In [1]:
```
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
```

In [ ]:
```

```

In [2]:
```
movies = pd.read_parquet('movies.parquet')
movies.head(1)
```

Out[2]:

imdb_title_id	title	original_title	year	date_published	genre	duration	country	language	dire
tt0000009	Miss Jerry	Miss Jerry	1894	1894-10-09	Romance	45	USA	None	Alexa B

1 rows × 21 columns

In [ ]:
```

```

In [3]:
```
names = pd.read_parquet('names.parquet')
names.head(1)
```

Out[3]:

imdb_name_id	name	birth_name	height	bio	birth_details	date_of_birth	place_of_birth	deat
nm0000001	Fred Astaire	Frederic Austerlitz Jr.	177.0	Fred Astaire was born in Omaha, Nebraska, to J...	May 10, 1899 in Omaha, Nebraska, USA	1899-05-10	Omaha, Nebraska, USA	Jun

In [ ]:
```

```

In [4]:
```
ratings = pd.read_parquet('ratings.parquet')
ratings.head(1)
```

Out[4]:

imdb_title_id	weighted_average_vote	total_votes	mean_vote	median_vote	votes_10	votes_9	votes_8
tt0000009	5.9	154	5.9	6.0	12	4	1(

1 rows × 48 columns

In [ ]:

```
```

In [5]:

```
title_principles = pd.read_parquet('title_principals.parquet')
title_principles.head(1)
```

Out[5]:

imdb_title_id	ordering	imdb_name_id	category	job	characters
tt0000009	1	nm0063086	actress	None	["Miss Geraldine Holbrook (Miss Jerry)"]

In [ ]:

```
```

In [6]:

```
combined_data = movies.join(title_principles).join(names, on='imdb_name_id')
combined_data.head()
```

Out[6]:

imdb_title_id	title	original_title	year	date_published	genre	duration	country	language	di
tt0000009	Miss Jerry	Miss Jerry	1894	1894-10-09	Romance	45	USA	None	Ale
tt0000009	Miss Jerry	Miss Jerry	1894	1894-10-09	Romance	45	USA	None	Ale
tt0000009	Miss Jerry	Miss Jerry	1894	1894-10-09	Romance	45	USA	None	Ale
tt0000009	Miss Jerry	Miss Jerry	1894	1894-10-09	Romance	45	USA	None	Ale
tt0000574	The Story of the Kelly Gang	The Story of the Kelly Gang	1906	1906-12-26	Biography, Crime, Drama	70	Australia	None	C

5 rows × 42 columns

In [ ]:

In [7]:
```
combined_data.isnull().sum()
```

Out[7]:

```
title                        0
original_title               0
year                         0
date_published               0
genre                        0
duration                     0
country                    545
language                  7807
director                   559
writer                   13832
production_company       41269
actors                     193
description              20392
avg_vote                     0
votes                        0
budget                  604544
usa_gross_income        683822
worlwide_gross_income   530122
metascore               703927
reviews_from_users       72656
reviews_from_critics    112714
ordering                     9
imdb_name_id                 9
category                     9
job                     622777
characters              494675
name                        10
birth_name                  10
height                  602680
bio                     190452
birth_details           333954
date_of_birth           333954
place_of_birth          351362
death_details           632089
date_of_death           632089
place_of_death          639913
reason_of_death         684371
spouses_string          563060
spouses                     10
divorces                    10
spouses_with_children       10
children                    10
dtype: int64
```

In [ ]:

In [8]:

```
movie_data = combined_data.dropna(axis='columns')
movie_data.head()
```

Out[8]:

imdb_title_id	title	original_title	year	date_published	genre	duration	avg_vote	votes
tt0000009	Miss Jerry	Miss Jerry	1894	1894-10-09	Romance	45	5.9	154
tt0000009	Miss Jerry	Miss Jerry	1894	1894-10-09	Romance	45	5.9	154
tt0000009	Miss Jerry	Miss Jerry	1894	1894-10-09	Romance	45	5.9	154
tt0000009	Miss Jerry	Miss Jerry	1894	1894-10-09	Romance	45	5.9	154
tt0000574	The Story of the Kelly Gang	The Story of the Kelly Gang	1906	1906-12-26	Biography, Crime, Drama	70	6.1	589

In [ ]:

In [9]:
```
movie_data.groupby('year').mean()
```

Out[9]:

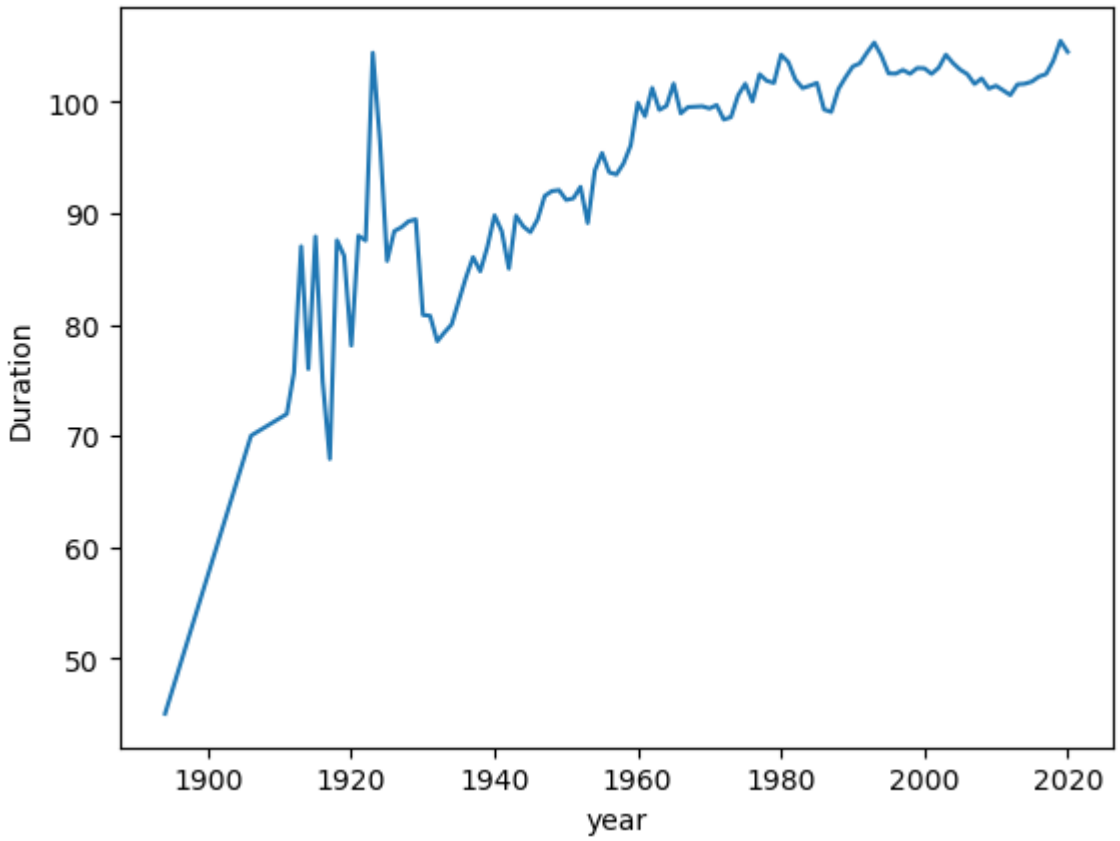
year	duration	avg_vote	votes
1894	45.000000	5.900000	154.000000
1906	70.000000	6.100000	589.000000
1911	71.956522	6.178261	607.000000
1912	75.666667	5.951111	342.866667
1913	87.025210	6.596639	787.689076
...	...	...	...
2016	102.293838	5.672411	9539.061293
2017	102.525504	5.706011	7859.460599
2018	103.712134	5.697565	6825.729651
2019	105.502008	5.791931	6945.636848
2020	104.513035	5.564219	3953.267358

112 rows × 3 columns

In [ ]:

In [10]:
```
movie_data.groupby('year').mean()['duration'].plot(ylabel='Duration')
```

Out[10]:
```
<AxesSubplot:xlabel='year', ylabel='Duration'>
```



In [ ]:

In [11]:
```
movie_data.explode('genre').groupby('genre').mean().sort_values('avg_vote', ascending=
```

Out[11]:

genre	year	duration	avg_vote	votes
Musical, Comedy, Family	2001.000000	184.000000	8.700000	3560.000000
Music, Musical	1974.000000	78.000000	8.500000	692.000000
Family, Sci-Fi, Adventure	1991.000000	140.000000	8.400000	2223.000000
Fantasy, Musical, Sci-Fi	2011.000000	172.000000	8.100000	626.000000
Fantasy, Drama, Romance	1950.000000	112.000000	8.000000	10117.000000
Animation, Fantasy, Mystery	2012.000000	109.000000	8.000000	1203.000000
Biography, History, Musical	1997.000000	150.000000	8.000000	1669.000000
Fantasy, Musical, Mystery	1958.000000	110.000000	8.000000	1168.000000
Crime, Film-Noir, Sport	1949.000000	73.000000	7.900000	7515.000000
Family, Musical, Comedy	1996.000000	105.000000	7.900000	388.000000
Animation, Drama, War	1999.620690	86.551724	7.837931	78253.482759
Drama, Fantasy, Family	1973.000000	75.000000	7.800000	8069.000000
Animation, Biography, Crime	2017.000000	94.000000	7.800000	48412.000000
Adventure, Comedy, Film-Noir	1944.000000	100.000000	7.800000	30249.000000
Action, Musical, War	1964.000000	184.000000	7.800000	439.000000
Adventure, Sport	2012.000000	86.500000	7.700000	240.500000
Family, Fantasy, History	2006.000000	120.000000	7.600000	577.000000
Drama, Musical, Family	1973.333333	148.000000	7.600000	3006.000000
Drama, Musical, Crime	1951.000000	90.000000	7.600000	232.000000
Animation, Drama, History	2006.128205	113.435897	7.582051	2268.410256

In [ ]:

In [12]:
```
data = movie_data.copy()
data['genre'] = data['genre'].str.split(',')
data = data.explode('genre')
data['genre'] = data['genre'].str.strip()
data.groupby('genre').mean().sort_values('avg_vote', ascending=False)
```

Out[12]:

genre	year	duration	avg_vote	votes
Documentary	2000.000000	88.500000	7.300000	615.500000
Film-Noir	1949.139478	84.960781	6.646315	4185.960476
Biography	1995.372979	114.124979	6.625196	22009.574383
History	1988.799592	117.187960	6.548684	10629.230354
War	1978.424390	106.463012	6.433162	7859.233684
News	2015.000000	82.000000	6.400000	105.000000
Animation	2003.070379	87.519645	6.380751	22914.737030
Musical	1971.924021	109.126046	6.250413	3933.559866
Music	1987.491646	100.368177	6.248827	9090.282164
Drama	1993.434491	103.870981	6.241178	8869.302638
Romance	1988.856128	103.891990	6.143578	7531.433537
Sport	1995.942269	103.805278	6.050500	13148.104982
Crime	1989.305463	100.267251	6.033713	13829.848456
Western	1966.563821	91.902807	5.988912	5076.806899
Family	1996.737833	97.083006	5.928564	10466.301872
Comedy	1993.497646	98.301741	5.869614	8520.361772
Adventure	1987.519356	99.430634	5.851632	29690.753398
Mystery	1992.418082	98.153549	5.835528	15351.315304
Fantasy	1994.657015	99.278913	5.752525	21427.210895
Action	1995.980256	106.404328	5.635853	21134.908142
Thriller	2000.770766	99.918057	5.486312	12581.362026
Sci-Fi	1995.418516	95.050151	5.087814	30604.171092
Horror	1999.155126	91.382721	4.853546	7914.089058
Adult	1979.000000	96.500000	4.550000	689.500000
Reality-TV	1998.333333	96.000000	3.800000	189.000000

In [ ]:

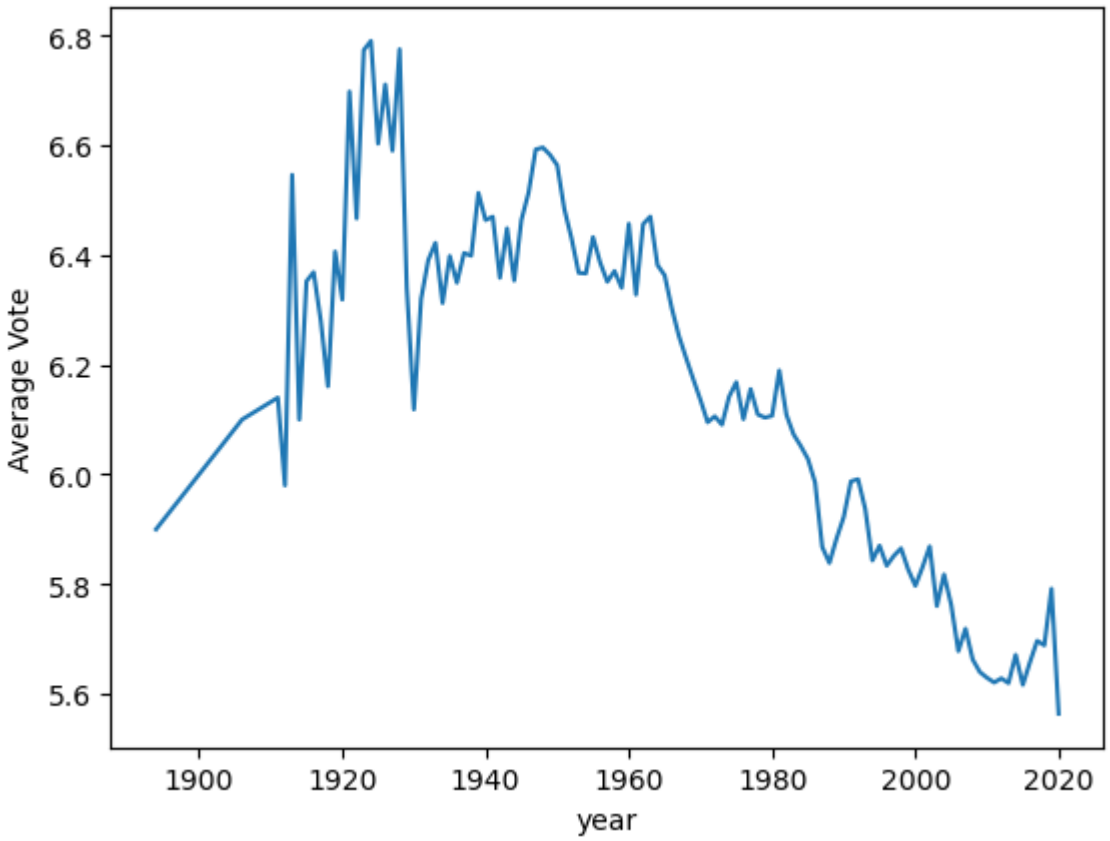
```
```

In [13]:

```
movies.groupby('year').mean()['avg_vote'].plot(ylabel='Average Vote')
```

Out[13]:

