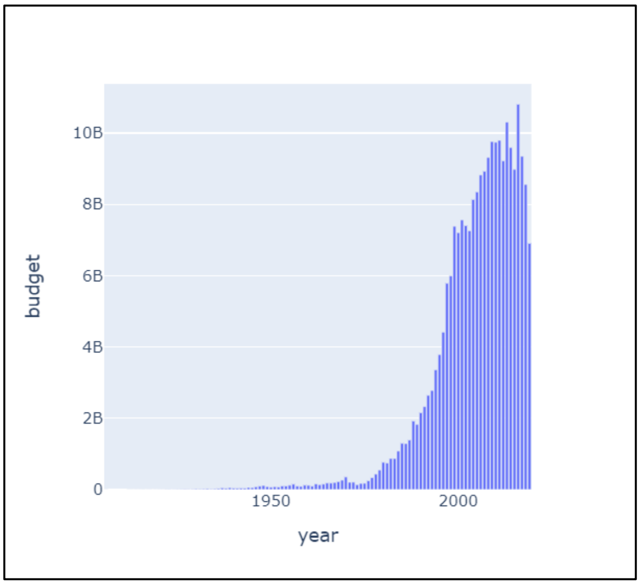
Budget: different valuta, and no numbers



votes	budget	u
537	$ 2250	
171		
420	$ 45000	
2019		
438		
709		
241	ITL 45000	
187	ROL 400000	
211	$ 30000	

Exercise 1

1. Read the file IMDB.csv into a pandas dataframe
2. Select only rows with USA $ sign
3. Remove the $ sign
4. Make the type numerical
5. Make a (bar) plot where for each year you could see the total budget



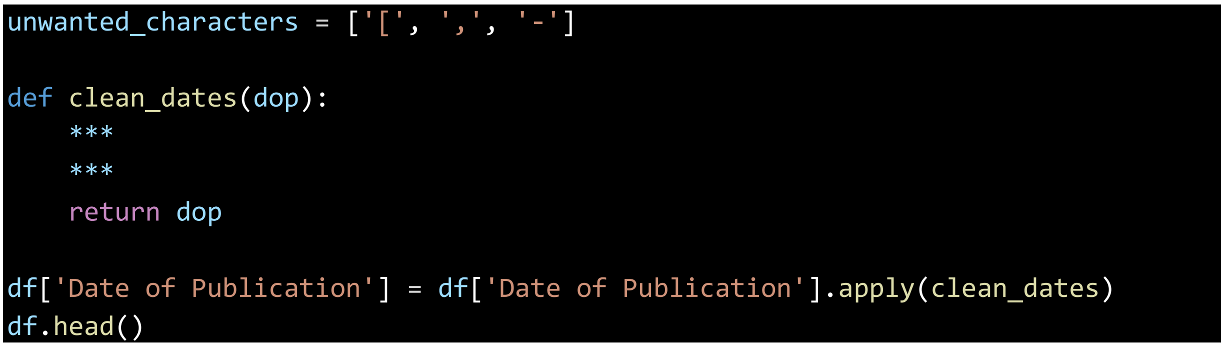
6. Split ‘genre’ in different columns
7. Add a column ‘numberOfActors’ containing the number of actors mentioned in the column ‘actors’
8. Add a column ‘mainActor’ containing only the first mentioned actor in the column ‘actors’

Exercise 2

1. Read the file BL-Flickr-Images-Book.csv into a pandas dataframe
2. Look at the 5 first rows
3. Look at the number of columns
4. We want to keep only the columns

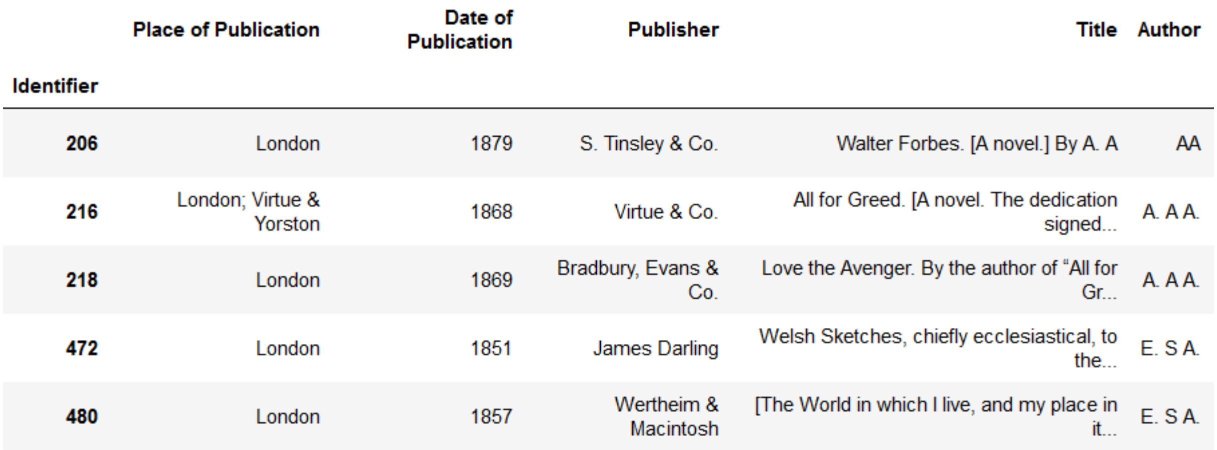
   `['Identifier', 'Place of Publication', 'Date of Publication', 'Publisher', 'Title', 'Author', 'Flickr URL']`

