Importing Pandas

In [2]:
```
import pandas as pd
```

Importing Data

In [3]:
```
names = ['id', 'title', 'year', 'rating', 'votes', 'length', 'genres']
data = pd.read_csv('imdb_top_10000.txt', sep="\t", names=names, index_col=0)
```

Exploring our Data

In [12]:
```
data.head()
```

Out[12]:

id	title	year	rating	votes	length	genres
tt0111161	The Shawshank Redemption (1994)	1994	9.2	619479	142 mins.	Crime|Drama
tt0110912	Pulp Fiction (1994)	1994	9.0	490065	154 mins.	Crime|Thriller
tt0137523	Fight Club (1999)	1999	8.8	458173	139 mins.	Drama|Mystery|Thriller
tt0133093	The Matrix (1999)	1999	8.7	448114	136 mins.	Action|Adventure|Sci-Fi
tt1375666	Inception (2010)	2010	8.9	385149	148 mins.	Action|Adventure|Sci-Fi|Thriller

In [13]:
```
data.head(3)
```

Out[13]:

id	title	year	rating	votes	length	genres
tt0111161	The Shawshank Redemption (1994)	1994	9.2	619479	142 mins.	Crime|Drama
tt0110912	Pulp Fiction (1994)	1994	9.0	490065	154 mins.	Crime|Thriller
tt0137523	Fight Club (1999)	1999	8.8	458173	139 mins.	Drama|Mystery|Thriller

In [14]:
```
data.tail()
```

Out[14]:

id	title	year	rating	votes	length	genres
tt0807721	Meduzot (2007)	2007	7.0	1357	78 mins.	Drama
tt0339642	Daltry Calhoun (2005)	2005	5.2	1357	100 mins.	Comedy|Drama|Music|Romance
tt0060880	The Quiller Memorandum (1966)	1966	6.5	1356	104 mins.	Drama|Mystery|Thriller
tt0152836	Taal (1999)	1999	6.5	1356	179 mins.	Musical|Romance
tt0279977	The Navigators (2001)	2001	6.9	1356	96 mins.	Comedy|Drama

In [15]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 10000 entries, tt0111161 to tt0279977
Data columns (total 6 columns):
 #   Column  Non-Null Count  Dtype  
---  ------  --------------  -----  
 0   title   10000 non-null  object 
 1   year    10000 non-null  int64  
 2   rating  10000 non-null  float64
 3   votes   10000 non-null  int64  
 4   length  10000 non-null  object 
 5   genres  9999 non-null   object 
dtypes: float64(1), int64(2), object(3)
memory usage: 546.9+ KB
```

In [16]:
```
data.describe()
```

Out[16]:

	year	rating	votes
count	10000.000000	10000.000000	10000.000000
mean	1993.472800	6.386070	16604.012800
std	14.829924	1.189933	34563.459698
min	1950.000000	1.500000	1356.000000
25%	1986.000000	5.700000	2333.750000
50%	1998.000000	6.600000	4980.500000
75%	2005.000000	7.200000	15277.750000
max	2011.000000	9.200000	619479.000000

Exporting Data

In [17]:
```
data.to_csv('test.csv', header=True, index=True, sep=',')
```

Sorting Data

In [18]:
```
data.sort_values(by='rating')
```

Out[18]:

id	title	year	rating	votes	length	genres
tt0270846	Superbabies: Baby Geniuses 2 (2004)	2004	1.5	13196	88 mins.	Comedy|Family
tt0059464	Monster a-Go Go (1965)	1965	1.5	3255	70 mins.	Sci-Fi|Horror
tt0364986	Ben & Arthur (2002)	2002	1.5	4675	85 mins.	Drama|Romance
tt0421051	Daniel the Wizard (2004)	2004	1.5	8271	81 mins.	Comedy|Crime|Family|Fantasy|Horror
tt1309000	Dream Well (2009)	2009	1.5	2848	00 mins.	Comedy|Romance|Sport
...	...	...	...	...	...	...
tt0071562	The Godfather: Part II (1974)	1974	9.0	291169	200 mins.	Crime|Drama
tt0060196	The Good, the Bad and the Ugly (1966)	1966	9.0	195238	161 mins.	Western
tt0110912	Pulp Fiction (1994)	1994	9.0	490065	154 mins.	Crime|Thriller
tt0068646	The Godfather (1972)	1972	9.2	474189	175 mins.	Crime|Drama
tt0111161	The Shawshank Redemption (1994)	1994	9.2	619479	142 mins.	Crime|Drama

10000 rows × 6 columns

In [20]:
```
data.sort_values(by='rating', ascending=False)
```

Out[20]:

id	title	year	rating	votes	length	genres
tt0111161	The Shawshank Redemption (1994)	1994	9.2	619479	142 mins.	Crime|Drama
tt0068646	The Godfather (1972)	1972	9.2	474189	175 mins.	Crime|Drama
tt0060196	The Good, the Bad and the Ugly (1966)	1966	9.0	195238	161 mins.	Western
tt0110912	Pulp Fiction (1994)	1994	9.0	490065	154 mins.	Crime|Thriller
tt0252487	Outrageous Class (1975)	1975	9.0	9823	87 mins.	Comedy|Drama
...	...	...	...	...	...	...
tt0364986	Ben & Arthur (2002)	2002	1.5	4675	85 mins.	Drama|Romance
tt0060753	Night Train to Mundo Fine (1966)	1966	1.5	3542	89 mins.	Action|Adventure|Crime|War
tt0421051	Daniel the Wizard (2004)	2004	1.5	8271	81 mins.	Comedy|Crime|Family|Fantasy|Horror
tt0059464	Monster a-Go Go (1965)	1965	1.5	3255	70 mins.	Sci-Fi|Horror
tt0060666	Manos: The Hands of Fate (1966)	1966	1.5	20927	74 mins.	Horror

10000 rows × 6 columns

Creating Data Frames from Scratch

In [22]:
```
sample_data = {
    'tv': [230, 44, 17],
    'radio': [37, 39, 45],
    'news': [69, 45, 69],
    'sales': [22, 10, 9]
}
```

In [23]:
```
data2 = pd.DataFrame(sample_data)
```

In [24]:
```
data2
```

Out[24]:

	tv	radio	news	sales
0	230	37	69	22
1	44	39	45	10
2	17	45	69	9

Selecting Data

In [25]:
```
data['title']
```

Out[25]:
```
id
tt0111161    The Shawshank Redemption (1994)
tt0110912                  Pulp Fiction (1994)
tt0137523                    Fight Club (1999)
tt0133093                    The Matrix (1999)
tt1375666                     Inception (2010)
                              ...
tt0807721                        Meduzot (2007)
tt0339642                Daltry Calhoun (2005)
tt0060880      The Quiller Memorandum (1966)
tt0152836                             Taal (1999)
tt0279977                The Navigators (2001)
Name: title, Length: 10000, dtype: object
```

In [27]:
```
data[['title', 'year']]
```

Out[27]:

id	title	year
tt0111161	The Shawshank Redemption (1994)	1994
tt0110912	Pulp Fiction (1994)	1994
tt0137523	Fight Club (1999)	1999
tt0133093	The Matrix (1999)	1999
tt1375666	Inception (2010)	2010
...	...	...
tt0807721	Meduzot (2007)	2007
tt0339642	Daltry Calhoun (2005)	2005
tt0060880	The Quiller Memorandum (1966)	1966
tt0152836	Taal (1999)	1999
tt0279977	The Navigators (2001)	2001

10000 rows × 2 columns

In [29]:
```
data['rating'].mean()
```

Out[29]:
```
6.38607
```

In [30]:
```
data['rating'].max()
```

Out[30]:
```
9.2
```

In [31]:
```
data['rating'].min()
```

Out[31]:
```
1.5
```

In [32]:
```
data['genres'].unique()
```

Out[32]:
```
array(['Crime|Drama', 'Crime|Thriller', 'Drama|Mystery|Thriller', ...,
       'Drama|War|Adventure|Romance', 'Western|Sci-Fi|Thriller',
       'Adventure|Comedy|Drama|War'], dtype=object)
```

In [33]:
```
data['rating'].value_counts()
```

Out[33]:
```
7.1    401
6.8    401
7.2    386
6.7    384
7.0    382
      ...
1.5      7
1.6      6
8.9      5
9.0      4
9.2      2
Name: rating, Length: 77, dtype: int64
```

In [35]:
```
data['rating'].value_counts().sort_index()
```

Out[35]:
```
1.5     7
1.6     6
1.7    12
1.8    12
1.9     9
       ..
8.7    13
8.8     9
8.9     5
9.0     4
9.2     2
Name: rating, Length: 77, dtype: int64
```

In [36]:
```
data['rating'].value_counts().sort_index(ascending=False)
```

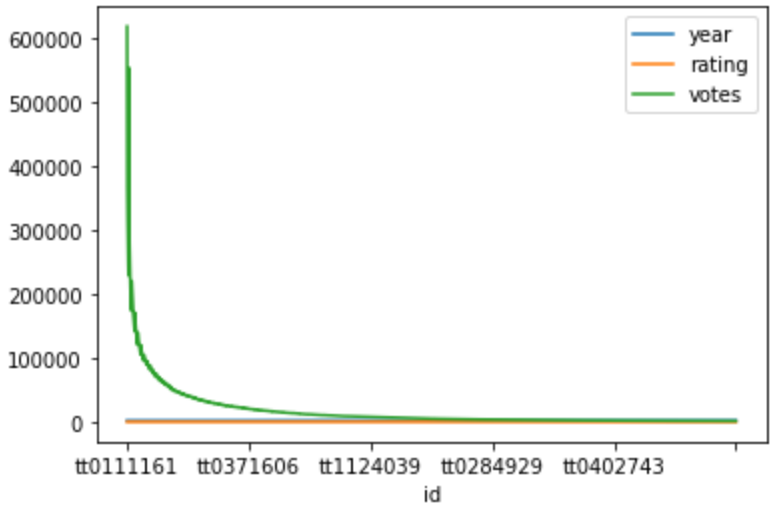
Out[36]:
```
9.2     2
9.0     4
8.9     5
8.8     9
8.7    13
       ..
1.9     9
1.8    12
1.7    12
1.6     6
1.5     7
Name: rating, Length: 77, dtype: int64
```

Plotting

In [1]:
```
%matplotlib inline
```

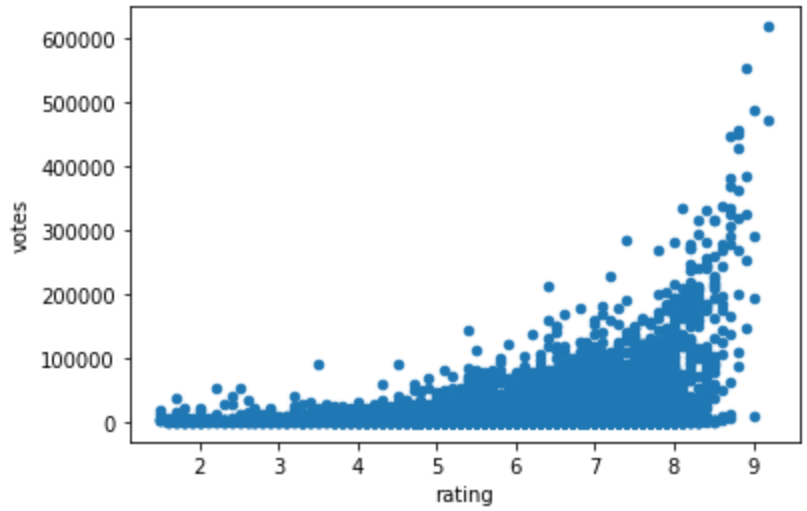
In [7]:
```
data.plot()
```

Out[7]:
```
<AxesSubplot:xlabel='id'>
```



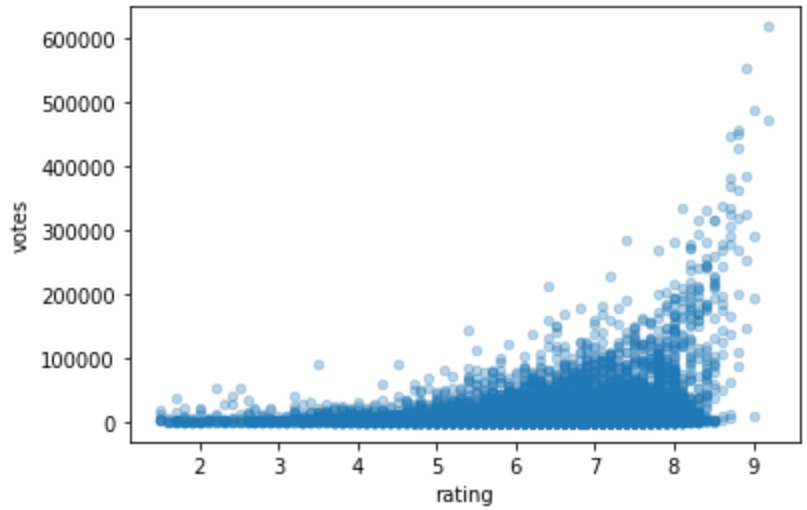
In [8]:
```
data.plot(kind='scatter', x='rating', y='votes')
```

Out[8]:
```
<AxesSubplot:xlabel='rating', ylabel='votes'>
```



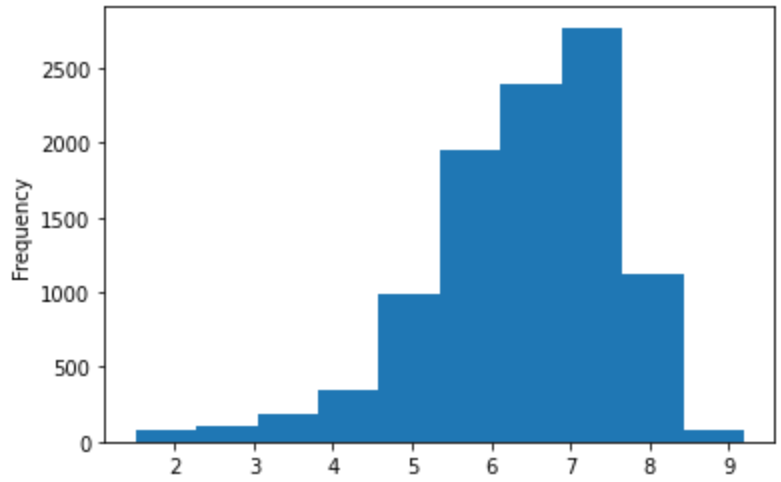
In [9]:
```
data.plot(kind='scatter', x='rating', y='votes', alpha=.3)
```

Out[9]:
```
<AxesSubplot:xlabel='rating', ylabel='votes'>
```



In [12]:
```
data['rating'].plot(kind='hist', x='rating', y='votes')
```

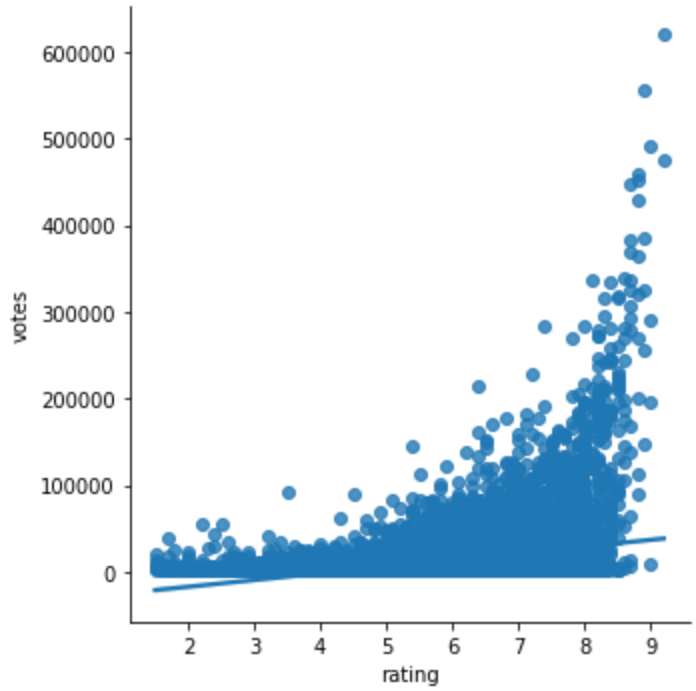
Out[12]:
```
<AxesSubplot:ylabel='Frequency'>
```



In [17]:
```
import seaborn as sns
```

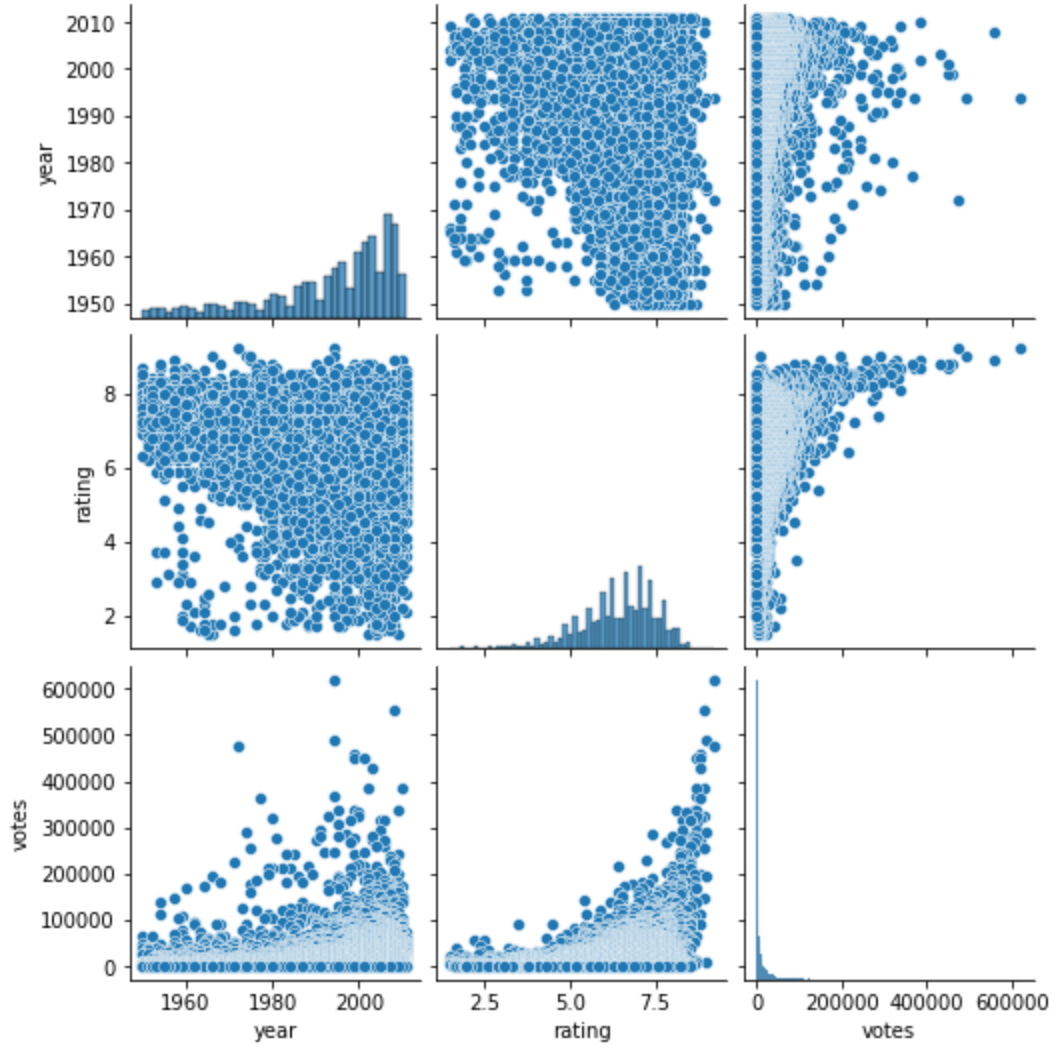
In [6]:
```
sns.lmplot(x='rating', y='votes', data=data)
```

Out[6]:
```
<seaborn.axisgrid.FacetGrid at 0x20317fe7790>
```



In [13]:
```
sns.pairplot(data)
```

Out[13]:
```
<seaborn.axisgrid.PairGrid at 0x20317ef3f40>
```



Ordinary Least Squares (OLS) Regression

In [4]:
```
import statsmodels.api as sm
```

In [8]:
```
results = sm.OLS(data['votes'], data['rating']).fit()
```

In [9]:
```
results.summary()
```

Out[9]:

OLS Regression Results

Dep. Variable:	votes	**R-squared (uncentered):**	0.221
Model:	OLS	**Adj. R-squared (uncentered):**	0.220
Method:	Least Squares	**F-statistic:**	2829.
Date:	Tue, 01 Feb 2022	**Prob (F-statistic):**	0.00
Time:	14:20:27	**Log-Likelihood:**	-1.1849e+05
No. Observations:	10000	**AIC:**	2.370e+05
Df Residuals:	9999	**BIC:**	2.370e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
rating	2771.9868	52.115	53.190	0.000	2669.831	2874.143

Omnibus:	11448.927	**Durbin-Watson:**	0.030
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1297144.076
Skew:	5.966	**Prob(JB):**	0.00
Kurtosis:	57.505	**Cond. No.**	1.00

Notes:
[1] R^2 is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Advance Data Selection

In [11]:
```
data[data['rating']>8.5]
```

Out[11]:

id	title	year	rating	votes	length	genres
tt0111161	The Shawshank Redemption (1994)	1994	9.2	619479	142 mins.	Crime|Drama
tt0110912	Pulp Fiction (1994)	1994	9.0	490065	154 mins.	Crime|Thriller
tt0137523	Fight Club (1999)	1999	8.8	458173	139 mins.	Drama|Mystery|Thriller
tt0133093	The Matrix (1999)	1999	8.7	448114	136 mins.	Action|Adventure|Sci-Fi
tt1375666	Inception (2010)	2010	8.9	385149	148 mins.	Action|Adventure|Sci-Fi|Thriller
tt0109830	Forrest Gump (1994)	1994	8.7	368994	142 mins.	Comedy|Drama|Romance
tt0169547	American Beauty (1999)	1999	8.6	338332	122 mins.	Drama
tt0108052	Schindler's List (1993)	1993	8.9	325888	195 mins.	Biography|Drama|History|War
tt0080684	Star Wars: Episode V - The Empire Strikes Back...	1980	8.8	320105	124 mins.	Action|Adventure|Family|Sci-Fi
tt0114814	The Usual Suspects (1995)	1995	8.7	306624	106 mins.	Crime|Mystery|Thriller
tt0102926	The Silence of the Lambs (1991)	1991	8.7	293081	118 mins.	Crime|Thriller
tt0099685	Goodfellas (1990)	1990	8.8	270728	146 mins.	Biography|Crime|Drama|Thriller
tt0110413	Leon: The Professional (1994)	1994	8.6	244568	110 mins.	Crime|Drama|Thriller
tt0468569	The Dark Knight (2008)	2008	8.9	555122	152 mins.	Action|Crime|Drama|Thriller
tt0068646	The Godfather (1972)	1972	9.2	474189	175 mins.	Crime|Drama
tt0120737	The Lord of the Rings: The Fellowship of the R...	2001	8.8	451263	178 mins.	Action|Adventure|Drama|Fantasy
tt0167260	The Lord of the Rings: The Return of the King ...	2003	8.8	428791	201 mins.	Action|Adventure|Drama|Fantasy
tt0167261	The Lord of the Rings: The Two Towers (2002)	2002	8.7	383113	179 mins.	Action|Adventure|Drama|Fantasy
tt0076759	Star Wars: Episode IV - A New Hope (1977)	1977	8.8	364211	121 mins.	Action|Adventure|Family|Fantasy|Sci-Fi
tt0114369	Se7en (1995)	1995	8.7	337198	127 mins.	Crime|Drama|Mystery|Thriller
tt0209144	Memento (2000)	2000	8.7	325663	113 mins.	Crime|Drama|Mystery|Thriller
tt0071562	The Godfather: Part II (1974)	1974	9.0	291169	200 mins.	Crime|Drama

id	title	year	rating	votes	length	genres
tt0103064	Terminator 2: Judgment Day (1991)	1991	8.6	280590	137 mins.	Action|Sci-Fi|Thriller
tt0082971	Raiders of the Lost Ark (1981)	1981	8.7	277941	115 mins.	Action|Adventure
tt0120586	American History X (1998)	1998	8.6	270082	119 mins.	Crime|Drama
tt0073486	One Flew Over the Cuckoo's Nest (1975)	1975	8.9	255503	133 mins.	Drama
tt0317248	City of God (2002)	2002	8.8	199917	130 mins.	Crime|Drama
tt0078788	Apocalypse Now (1979)	1979	8.6	198861	153 mins.	Drama|War
tt0060196	The Good, the Bad and the Ugly (1966)	1966	9.0	195238	161 mins.	Western
tt0075314	Taxi Driver (1976)	1976	8.6	186983	113 mins.	Drama|Thriller
tt0057012	Dr. Strangelove or: How I Learned to Stop Worr...	1964	8.6	174723	95 mins.	Comedy|Drama
tt0435761	Toy Story 3 (2010)	2010	8.6	144200	103 mins.	Animation|Adventure|Comedy|Family|Fantasy
tt0054215	Psycho (1960)	1960	8.7	168286	109 mins.	Horror|Mystery|Thriller
tt0050083	12 Angry Men (1957)	1957	8.9	148155	96 mins.	Drama|Mystery
tt0245429	Spirited Away (2001)	2001	8.6	125718	125 mins.	Animation|Adventure|Family|Fantasy
tt0047396	Rear Window (1954)	1954	8.7	137663	112 mins.	Mystery|Romance|Thriller
tt0047478	Seven Samurai (1954)	1954	8.8	111707	207 mins.	Adventure|Drama
tt0053125	North by Northwest (1959)	1959	8.6	106359	131 mins.	Adventure|Drama|Mystery|Romance|Thriller
tt0064116	Once Upon a Time in the West (1968)	1968	8.8	89764	175 mins.	Western
tt0043014	Sunset Blvd. (1950)	1950	8.7	64363	110 mins.	Drama|Film-Noir
tt0050825	Paths of Glory (1957)	1957	8.6	52498	88 mins.	Crime|Drama|War
tt0476735	My Father and My Son (2005)	2005	8.7	14080	108 mins.	Drama
tt1832382	A Separation (2011)	2011	8.6	11954	120 mins.	Drama
tt0252487	Outrageous Class (1975)	1975	9.0	9823	87 mins.	Comedy|Drama

id	title	year	rating	votes	length	genres
tt0253828	Tosun Pasa (1976)	1976	8.7	6559	90 mins.	Comedy|History
tt0076276	Who Sings Over There (1980)	1980	8.6	3868	86 mins.	Comedy

In [18]:
```
data[data['year'] > 2000]
```

Out[18]:

id	title	year	rating	votes	length	genres
tt1375666	Inception (2010)	2010	8.9	385149	148 mins.	Action|Adventure|Sci-Fi|Thriller
tt0499549	Avatar (2009)	2009	8.1	336855	162 mins.	Action|Adventure|Fantasy|Sci-Fi
tt0372784	Batman Begins (2005)	2005	8.3	316613	140 mins.	Action|Crime|Drama|Thriller
tt0266697	Kill Bill: Vol. 1 (2003)	2003	8.2	272983	111 mins.	Action|Crime|Thriller
tt0416449	300 (2006)	2006	7.8	269328	117 mins.	Action|Fantasy|History|War
...	...	...	...	...	...	...
tt0421090	Zerophilia (2005)	2005	6.3	1359	90 mins.	Comedy|Romance
tt0339727	Stateside (2004)	2004	5.8	1358	97 mins.	Drama|Music|Romance
tt0807721	Meduzot (2007)	2007	7.0	1357	78 mins.	Drama
tt0339642	Daltry Calhoun (2005)	2005	5.2	1357	100 mins.	Comedy|Drama|Music|Romance
tt0279977	The Navigators (2001)	2001	6.9	1356	96 mins.	Comedy|Drama

4261 rows × 6 columns

In [20]:
```
data[(data['year'] == 1966) & (data['year'])]
```

Out[20]:

id	title	year	rating	votes	length	genres
tt0060196	The Good, the Bad and the Ugly (1966)	1966	9.0	195238	161 mins.	Western
tt0061184	Who's Afraid of Virginia Woolf? (1966)	1966	8.2	23811	131 mins.	Drama
tt0060666	Manos: The Hands of Fate (1966)	1966	1.5	20927	74 mins.	Horror
tt0060827	Persona (1966)	1966	8.2	20157	85 mins.	Drama|Fantasy
tt0060176	Blow-Up (1966)	1966	7.6	18679	111 mins.	Drama|Mystery|Thriller
...	...	...	...	...	...	...
tt0060214	Carry on Screaming! (1966)	1966	6.7	1427	97 mins.	Comedy|Horror
tt0060305	Le Deuxieme Souffle (1966)	1966	8.0	1393	150 mins.	Crime|Drama
tt0060841	The Plague of the Zombies (1966)	1966	6.7	1386	91 mins.	Horror
tt0061204	The Wrong Box (1966)	1966	6.9	1372	105 mins.	Comedy
tt0060880	The Quiller Memorandum (1966)	1966	6.5	1356	104 mins.	Drama|Mystery|Thriller

69 rows × 6 columns

In [21]:
```
data[data['year'] > 1995 ]
```

Out[21]:

id	title	year	rating	votes	length	genres
tt0137523	Fight Club (1999)	1999	8.8	458173	139 mins.	Drama|Mystery|Thriller
tt0133093	The Matrix (1999)	1999	8.7	448114	136 mins.	Action|Adventure|Sci-Fi
tt1375666	Inception (2010)	2010	8.9	385149	148 mins.	Action|Adventure|Sci-Fi|Thriller
tt0169547	American Beauty (1999)	1999	8.6	338332	122 mins.	Drama
tt0499549	Avatar (2009)	2009	8.1	336855	162 mins.	Action|Adventure|Fantasy|Sci-Fi
...	...	...	...	...	...	...
tt0118635	Aprile (1998)	1998	6.7	1358	78 mins.	Comedy
tt0807721	Meduzot (2007)	2007	7.0	1357	78 mins.	Drama
tt0339642	Daltry Calhoun (2005)	2005	5.2	1357	100 mins.	Comedy|Drama|Music|Romance
tt0152836	Taal (1999)	1999	6.5	1356	179 mins.	Musical|Romance
tt0279977	The Navigators (2001)	2001	6.9	1356	96 mins.	Comedy|Drama

5710 rows × 6 columns

In [27]:
```
data[(data['year'] > 1995) & (data['year'] < 2000)].sort_values(by='rating', ascending=F
```

Out[27]:

id	title	year	rating	votes	length	genres
tt0137523	Fight Club (1999)	1999	8.8	458173	139 mins.	Drama|Mystery|Thriller
tt0133093	The Matrix (1999)	1999	8.7	448114	136 mins.	Action|Adventure|Sci-Fi
tt0120586	American History X (1998)	1998	8.6	270082	119 mins.	Crime|Drama
tt0169547	American Beauty (1999)	1999	8.6	338332	122 mins.	Drama
tt0118799	Life Is Beautiful (1997)	1997	8.5	131578	116 mins.	Comedy|Drama|Romance|War
tt0120815	Saving Private Ryan (1998)	1998	8.5	317912	169 mins.	Action|Drama|History|War
tt0119488	L.A. Confidential (1997)	1997	8.4	187115	138 mins.	Crime|Drama|Mystery|Thriller
tt0119698	Princess Mononoke (1997)	1997	8.4	77859	134 mins.	Animation|Adventure|Fantasy
tt0120689	The Green Mile (1999)	1999	8.4	243660	189 mins.	Crime|Drama|Fantasy|Mystery
tt0128332	Innocence (1997)	1997	8.3	2402	110 mins.	Drama

Grouping

In [29]:
```
data.groupby(data['year'])['rating'].mean()
```

Out[29]:
```
year
1950    7.545161
1951    7.478125
1952    7.475676
1953    7.106383
1954    7.371795
          ...
2007    6.303831
2008    6.275260
2009    6.287290
2010    6.340635
2011    6.357143
Name: rating, Length: 62, dtype: float64
```

In [30]:
```
data.groupby(data['year'])['rating'].min()
```

Out[30]:
```
year
1950    6.3
1951    6.2
1952    6.3
1953    2.9
1954    6.1
       ...
2007    1.6
2008    1.7
2009    1.5
2010    2.2
2011    2.1
Name: rating, Length: 62, dtype: float64
```

In [31]:
```
data.groupby(data['year'])['rating'].max()
```

Out[31]:
```
year
1950    8.7
1951    8.3
1952    8.4
1953    8.3
1954    8.8
       ...
2007    8.3
2008    8.9
2009    8.4
2010    8.9
2011    8.6
Name: rating, Length: 62, dtype: float64
```

In [32]:
```
data.groupby(data['year'])['rating'].std()
```

Out[32]:
```
year
1950    0.544572
1951    0.542833
1952    0.524619
1953    1.000957
1954    0.623635
          ...
2007    1.162358
2008    1.113193
2009    1.044310
2010    1.104738
2011    1.302258
Name: rating, Length: 62, dtype: float64
```

Challenges

1. What was the highest scoring movie in 1996?
   `A: Fargo & The Bandit at 8.3`

In [81]:
```
data[(data['year'] == 1996)].sort_values(by='rating', ascending=False)
```

Out[81]:

id	title	year	rating	votes	length	genres
tt0116282	Fargo (1996)	1996	8.3	187498	98 mins.	Crime|Drama|Thriller
tt0116231	The Bandit (1996)	1996	8.3	13288	121 mins.	Action|Crime|Drama|Romance|Thriller
tt0117951	Trainspotting (1996)	1996	8.2	175993	94 mins.	Crime|Drama
tt0117666	Sling Blade (1996)	1996	8.0	44316	135 mins.	Drama
tt0117589	Secrets & Lies (1996)	1996	8.0	15435	142 mins.	Comedy|Drama
...	...	...	...	...	...	...
tt0116165	Ed (1996)	1996	2.4	4085	94 mins.	Comedy|Family|Sport
tt0116839	Lawnmower Man 2: Beyond Cyberspace (1996)	1996	2.2	5299	92 mins.	Action|Sci-Fi|Thriller
tt0117676	Snowboard Academy (1996)	1996	2.2	1776	88 mins.	Comedy|Sport
tt0117550	Santa with Muscles (1996)	1996	2.1	6420	97 mins.	Family|Comedy
tt0174917	Merlin's Shop of Mystical Wonders (1996)	1996	1.7	2362	92 mins.	Fantasy|Horror

265 rows × 6 columns

1. In what year was the highest rated movie of all time made?
   A: The Shawshank Redemption (1994) & The Godfather (1972) both at 9.2 rating

In [82]:
```
data.sort_values(by='rating', ascending=False)
```

Out[82]:

id	title	year	rating	votes	length	genres
tt0111161	The Shawshank Redemption (1994)	1994	9.2	619479	142 mins.	Crime|Drama
tt0068646	The Godfather (1972)	1972	9.2	474189	175 mins.	Crime|Drama
tt0060196	The Good, the Bad and the Ugly (1966)	1966	9.0	195238	161 mins.	Western
tt0110912	Pulp Fiction (1994)	1994	9.0	490065	154 mins.	Crime|Thriller
tt0252487	Outrageous Class (1975)	1975	9.0	9823	87 mins.	Comedy|Drama
...	...	...	...	...	...	...
tt0364986	Ben & Arthur (2002)	2002	1.5	4675	85 mins.	Drama|Romance
tt0060753	Night Train to Mundo Fine (1966)	1966	1.5	3542	89 mins.	Action|Adventure|Crime|War
tt0421051	Daniel the Wizard (2004)	2004	1.5	8271	81 mins.	Comedy|Crime|Family|Fantasy|Horror
tt0059464	Monster a-Go Go (1965)	1965	1.5	3255	70 mins.	Sci-Fi|Horror
tt0060666	Manos: The Hands of Fate (1966)	1966	1.5	20927	74 mins.	Horror

10000 rows × 6 columns

In [83]:
```
data[data['rating'] == data['rating'].max()]
```

Out[83]:

id	title	year	rating	votes	length	genres
tt0111161	The Shawshank Redemption (1994)	1994	9.2	619479	142 mins.	Crime|Drama
tt0068646	The Godfather (1972)	1972	9.2	474189	175 mins.	Crime|Drama

1. What five movies have the most votes ever?
   A: The Shawshank Redemption (619479)
   The Dark Knight (555122)
   Pulp Fiction (490065)
   The Godfather (474189)
   Fight Club (458173)

In [84]:
```
data.sort_values(by='votes', ascending=False).head()
```

Out[84]:

id	title	year	rating	votes	length	genres
tt0111161	The Shawshank Redemption (1994)	1994	9.2	619479	142 mins.	Crime|Drama
tt0468569	The Dark Knight (2008)	2008	8.9	555122	152 mins.	Action|Crime|Drama|Thriller
tt0110912	Pulp Fiction (1994)	1994	9.0	490065	154 mins.	Crime|Thriller
tt0068646	The Godfather (1972)	1972	9.2	474189	175 mins.	Crime|Drama
tt0137523	Fight Club (1999)	1999	8.8	458173	139 mins.	Drama|Mystery|Thriller

1. What year in the 1960s had the highest average movie rating?
   A: 1962 had the highest average rating at 7.34

In [49]:
```
data[(data['year'] > 1960) & (data['year'] < 1969)].sort_values(by='rating', ascending=F
```

Out[49]:

id	title	year	rating	votes	length	genres
tt0060196	The Good, the Bad and the Ugly (1966)	1966	9.0	195238	161 mins.	Western
tt0064116	Once Upon a Time in the West (1968)	1968	8.8	89764	175 mins.	Western
tt0057012	Dr. Strangelove or: How I Learned to Stop Worr...	1964	8.6	174723	95 mins.	Comedy|Drama
tt0056592	To Kill a Mockingbird (1962)	1962	8.5	93918	129 mins.	Crime|Drama|Mystery
tt0056172	Lawrence of Arabia (1962)	1962	8.5	89628	216 mins.	Adventure|Biography|Drama|History|War
tt0055630	Yojimbo (1961)	1961	8.4	33878	110 mins.	Action|Crime|Drama|Thriller
tt0062622	2001: A Space Odyssey (1968)	1968	8.4	183207	141 mins.	Adventure|Mystery|Sci-Fi
tt0056058	Harakiri (1962)	1962	8.4	6390	133 mins.	Drama
tt0059578	For a Few Dollars More (1965)	1965	8.3	54321	132 mins.	Action|Crime|Western
tt0061512	Cool Hand Luke (1967)	1967	8.3	52791	126 mins.	Crime|Drama

In [56]:
```
data[(data['year'] > 1960) & (data['year'] < 1969)].sort_values(by='rating', ascending=F
```

Out[56]:

id	title	year	rating	votes	length	genres
tt0060196	The Good, the Bad and the Ugly (1966)	1966	9.0	195238	161 mins.	Western
tt0064116	Once Upon a Time in the West (1968)	1968	8.8	89764	175 mins.	Western
tt0057012	Dr. Strangelove or: How I Learned to Stop Worr...	1964	8.6	174723	95 mins.	Comedy|Drama
tt0056592	To Kill a Mockingbird (1962)	1962	8.5	93918	129 mins.	Crime|Drama|Mystery
tt0056172	Lawrence of Arabia (1962)	1962	8.5	89628	216 mins.	Adventure|Biography|Drama|History|War
tt0055630	Yojimbo (1961)	1961	8.4	33878	110 mins.	Action|Crime|Drama|Thriller
tt0062622	2001: A Space Odyssey (1968)	1968	8.4	183207	141 mins.	Adventure|Mystery|Sci-Fi
tt0056058	Harakiri (1962)	1962	8.4	6390	133 mins.	Drama
tt0059578	For a Few Dollars More (1965)	1965	8.3	54321	132 mins.	Action|Crime|Western
tt0061512	Cool Hand Luke (1967)	1967	8.3	52791	126 mins.	Crime|Drama

In [69]:
```
data[(data['year'] == 1960)].mean()
```

```
C:\Users\Admin\AppData\Local\Temp\ipykernel_21468\3872299459.py:1: FutureWarning: Droppi
ng of nuisance columns in DataFrame reductions (with 'numeric_only=None') is deprecated;
in a future version this will raise TypeError.  Select only valid columns before calling
the reduction.
  data[(data['year'] == 1960)].mean()
```

Out[69]:
```
year       1960.000000
rating        7.247917
votes     10692.104167
dtype: float64
```

In [70]:
```
data[(data['year'] == 1961)].mean()
```

```
C:\Users\Admin\AppData\Local\Temp\ipykernel_21468\3184299725.py:1: FutureWarning: Droppi
ng of nuisance columns in DataFrame reductions (with 'numeric_only=None') is deprecated;
in a future version this will raise TypeError.  Select only valid columns before calling
the reduction.
  data[(data['year'] == 1961)].mean()
```

Out[70]:
```
year      1961.000000
rating       7.195349
votes     7551.651163
dtype: float64
```

In [71]:
```
data[(data['year'] == 1962)].mean()
```

```
C:\Users\Admin\AppData\Local\Temp\ipykernel_21468\4112441894.py:1: FutureWarning: Droppi
ng of nuisance columns in DataFrame reductions (with 'numeric_only=None') is deprecated;
in a future version this will raise TypeError.  Select only valid columns before calling
the reduction.
  data[(data['year'] == 1962)].mean()
```

Out[71]:
```
year       1962.000000
rating        7.349057
votes     10355.943396
dtype: float64
```

In [72]:
```
data[(data['year'] == 1963)].mean()
```

```
C:\Users\Admin\AppData\Local\Temp\ipykernel_21468\1502393903.py:1: FutureWarning: Droppi
ng of nuisance columns in DataFrame reductions (with 'numeric_only=None') is deprecated;
in a future version this will raise TypeError.  Select only valid columns before calling
the reduction.
  data[(data['year'] == 1963)].mean()
```

Out[72]:
```
year      1963.000000
rating       7.103922
votes     8699.352941
dtype: float64
```

In [73]:
```
data[(data['year'] == 1964)].mean()
```

```
C:\Users\Admin\AppData\Local\Temp\ipykernel_21468\3766920186.py:1: FutureWarning: Droppi
ng of nuisance columns in DataFrame reductions (with 'numeric_only=None') is deprecated;
in a future version this will raise TypeError.  Select only valid columns before calling
the reduction.
  data[(data['year'] == 1964)].mean()
```

Out[73]:
```
year      1964.000000
rating       6.865217
votes     8543.318841
dtype: float64
```

In [74]:
```
data[(data['year'] == 1965)].mean()
```

```
C:\Users\Admin\AppData\Local\Temp\ipykernel_21468\3368280786.py:1: FutureWarning: Droppi
ng of nuisance columns in DataFrame reductions (with 'numeric_only=None') is deprecated;
in a future version this will raise TypeError.  Select only valid columns before calling
the reduction.
  data[(data['year'] == 1965)].mean()
```

Out[74]:
```
year      1965.000000
rating       6.996875
votes     5933.921875
dtype: float64
```

In [75]:
```
data[(data['year'] == 1966)].mean()
```

```
C:\Users\Admin\AppData\Local\Temp\ipykernel_21468\211241738.py:1: FutureWarning: Droppin
g of nuisance columns in DataFrame reductions (with 'numeric_only=None') is deprecated;
in a future version this will raise TypeError.  Select only valid columns before calling
the reduction.
  data[(data['year'] == 1966)].mean()
```

Out[75]:
```
year      1966.000000
rating       6.895652
votes     7826.028986
dtype: float64
```

In [76]:
```
data[(data['year'] == 1967)].mean()
```

```
C:\Users\Admin\AppData\Local\Temp\ipykernel_21468\32996977.py:1: FutureWarning: Dropping
of nuisance columns in DataFrame reductions (with 'numeric_only=None') is deprecated; in
a future version this will raise TypeError.  Select only valid columns before calling th
e reduction.
  data[(data['year'] == 1967)].mean()
```

Out[76]:
```
year      1967.000000
rating       7.169091
votes     8904.909091
dtype: float64
```

In [ ]:

In [77]:
```
data[(data['year'] == 1968)].mean()
```

```
C:\Users\Admin\AppData\Local\Temp\ipykernel_21468\4059933262.py:1: FutureWarning: Droppi
ng of nuisance columns in DataFrame reductions (with 'numeric_only=None') is deprecated;
in a future version this will raise TypeError.  Select only valid columns before calling
the reduction.
  data[(data['year'] == 1968)].mean()
```

Out[77]:
```
year       1968.000000
rating        6.967213
votes     12314.229508
dtype: float64
```

In [78]:
```
data[(data['year'] == 1969)].mean()
```

```
C:\Users\Admin\AppData\Local\Temp\ipykernel_21468\3712091619.py:1: FutureWarning: Droppi
ng of nuisance columns in DataFrame reductions (with 'numeric_only=None') is deprecated;
in a future version this will raise TypeError.  Select only valid columns before calling
the reduction.
  data[(data['year'] == 1969)].mean()
```

Out[78]:
```
year       1969.000000
rating        7.011321
votes      7267.150943
dtype: float64
```

In [85]:
```
data[(data['year'] >= 1960) & (data['year'] < 1970)].groupby(data['year']).mean()
```

Out[85]:

year	year	rating	votes
1960	1960.0	7.247917	10692.104167
1961	1961.0	7.195349	7551.651163
1962	1962.0	7.349057	10355.943396
1963	1963.0	7.103922	8699.352941
1964	1964.0	6.865217	8543.318841
1965	1965.0	6.996875	5933.921875
1966	1966.0	6.895652	7826.028986
1967	1967.0	7.169091	8904.909091
1968	1968.0	6.967213	12314.229508
1969	1969.0	7.011321	7267.150943

Cleaning Data

In [89]:
```
data['formatted title'] = data['title'].str[:-7]
```

In [90]:
```
data.head()
```

Out[90]:

id	title	year	rating	votes	length	genres	formatted title
tt0111161	The Shawshank Redemption (1994)	1994	9.2	619479	142 mins.	Crime|Drama	The Shawshank Redemption
tt0110912	Pulp Fiction (1994)	1994	9.0	490065	154 mins.	Crime|Thriller	Pulp Fiction
tt0137523	Fight Club (1999)	1999	8.8	458173	139 mins.	Drama|Mystery|Thriller	Fight Club
tt0133093	The Matrix (1999)	1999	8.7	448114	136 mins.	Action|Adventure|Sci-Fi	The Matrix
tt1375666	Inception (2010)	2010	8.9	385149	148 mins.	Action|Adventure|Sci-Fi|Thriller	Inception

In [97]:
```
data['formatted title'] = data['title'].str.split('\(').str[0]
```

In [98]:
```
data.head()
```

Out[98]:

id	title	year	rating	votes	length	genres	formatted title
tt0111161	The Shawshank Redemption (1994)	1994	9.2	619479	142 mins.	Crime|Drama	The Shawshank Redemption
tt0110912	Pulp Fiction (1994)	1994	9.0	490065	154 mins.	Crime|Thriller	Pulp Fiction
tt0137523	Fight Club (1999)	1999	8.8	458173	139 mins.	Drama|Mystery|Thriller	Fight Club
tt0133093	The Matrix (1999)	1999	8.7	448114	136 mins.	Action|Adventure|Sci-Fi	The Matrix
tt1375666	Inception (2010)	2010	8.9	385149	148 mins.	Action|Adventure|Sci-Fi|Thriller	Inception

In [104…
```
data['formatted length'] = data['length'].str.split('m').str[0]
```

In [105…
```
data.head()
```

Out[105]:

id	title	year	rating	votes	length	genres	formatted title	formatted length
tt0111161	The Shawshank Redemption (1994)	1994	9.2	619479	142 mins.	Crime|Drama	142	142
tt0110912	Pulp Fiction (1994)	1994	9.0	490065	154 mins.	Crime|Thriller	154	154
tt0137523	Fight Club (1999)	1999	8.8	458173	139 mins.	Drama|Mystery|Thriller	139	139
tt0133093	The Matrix (1999)	1999	8.7	448114	136 mins.	Action|Adventure|Sci-Fi	136	136
tt1375666	Inception (2010)	2010	8.9	385149	148 mins.	Action|Adventure|Sci-Fi|Thriller	148	148

In [6]:
```
data['length'].str.split().str.get(0)
```

Out[6]:
```
id
tt0111161    142
tt0110912    154
tt0137523    139
tt0133093    136
tt1375666    148
             ...
tt0807721     78
tt0339642    100
tt0060880    104
tt0152836    179
tt0279977     96
Name: length, Length: 10000, dtype: object
```

In [14]:
```
data['formatted length'] = data['length'].str.replace('mins.', '').astype('int')
```

```
C:\Users\Admin\AppData\Local\Temp\ipykernel_14332\1196249904.py:1: FutureWarning: The de
fault value of regex will change from True to False in a future version.
  data['formatted length'] = data['length'].str.replace('mins.', '').astype('int')
```

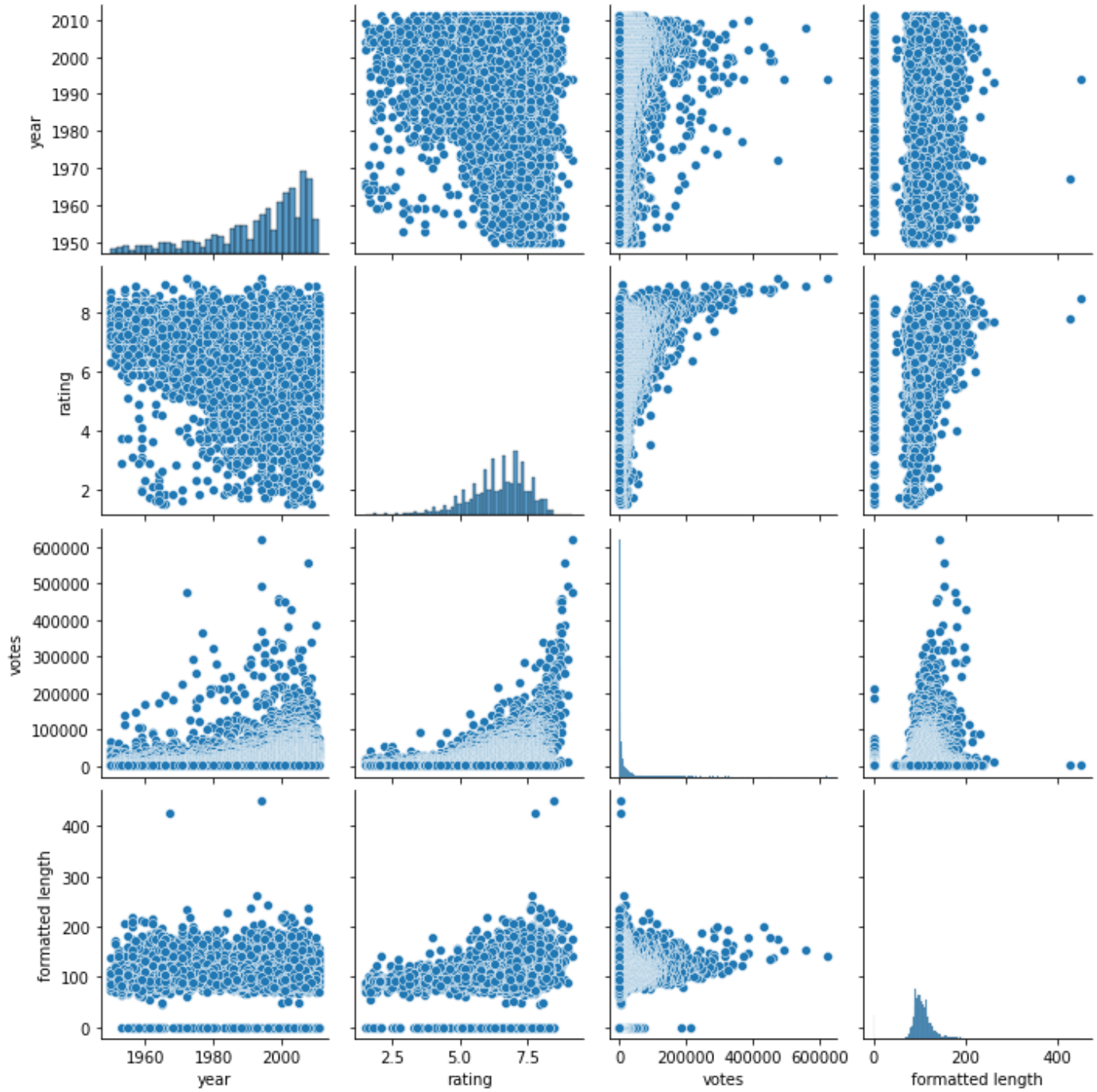
In [15]:
```
data.head()
```

Out[15]:

id	title	year	rating	votes	length	genres	formatted length
tt0111161	The Shawshank Redemption (1994)	1994	9.2	619479	142 mins.	Crime|Drama	142
tt0110912	Pulp Fiction (1994)	1994	9.0	490065	154 mins.	Crime|Thriller	154
tt0137523	Fight Club (1999)	1999	8.8	458173	139 mins.	Drama|Mystery|Thriller	139
tt0133093	The Matrix (1999)	1999	8.7	448114	136 mins.	Action|Adventure|Sci-Fi	136
tt1375666	Inception (2010)	2010	8.9	385149	148 mins.	Action|Adventure|Sci-Fi|Thriller	148

In [21]:
```
sns.pairplot(data)
```

Out[21]:
```
<seaborn.axisgrid.PairGrid at 0x289d1f54f70>
```



In [22]:
```
data[data['formatted length'] == 0]
```

Out[22]:

id	title	year	rating	votes	length	genres	formatted length
tt0457430	Pan's Labyrinth (2006)	2006	8.4	186080	00 mins.	Drama|Fantasy|Mystery	0
tt0081505	The Shining (1980)	1980	8.5	212988	00 mins.	Drama|Horror|Mystery	0
tt1077258	Planet Terror (2007)	2007	7.5	74950	00 mins.	Action|Horror|Sci-Fi	0
tt0075860	Close Encounters of the Third Kind (1977)	1977	7.8	65768	00 mins.	Drama|Sci-Fi	0
tt0054331	Spartacus (1960)	1960	8.0	55504	00 mins.	Action|Adventure|Biography|Drama|History	0
...	...	...	...	...	...	...	...
tt1042499	Filth and Wisdom (2008)	2008	5.1	1367	00 mins.	Comedy|Drama|Music|Romance	0
tt0094642	American Gothic (1988)	1988	5.1	1359	00 mins.	Horror	0
tt1926313	Pyaar Ka Punchnama (2011)	2011	8.0	1359	00 mins.	Comedy|Drama|Romance	0
tt0878674	Garage (2007)	2007	6.9	1356	00 mins.	Drama	0
tt0197094	What Becomes of the Broken Hearted? (1999)	1999	6.3	1365	00 mins.	Drama	0

282 rows × 7 columns

In [ ]: