Imporing - Cleaning - Exporting

In [3]:
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

In [4]:
```
#Importing Movie Titles, Ratings, Users
movies_df = pd.read_csv('../Data/movies.csv')
ratings_df = pd.read_csv('../Data/ratings.csv')
```

In [5]:
```
#Merging Movies and Ratings
full = movies_df.merge(ratings_df, how = 'right', on = 'movieId')
```

In [6]:
```
# Creating a Year column from movie titles and setting year as an integer
full['year'] = full.title.str.extract('(\d+)')
```

In [7]:
```
# Dropping Several null values created from extraction process
# setting year values as integers
full.dropna(inplace = True)
full.year = full.year.astype(int)
```

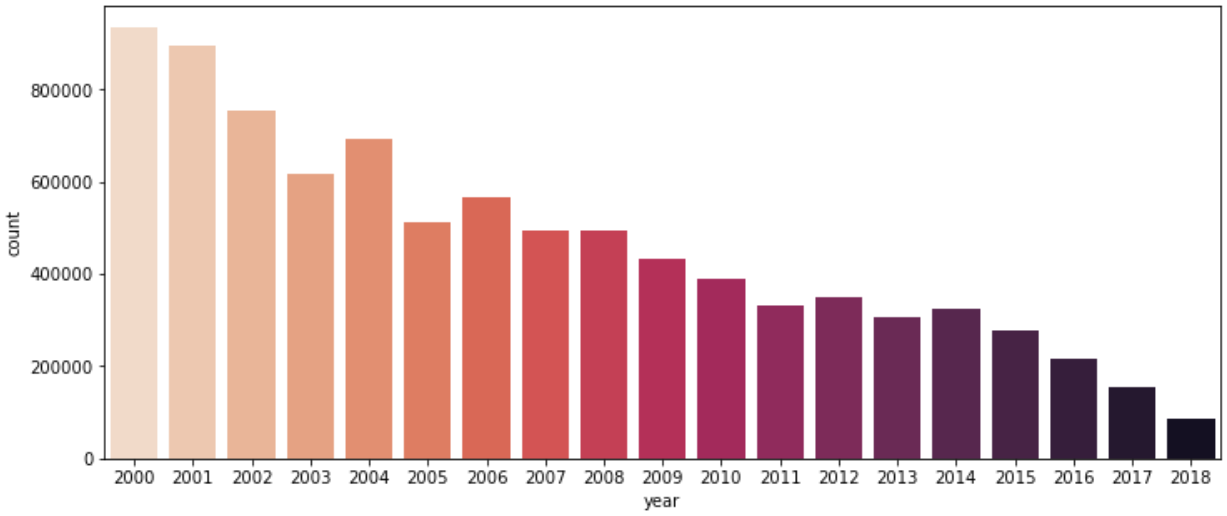
In [8]:
```
# Filtering for years between 2000 and 2018
full = full.loc[full.year >= 2000]
full = full.loc[full.year <= 2018]
```

In [9]:
```
# Getting rating frequency per movie
freq = full.groupby(full['movieId']).count()

# Creating a series of frequency counts
frequency = pd.Series(freq.year, name = 'frequency')

# merging frequency series into data
freq_df = full.merge(frequency, left_on = 'movieId', right_index = True)
```

In [10]:
```
# Filtering Movies with less than 10 reviews
filtered = freq_df.loc[freq_df['frequency'] >= 50]
```

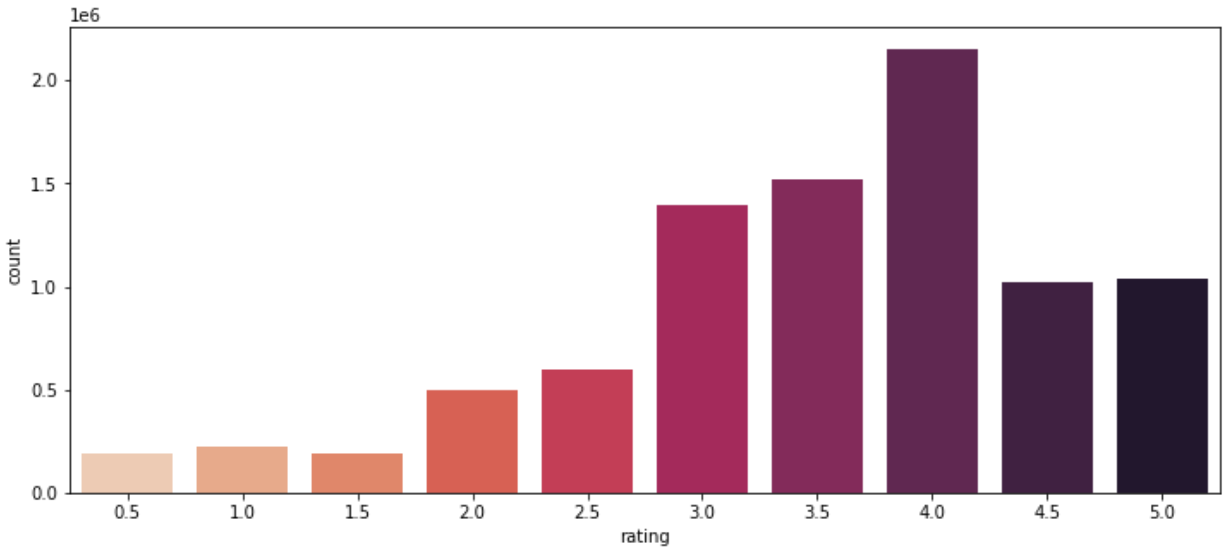
Year Distribution

In [11]:
```
years = full.year
plt.figure(figsize = (12, 5))
g = sns.countplot(x = years, palette = 'rocket_r');
```



Rating Distribution

In [12]:
```
ratings = full.rating
plt.figure(figsize = (12, 5))
r = sns.countplot(x = ratings, palette = 'rocket_r')
```



In [13]:
```
sample = filtered.sample(2000000)
```

In [14]:
```
sample.to_csv('../Data/filtered-cleaned')
```

In [ ]:
```

```