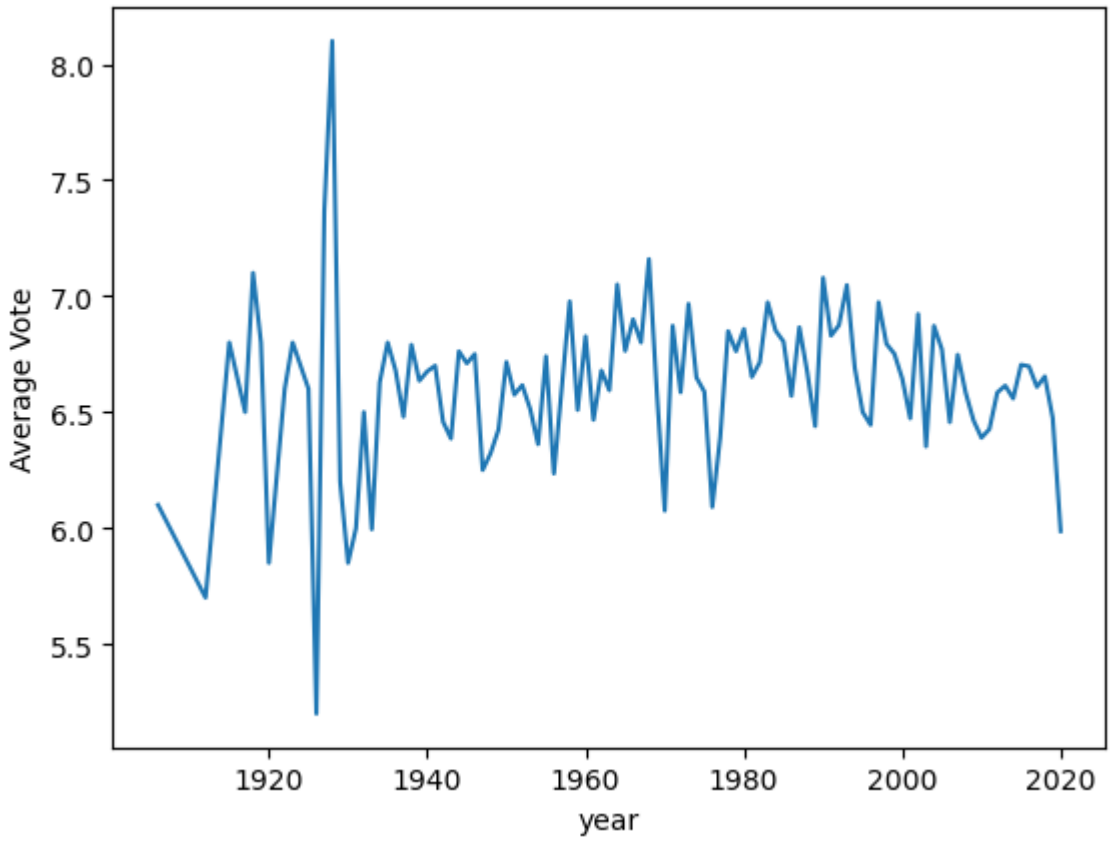
```
<AxesSubplot:xlabel='year', ylabel='Average Vote'>
```



In [ ]:

In [14]:
```
data[data['genre']=='Biography'].groupby('year').mean()['avg_vote'].plot(ylabel='Avera
```

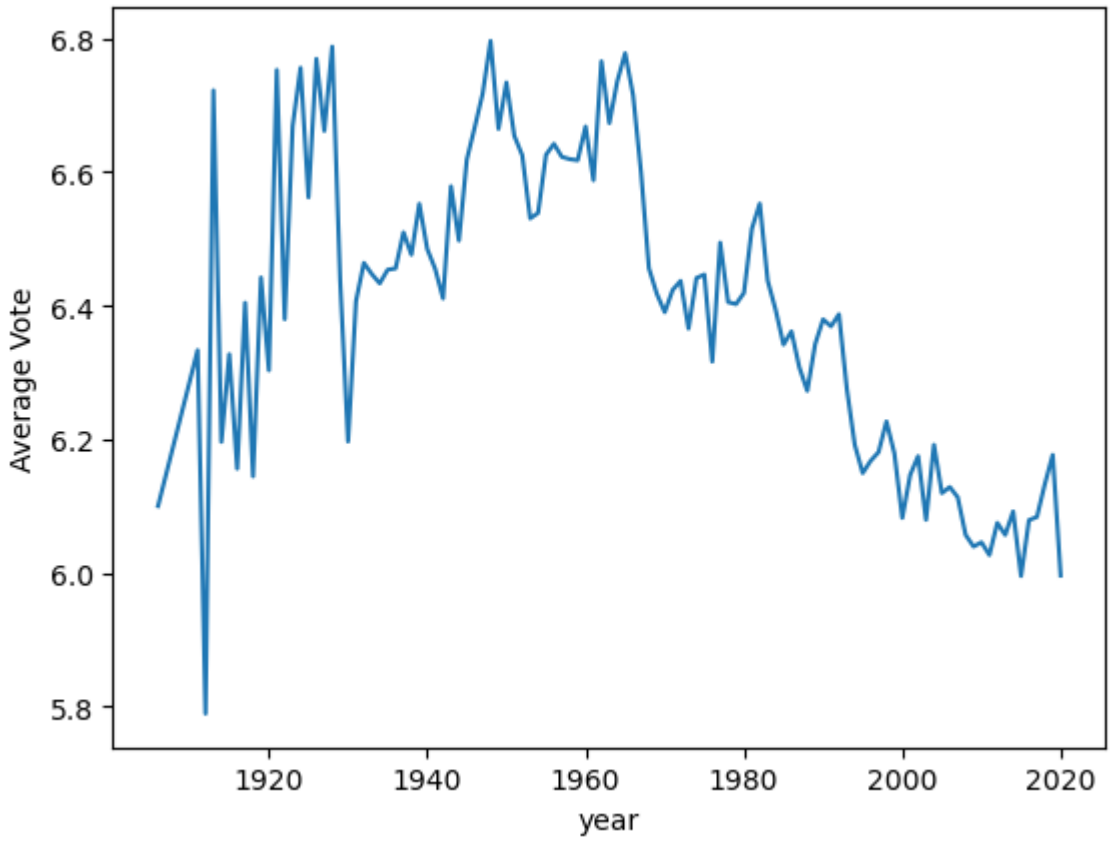
Out[14]:
```
<AxesSubplot:xlabel='year', ylabel='Average Vote'>
```



In [ ]:

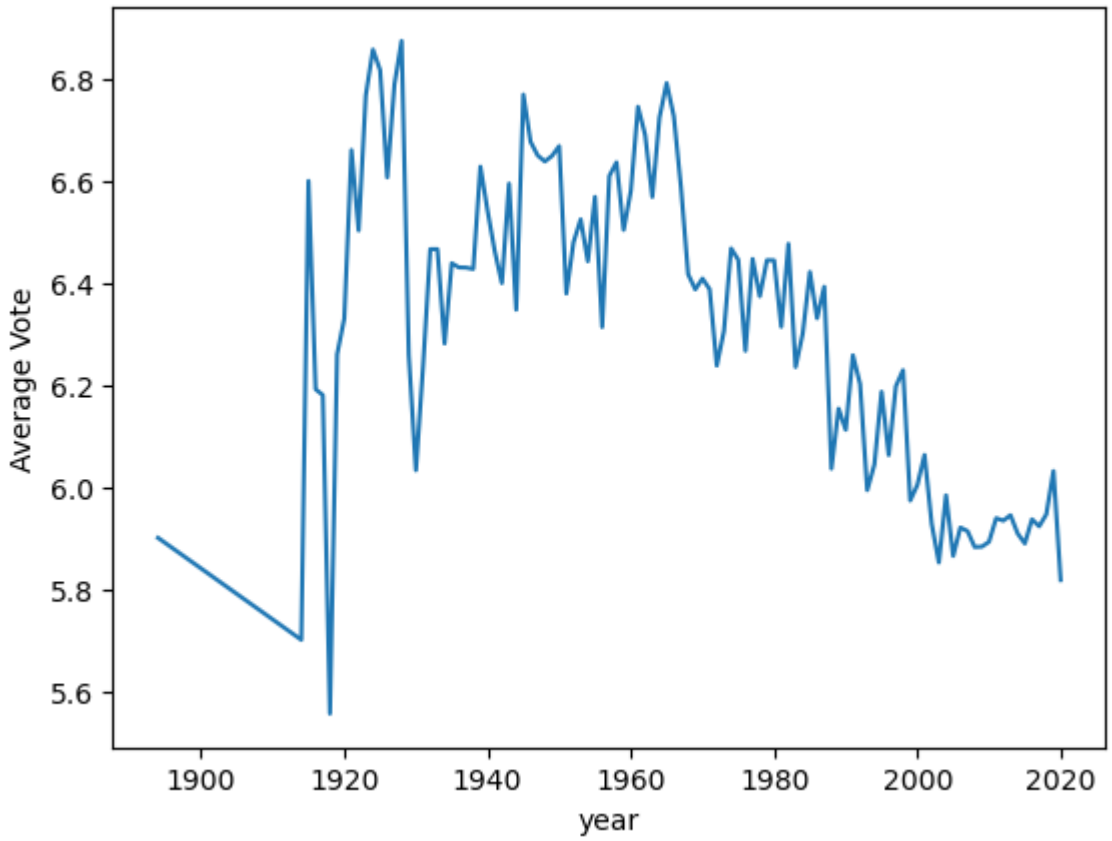
In [15]:
```
data[data['genre']=='Drama'].groupby('year').mean()['avg_vote'].plot(ylabel='Average V
```