5. Look at the first 25 rows of `'Date of Publication'`. Use a function to clean up in the following way



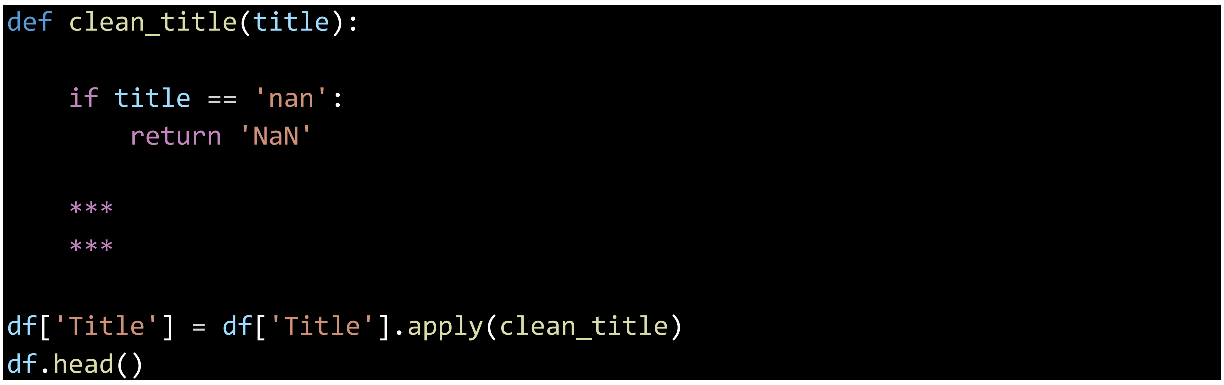
```
unwanted_characters = ['[', ',', '-']

def clean_dates(dop):
    ***
    ***
    return dop

df['Date of Publication'] = df['Date of Publication'].apply(clean_dates)
df.head()
```

it should look like



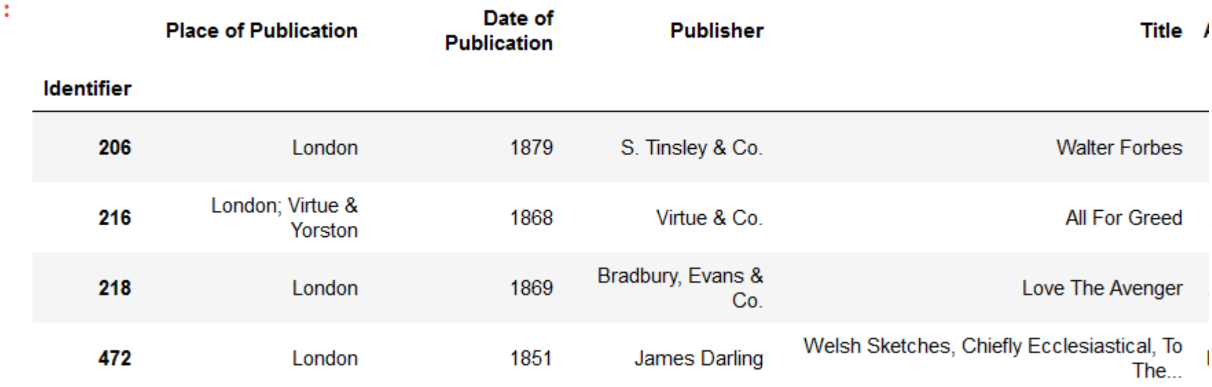
	Place of Publication	Date of Publication	Publisher	Title	Author
Identifier					
206	London	1879	S. Tinsley & Co.	Walter Forbes. [A novel.] By A. A	AA
216	London; Virtue & Yorston	1868	Virtue & Co.	All for Greed. [A novel. The dedication signed...	A. A A.
218	London	1869	Bradbury, Evans & Co.	Love the Avenger. By the author of “All for Gr...	A. A A.
472	London	1851	James Darling	Welsh Sketches, chiefly ecclesiastical, to the...	E. S A.
480	London	1857	Wertheim & Macintosh	[The World in which I live, and my place in it...	E. S A.

6. Use a function to clean up the column ‘Title’



```
def clean_title(title):

    if title == 'nan':
        return 'NaN'

    ***
    ***

df['Title'] = df['Title'].apply(clean_title)
df.head()
```

It should look like



	Place of Publication	Date of Publication	Publisher	Title
Identifier				
206	London	1879	S. Tinsley & Co.	Walter Forbes
216	London; Virtue & Yorston	1868	Virtue & Co.	All For Greed
218	London	1869	Bradbury, Evans & Co.	Love The Avenger
472	London	1851	James Darling	Welsh Sketches, Chiefly Ecclesiastical, To The...