Data Cleaning With pandas

Muhammad Huzaifa

- Here we have an uncleaned data set about movies, we will try to clean it step by step.

	MOVIES	YEAR	GENRE	RATING	
0	Blood Red Sky	(2021)	\nAction, Horror, Thriller	6.1	\nA w
1	Masters of the Universe: Revelation	(2021–)	\nAnimation, Action, Adventure	5.0	\nT
2	The Walking Dead	(2010–2022)	\nDrama, Horror, Thriller	8.2	\nSh
3	Rick and Morty	(2013–)	\nAnimation, Adventure, Comedy	9.2	
4	Army of Thieves	(2021)	\nAction, Crime, Horror	NaN	\nA p
5	Outer Banks	(2020–)	\nAction, Crime, Drama	7.6	\nA gr
6	The Last Letter from Your Lover	(2021)	\nDrama, Romance	6.8	\nA pa
7	Dexter	(2006–2013)	\nCrime, Drama, Mystery	8.6	\nB
8	Never Have I Ever	(2020–)	\nComedy	7.9	\nT
9	Virgin River	(2019–)	\nDrama, Romance	7.4	\n

- Importing the libraries

```
[117] import pandas as pd
```

- Loading our movies dataset

```
[100] df = pd.read_csv('/content/movies.csv')
```

- Renaming the columns properly

```
[104] df = df.rename(columns={'MOVIES': 'Movies',
                              'YEAR': 'Year',
                              'GENRE':'Genre',
                              'RATING': 'Rating',
                              'ONE-LINE': 'One-Line',
                              'STARS': 'Stars',
                              'VOTES': 'Votes',
                              'Runtime': 'Runtime',
                              'Gross': 'Gross'})
```

- Dropping the duplicate & NAN values

```
[104] df = df.rename(columns={'MOVIES': 'Movies',
                              'YEAR': 'Year',
                              'GENRE':'Genre',
                              'RATING': 'Rating',
                              'ONE-LINE': 'One-Line',
                              'STARS': 'Stars',
                              'VOTES': 'Votes',
                              'Runtime': 'Runtime',
                              'Gross': 'Gross'})
```

- Cleaning all the columns

```
df['Year'] = df['Year'].str.replace('(', '')
```

```
[107] df['Year'] = df['Year'].str.replace('\n', '')
```

```
[108] df['Year'] = df['Year'].str.replace(')', '')
```

```
[109] df['Year'] = df['Year'].str.replace('I', '')
```

```
[110] df['Stars'] = df['Stars'].str.replace('\n', '')
```

- Dropping the columns that we don't need

```
df = df.drop(['One-Line', 'Stars'], axis=1)
```

- Resetting our index

```
df = df.reset_index()
```

```
[115] df = df.drop(['index'], axis=1)
      df.index += 1
```

- Here is our final cleaned dataset.

	Movies	Year	Genre	Rating	Votes	RunTime	Gross
1	The Hitman's Bodyguard	2017	Action, Comedy, Crime	6.9	205979	118.0	$75.47M
2	Jurassic Park	1993	Action, Adventure, Sci-Fi	8.1	897444	127.0	$402.45M
3	Don't Breathe	2016	Crime, Horror, Thriller	7.1	237601	88.0	$89.22M
4	The Lord of the Rings: The Fellowship of the Ring	2001	Action, Adventure, Drama	8.8	1713028	178.0	$315.54M
5	Escape Room	2019	Action, Adventure, Horror	6.4	99351	99.0	$57.01M
...	...	...	...	...	...	...	...
456	Vidal Sassoon: The Movie	2010	Documentary	6.5	245	90.0	$0.09M
457	Men at Lunch	2012	Documentary, Mystery	6.3	331	75.0	$0.00M
458	Decoding Deepak	2012	Documentary	5.5	124	83.0	$0.01M
459	Theo Who Lived	2016	Documentary	6.8	111	86.0	$0.01M
460	Southern Justice	2006	Action, Adventure, Thriller	3.1	126	96.0	$0.14M

460 rows × 7 columns