Out[15]:
```
<AxesSubplot:xlabel='year', ylabel='Average Vote'>
```



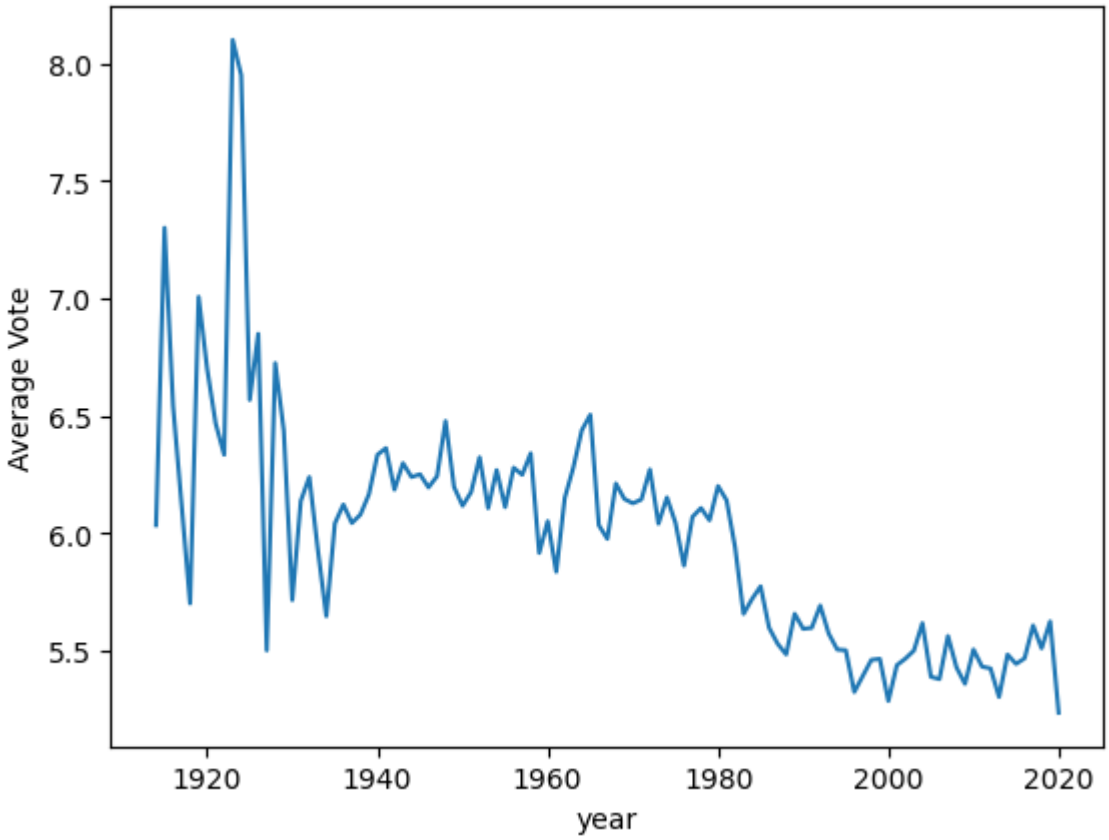
In [16]:
```
data[data['genre']=='Romance'].groupby('year').mean()['avg_vote'].plot(ylabel='Average
```

Out[16]:
```
<AxesSubplot:xlabel='year', ylabel='Average Vote'>
```



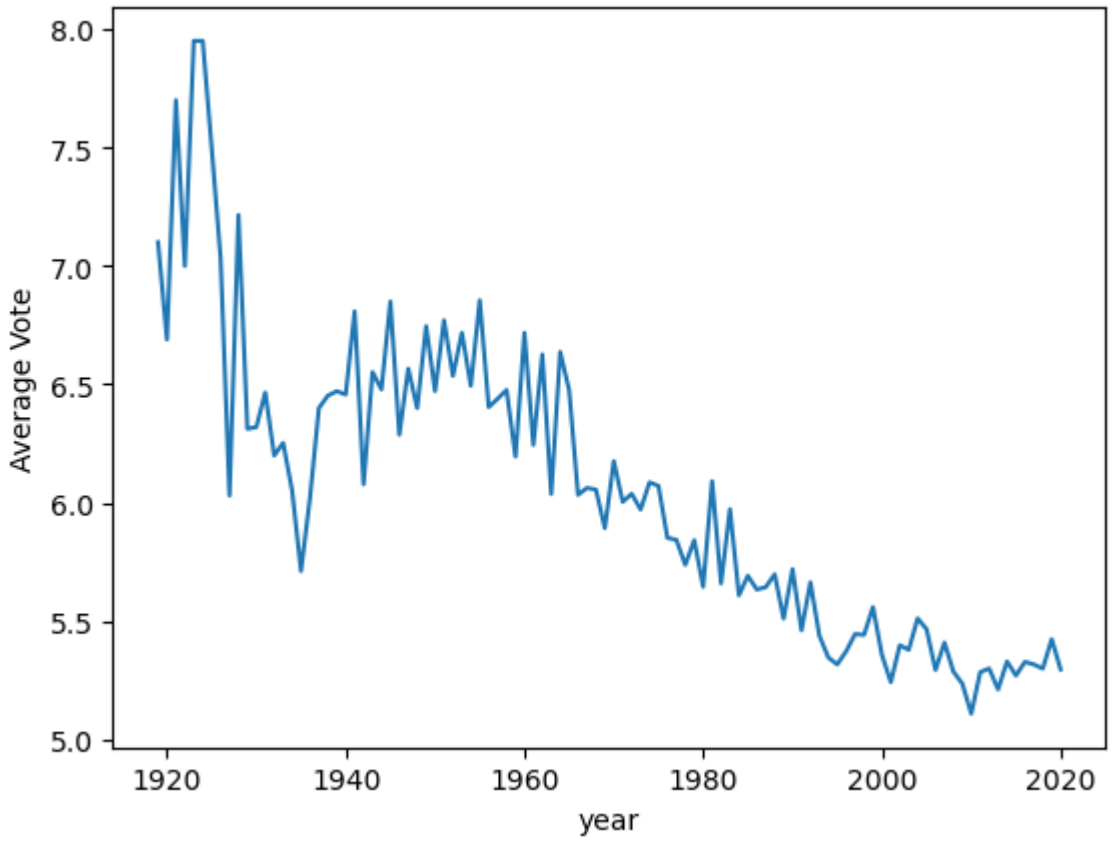
In [17]:
```
data[data['genre']=='Action'].groupby('year').mean()['avg_vote'].plot(ylabel='Average
```

Out[17]:
```
<AxesSubplot:xlabel='year', ylabel='Average Vote'>
```



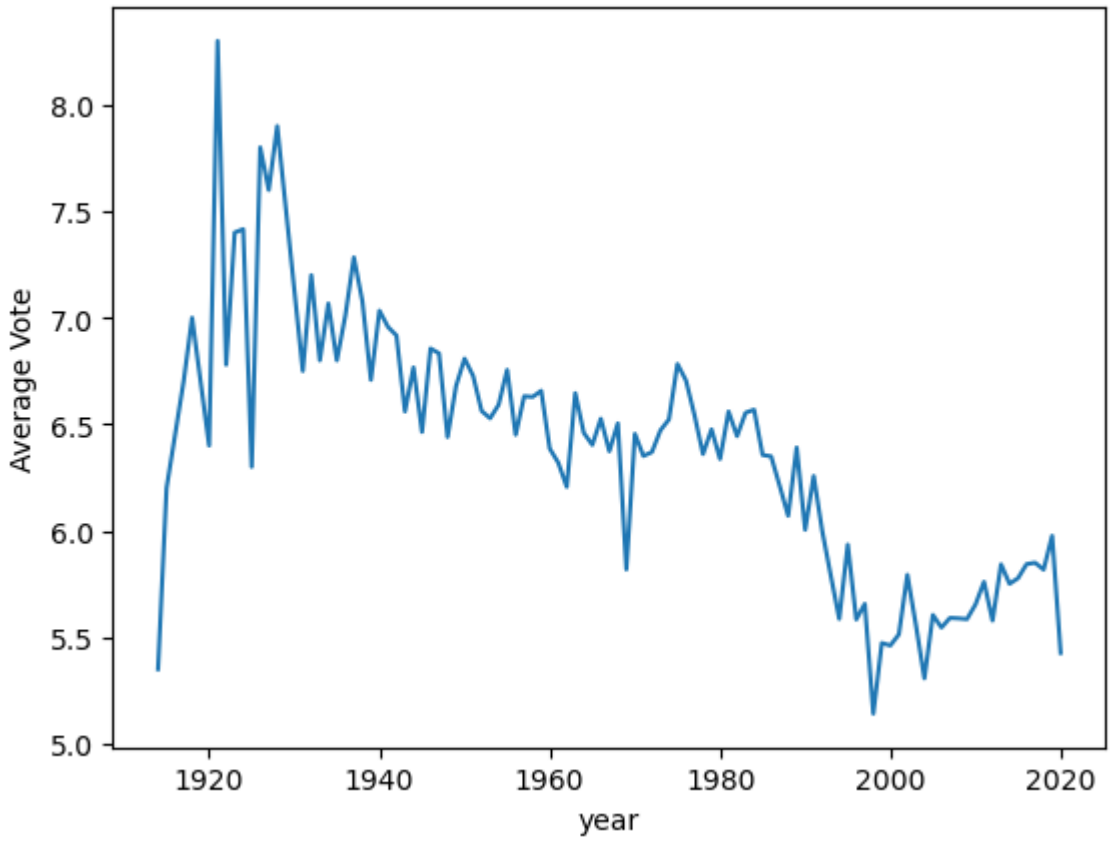
In [18]:
```
data[data['genre']=='Thriller'].groupby('year').mean()['avg_vote'].plot(ylabel='Averag
```

Out[18]:
```
<AxesSubplot:xlabel='year', ylabel='Average Vote'>
```



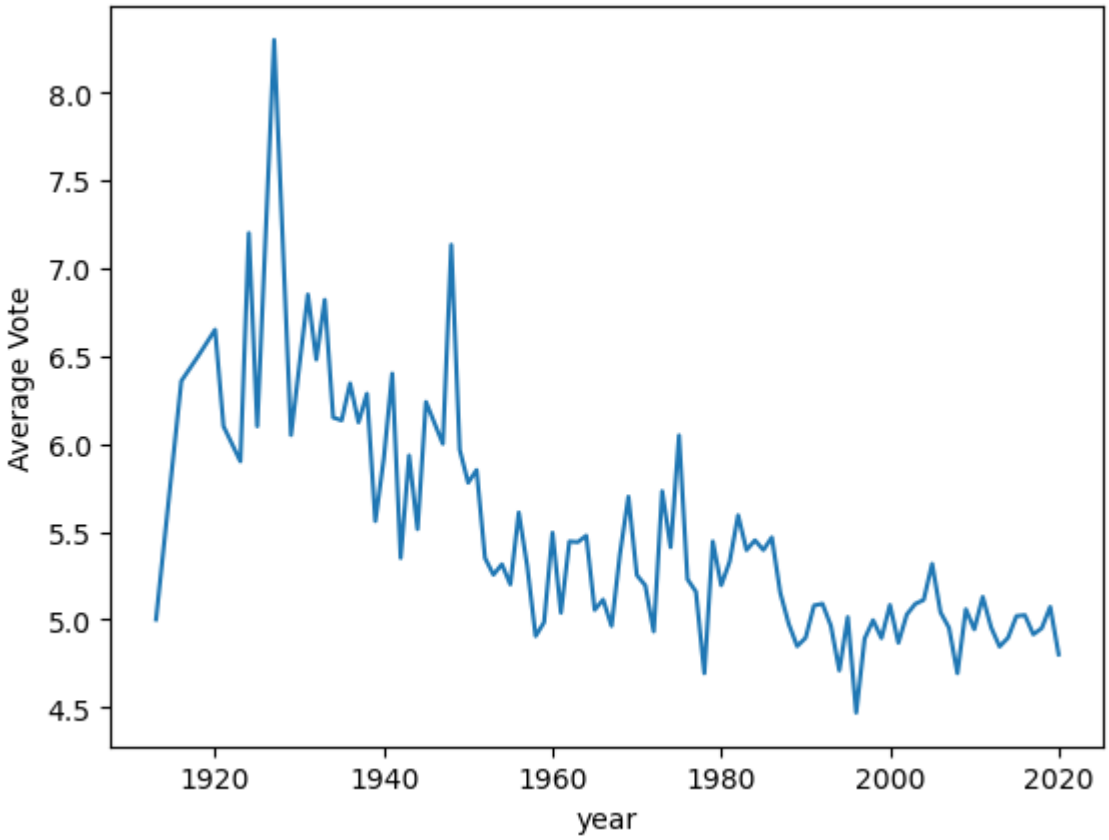
In [19]:
```
data[data['genre']=='Family'].groupby('year').mean()['avg_vote'].plot(ylabel='Average
```

Out[19]:
```
<AxesSubplot:xlabel='year', ylabel='Average Vote'>
```



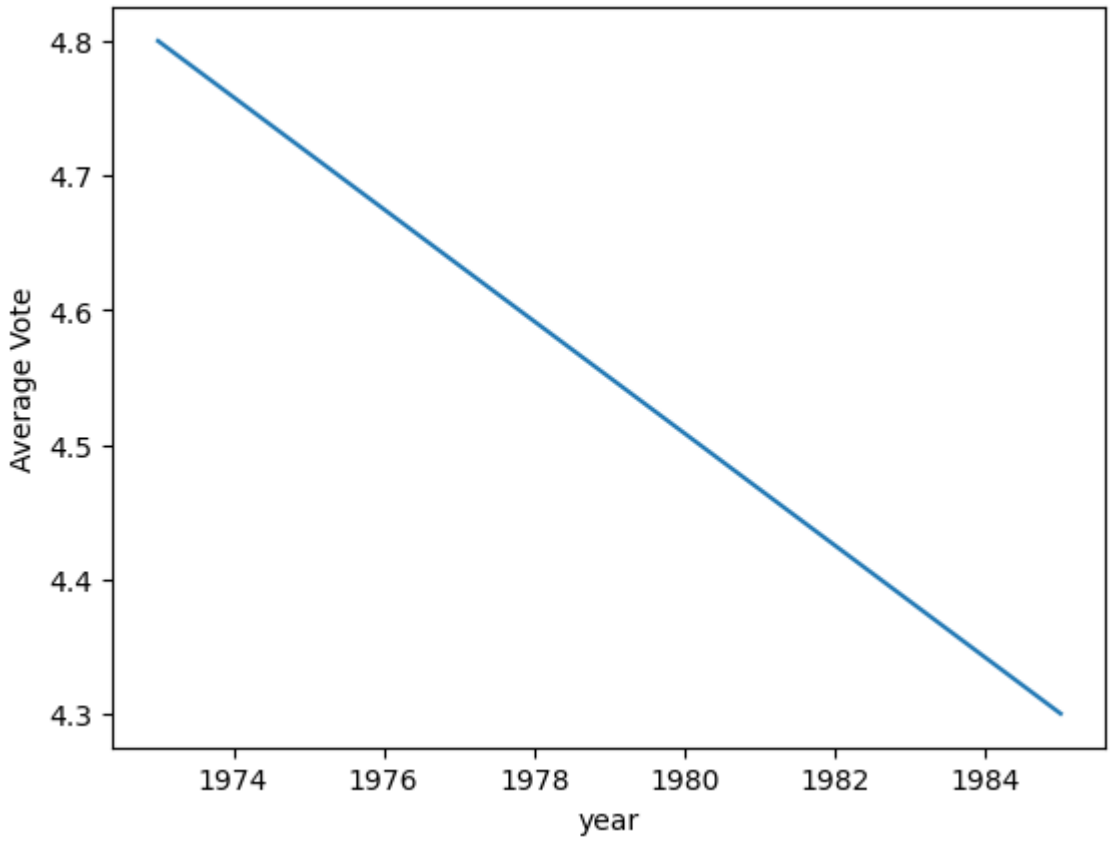
In [20]:
```
data[data['genre']=='Sci-Fi'].groupby('year').mean()['avg_vote'].plot(ylabel='Average
```

Out[20]:
```
<AxesSubplot:xlabel='year', ylabel='Average Vote'>
```



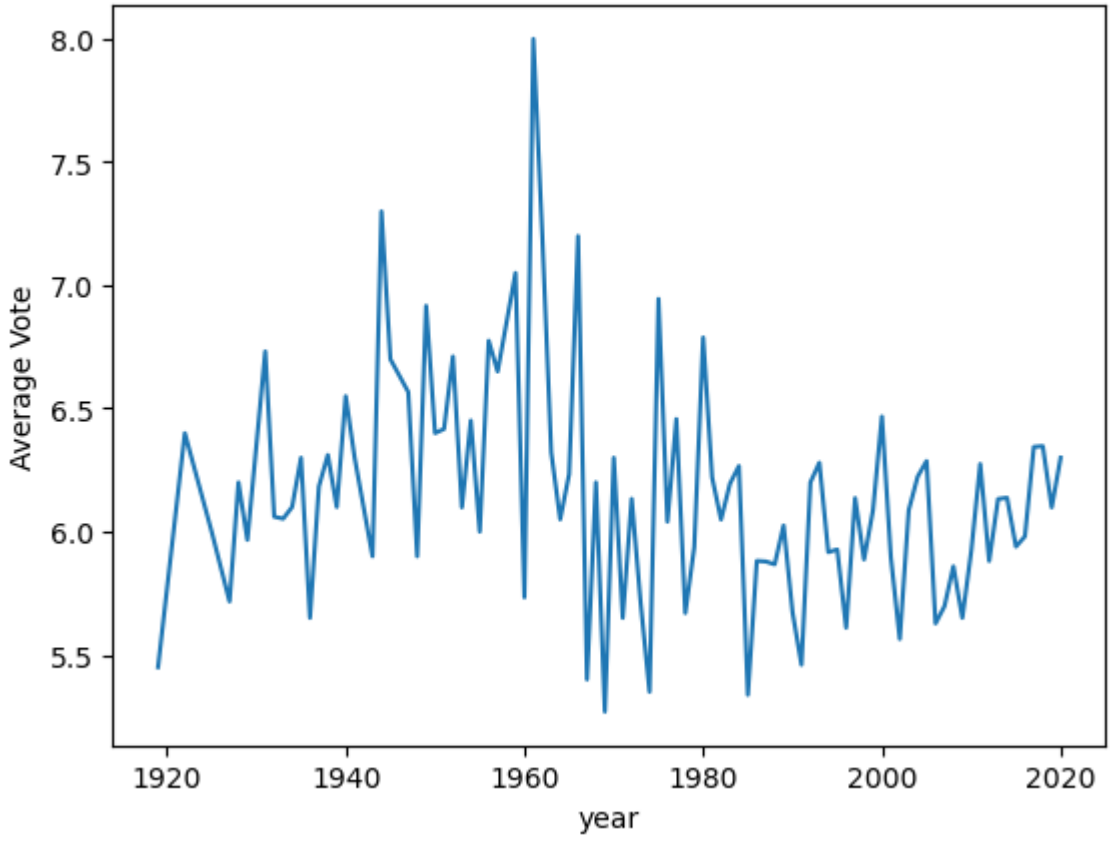
In [21]:
```
data[data['genre']=='Adult'].groupby('year').mean()['avg_vote'].plot(ylabel='Average \
```

Out[21]:
```
<AxesSubplot:xlabel='year', ylabel='Average Vote'>
```



In [22]:
```
data[data['genre']=='Sport'].groupby('year').mean()['avg_vote'].plot(ylabel='Average \
```

Out[22]:
```
<AxesSubplot:xlabel='year', ylabel='Average Vote'>
```



In [ ]:

In [23]:
```
all_df = movies.groupby('year').mean()
bio_df = data[data['genre']=='Biography'].groupby('year').mean()
sport_df = data[data['genre']=='Sport'].groupby('year').mean()

X_all = pd.DataFrame(all_df.index)
y_all = all_df['avg_vote']


X_bio = pd.DataFrame(bio_df.index)
y_bio = bio_df['avg_vote']

X_sport = pd.DataFrame(sport_df.index)
y_sport = sport_df['avg_vote']
```

In [ ]:

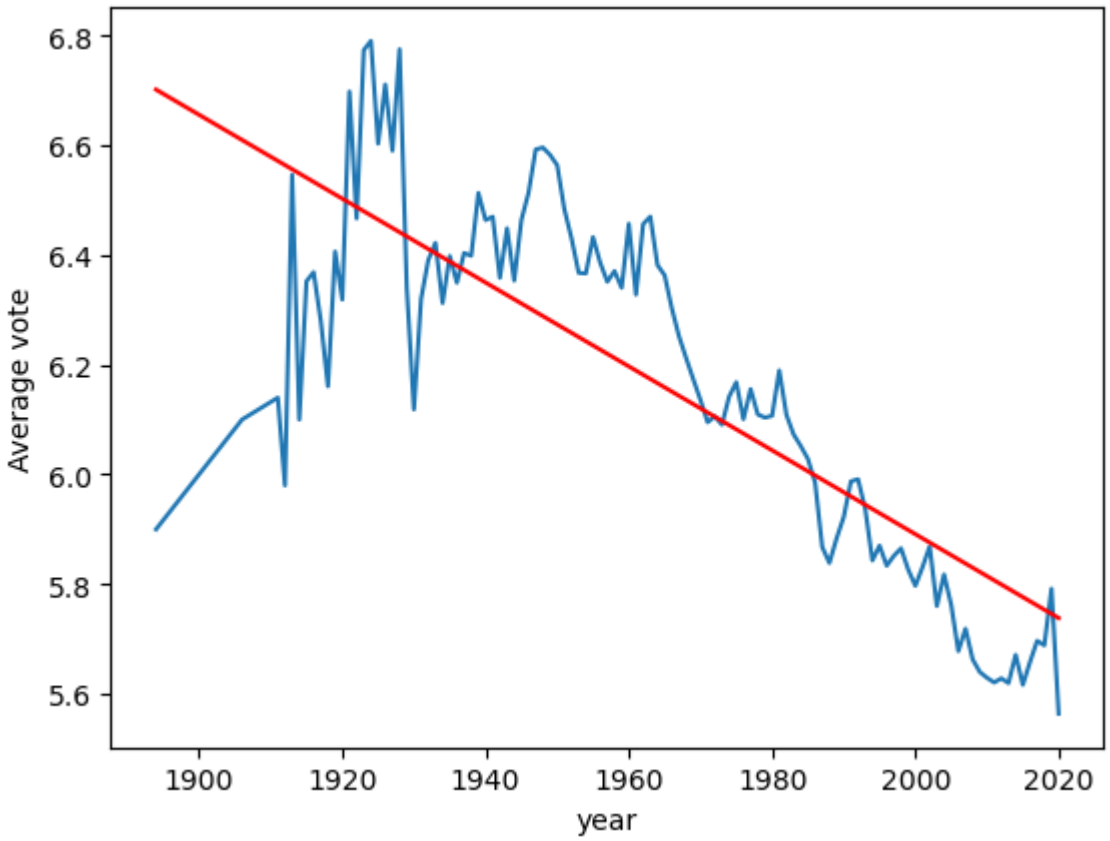
In [24]:
```
from sklearn.linear_model import LinearRegression
lin_all = LinearRegression()
lin_all.fit(X_all, y_all)

lin_bio = LinearRegression()
lin_bio.fit(X_bio, y_bio)

lin_sport = LinearRegression()
lin_sport.fit(X_sport, y_sport)
```

Out[24]:
```
LinearRegression()
```

In [ ]:

In [25]:
```
ax = all_df['avg_vote'].plot(ylabel='Average vote')
ax.plot(all_df.index, lin_all.predict(X_all), c='r')
```

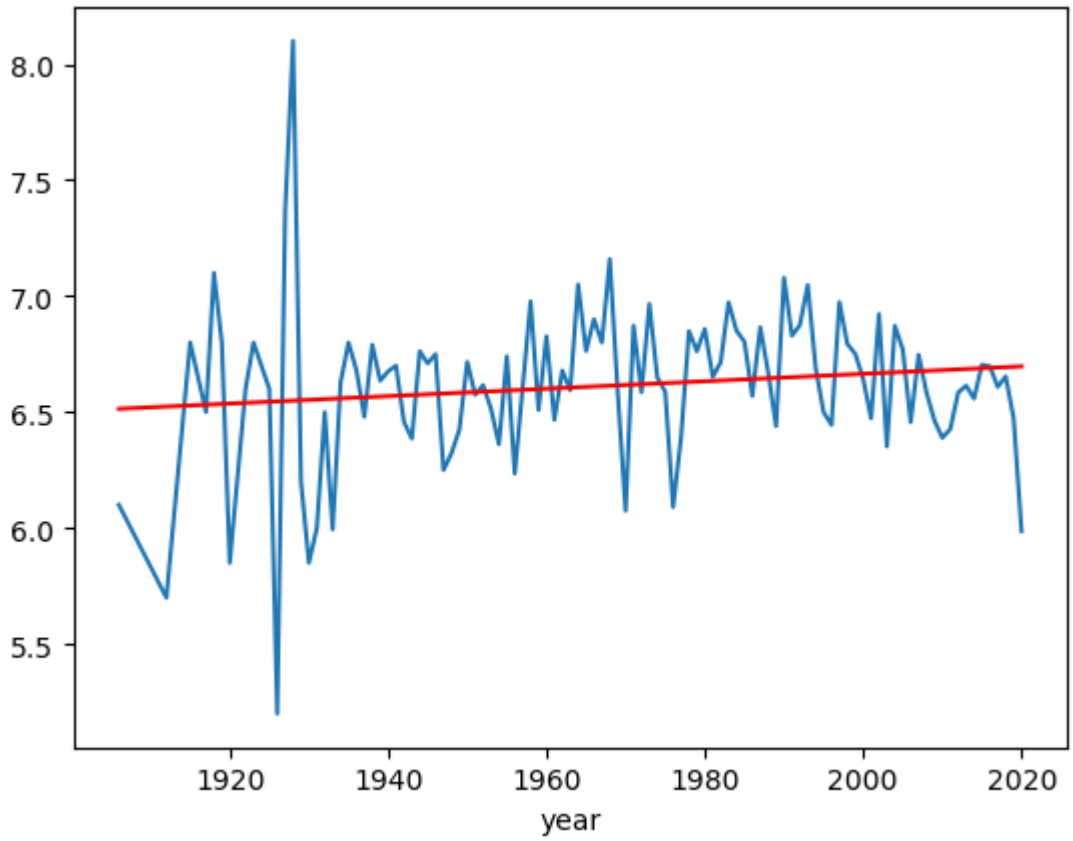
Out[25]:
```
[<matplotlib.lines.Line2D at 0x17819b3d970>]
```



In [ ]:

In [26]:
```
ax = bio_df['avg_vote'].plot()
ax.plot(bio_df.index, lin_bio.predict(X_bio), c='r')
```

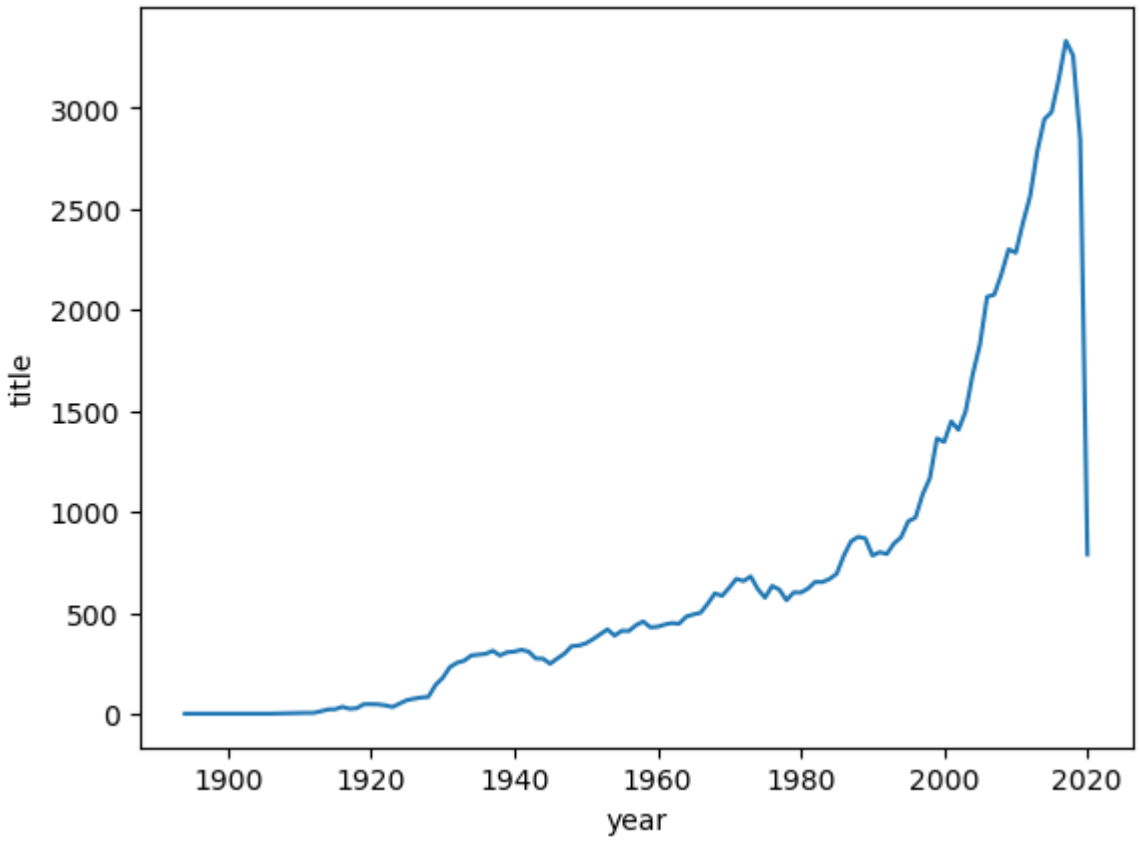
Out[26]:
```
[<matplotlib.lines.Line2D at 0x178199d2d30>]
```



In [ ]:

In [27]:
```
movies.groupby('year').count()['title'].plot(ylabel='title')
```

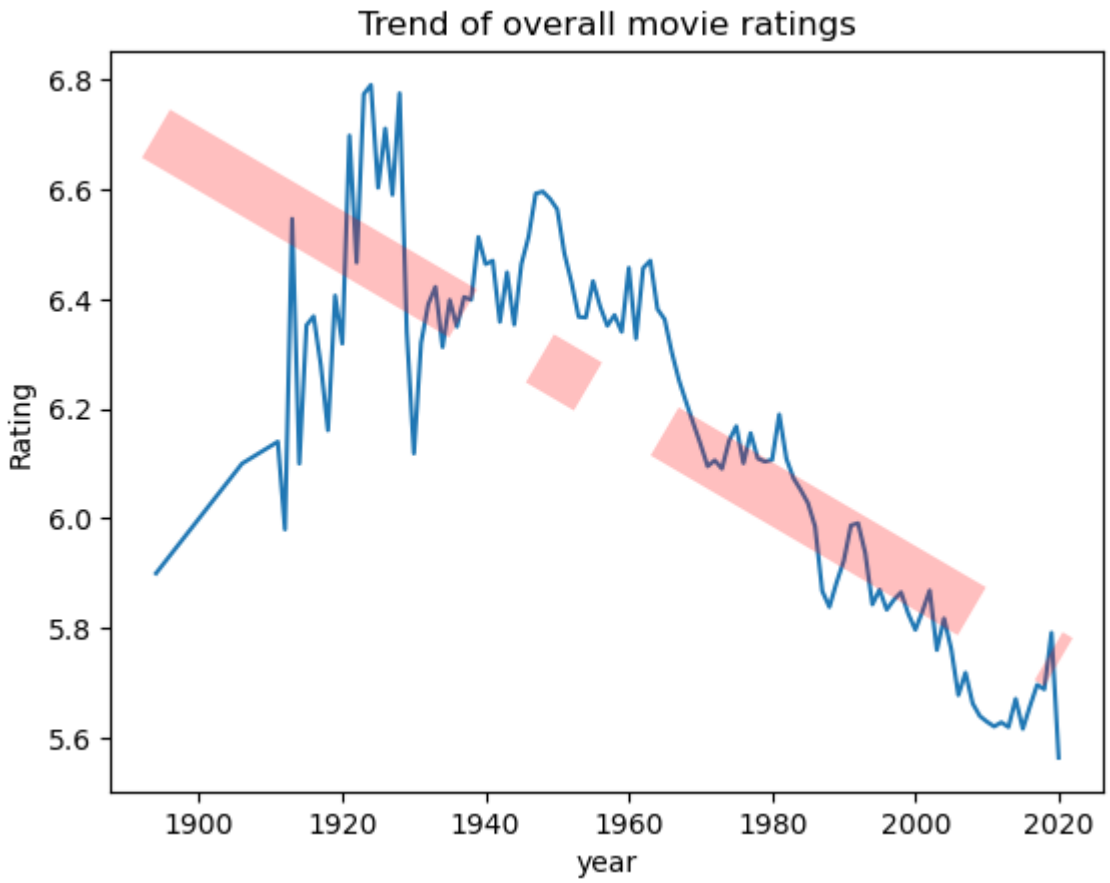
Out[27]:
```
<AxesSubplot:xlabel='year', ylabel='title'>
```



In [ ]:

In [28]:
```
ax = all_df['avg_vote'].plot()
ax.plot(all_df.index, lin_all.predict(X_all), c='r', alpha=.25, linewidth=20, linestyl
ax.set_ylabel('Rating')
ax.set_title('Trend of overall movie ratings')
```

Out[28]:
```
Text(0.5, 1.0, 'Trend of overall movie ratings')
```



In [ ]:

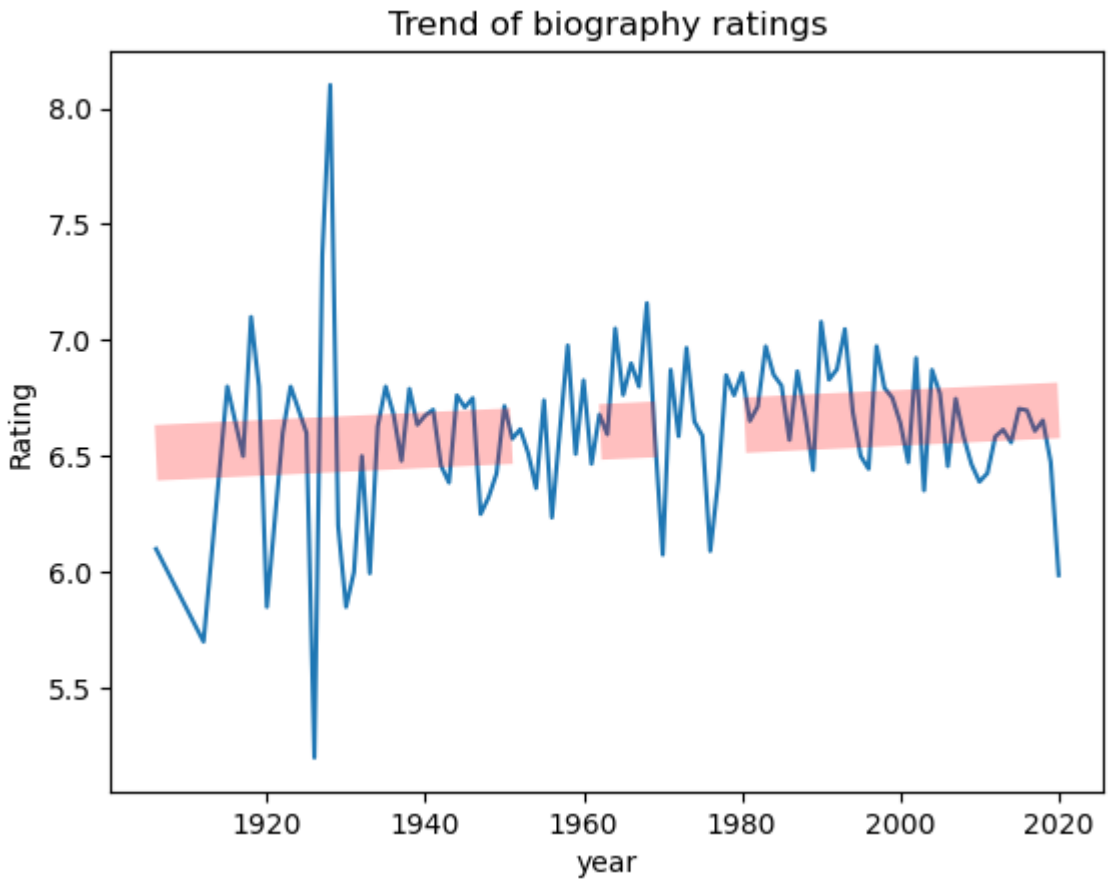
```
```

In [29]:

```
ax = bio_df['avg_vote'].plot()
ax.plot(bio_df.index, lin_bio.predict(X_bio), c='r', alpha=.25, linewidth=20, linestyl
ax.set_ylabel('Rating')
ax.set_title('Trend of biography ratings')
```

Out[29]:

```
Text(0.5, 1.0, 'Trend of biography ratings')
```



In [ ]:

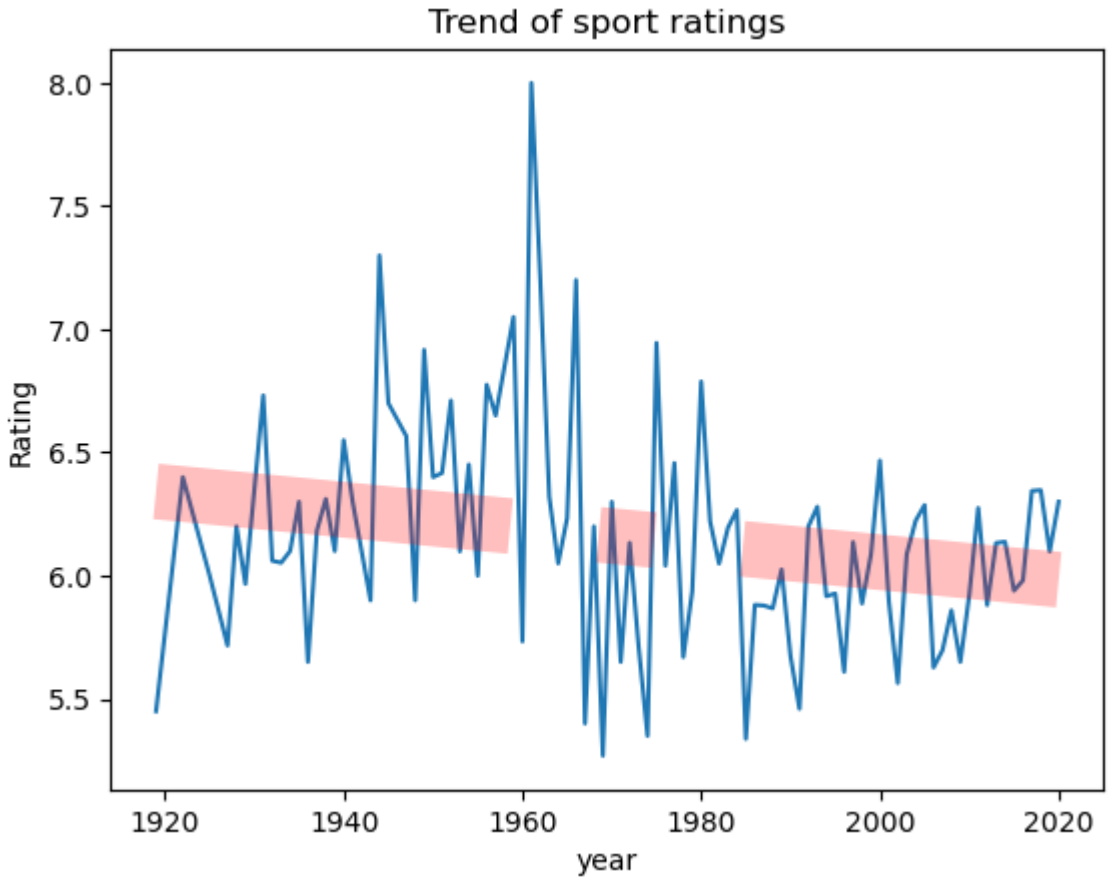
```
```

In [30]:

```
ax = sport_df['avg_vote'].plot()
ax.plot(sport_df.index, lin_sport.predict(X_sport), c='r', alpha=.25, linewidth=20, li
ax.set_ylabel('Rating')
ax.set_title('Trend of sport ratings')
```

Out[30]:

```
Text(0.5, 1.0, 'Trend of sport ratings')
```


Trend of sport ratings
8.0
7.5
7.0
6.5
6.0
5.5
Rating
1920
1940
1960
1980
2000
2020
year

In [ ]: