Importing Libraries ¶

In [21]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

In [2]:

```
#Loading  and reading the data

ratings = pd.read_csv('./ratings_small.csv')
movies = pd.read_csv('./movies_metadata.csv')
```

In [3]:

```
ratings.head()
```

Out[3]:

	userId	movieId	rating	timestamp
0	1	31	2.5	1260759144
1	1	1029	3.0	1260759179
2	1	1061	3.0	1260759182
3	1	1129	2.0	1260759185
4	1	1172	4.0	1260759205

```
ratings.info()
```

In [10]:

```
movies.head()
```

Out[10]:

	adult	belongs_to_collection	budget	genres	homepage	id	ir
0	False	{'id': 10194, 'name': 'Toy Story Collection', ...	30000000	[{'id': 16, 'name': 'Animation'}, {'id': 35, '...	http://toystory.disney.com/toy-story	862	tt0
1	False	NaN	65000000	[{'id': 12, 'name': 'Adventure'}, {'id': 14, '...	NaN	8844	tt0
2	False	{'id': 119050, 'name': 'Grumpy Old Men Collect...	0	[{'id': 10749, 'name': 'Romance'}, {'id': 35, ...	NaN	15602	tt0
3	False	NaN	16000000	[{'id': 35, 'name': 'Comedy'}, {'id': 18, 'nam...	NaN	31357	tt0
4	False	{'id': 96871, 'name': 'Father of the Bride Col...	0	[{'id': 35, 'name': 'Comedy'}]	NaN	11862	tt0

5 rows × 24 columns

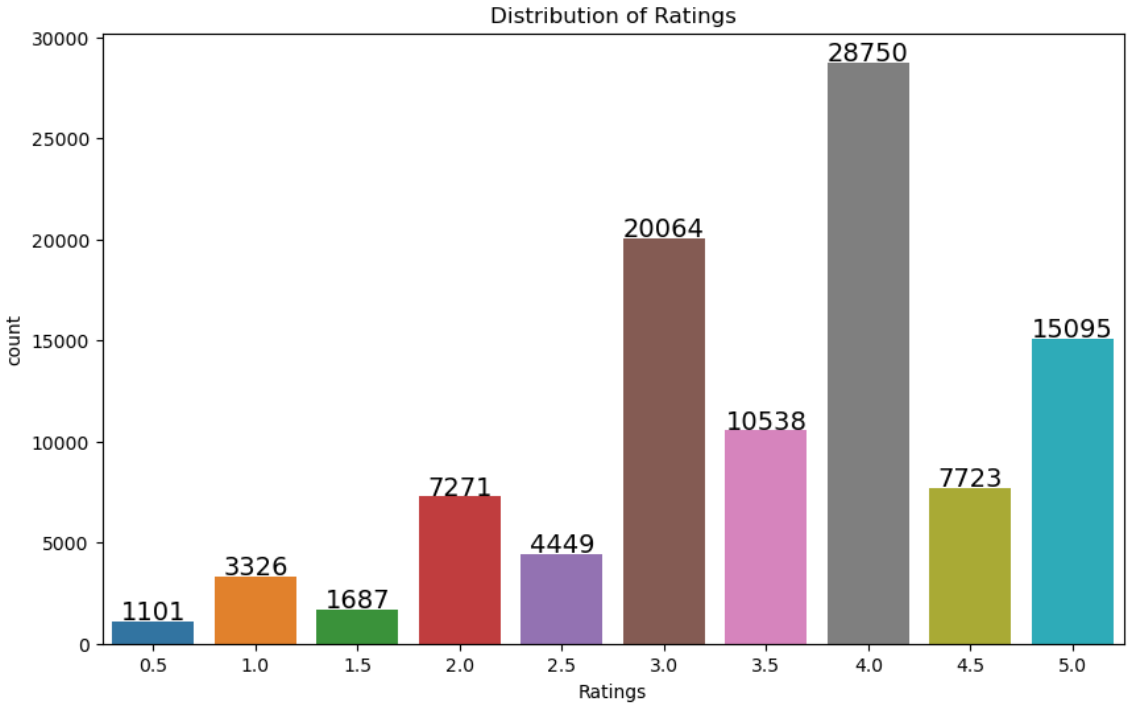
In [4]:

```
movies.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 45466 entries, 0 to 45465
Data columns (total 24 columns):
 #   Column                 Non-Null Count  Dtype  
---  ------                 --------------  -----  
 0   adult                  45466 non-null  object 
 1   belongs_to_collection  4494 non-null   object 
 2   budget                 45466 non-null  object 
 3   genres                 45466 non-null  object 
 4   homepage               7782 non-null   object 
 5   id                     45466 non-null  object 
 6   imdb_id                45449 non-null  object 
 7   original_language      45455 non-null  object 
 8   original_title         45466 non-null  object 
 9   overview               44512 non-null  object 
 10  popularity             45461 non-null  object 
 11  poster_path            45080 non-null  object 
 12  production_companies   45463 non-null  object 
 13  production_countries   45463 non-null  object 
 14  release_date           45379 non-null  object 
 15  revenue                45460 non-null  float64
 16  runtime                45203 non-null  float64
 17  spoken_languages       45460 non-null  object 
 18  status                 45379 non-null  object 
 19  tagline                20412 non-null  object 
 20  title                  45460 non-null  object 
 21  video                  45460 non-null  object 
 22  vote_average           45460 non-null  float64
 23  vote_count             45460 non-null  float64
dtypes: float64(4), object(20)
memory usage: 8.3+ MB
```

In [5]:

```
plt.figure(figsize = (10,6))
X = sns.countplot(data = ratings, x='rating')
labels = (ratings['rating'].value_counts().sort_index())
plt.title('Distribution of Ratings')
plt.xlabel('Ratings')
for i,v in enumerate(labels):
    X.text(i, v+100, str(v), size=14, color='k',
          horizontalalignment = 'center')
plt.show()
```



Cleaning data

In [6]:

```
t_mask = movies['title'].isnull()
```

In [7]:

```
movies = movies.loc[t_mask == False]
```

In [8]:

```
movies = movies.astype({'id':'int64'})
```

In [9]:

```
df = pd.merge(ratings, movies[['id','title']],
              left_on = 'movieId',
              right_on = 'id')
df.head()
```

Out[9]:

	userId	movieId	rating	timestamp	id	title
0	1	1371	2.5	1260759135	1371	Rocky III
1	4	1371	4.0	949810302	1371	Rocky III
2	7	1371	3.0	851869160	1371	Rocky III
3	19	1371	4.0	855193404	1371	Rocky III
4	21	1371	3.0	853852263	1371	Rocky III

In [10]:

```
df.drop(['timestamp','id'],axis=1, inplace=True)
```

In [11]:

```
df = df.drop_duplicates(['userId','title'])
```

In [12]:

```
dfp = df.pivot(index = 'userId', columns = 'title', values = 'rating').fillna(0)
```

In [14]:

```
dfp = dfp.astype('int64')
```

In [15]:

```
def encode_ratings(x):
    if x<=0:
        return 0
    if x>=1:
        return 1

dfp = dfp.applymap(encode_ratings)
```

In [16]:

```
dfp.head()
```

Out[16]:

vas the Night Before istmas	...And God Created Woman	00 Schneider - Jagd auf Nihil Baxter	10 Items or Less	10 Things I Hate About You	10,000 BC	11'09"01 - September 11	12 Angry Men	...	Zodiac	Zombie Flesh Eaters	Hc
0	0	0	0	0	0	0	0	...	0	0	
0	0	0	0	0	0	0	0	...	0	0	
0	0	0	0	0	0	0	0	...	0	0	
0	0	0	0	0	0	0	0	...	0	0	
0	0	0	0	0	0	0	0	...	0	0	

In [ ]: