MovieLens Data Analysis

Chua Yeow Long

Dataset(s)

Which dataset did you use of the following:

- IMDB Movie Dataset

Motivation

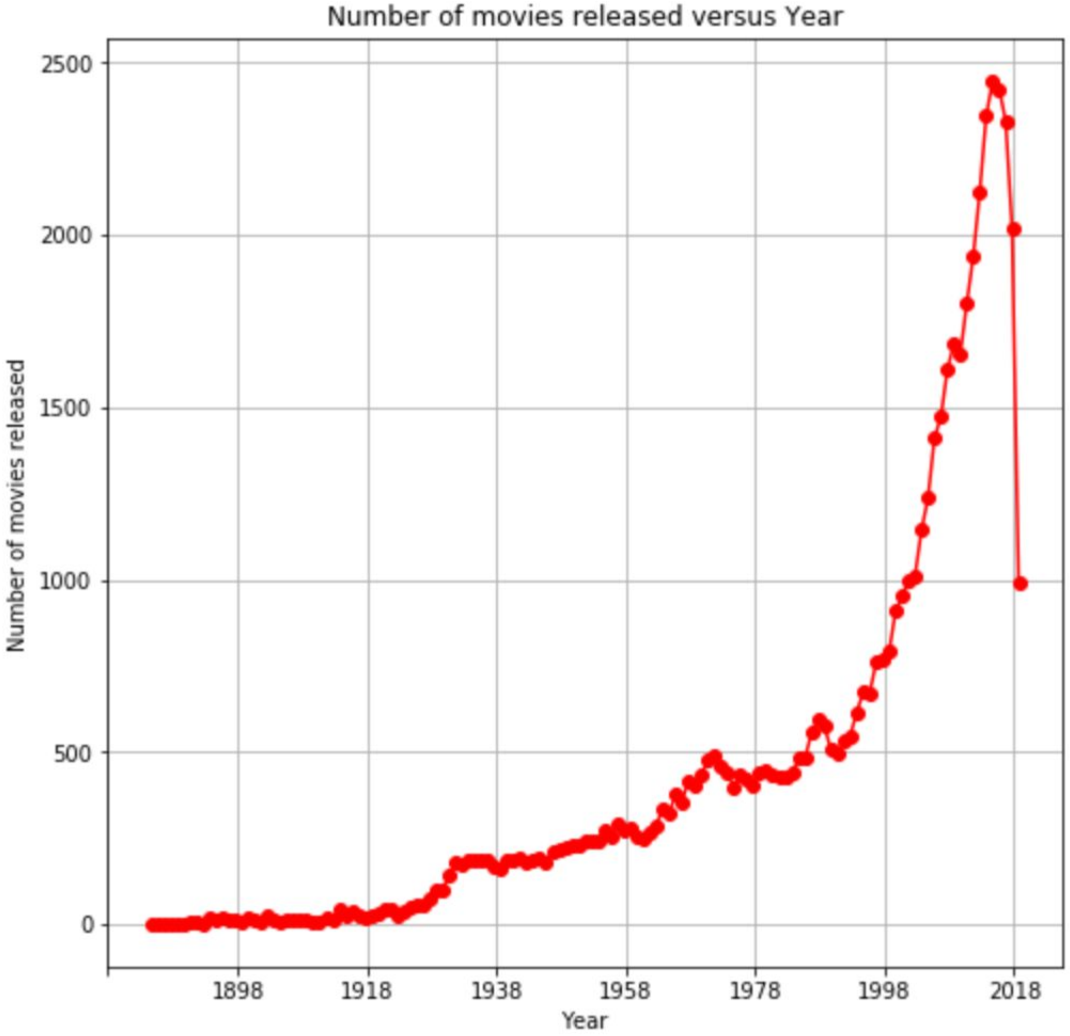
Explore the correlation between the genre of movies being released and the number of ratings given to movies.

Research Question(s)

Which movie genre tend to be rated more highly than other movie genres?

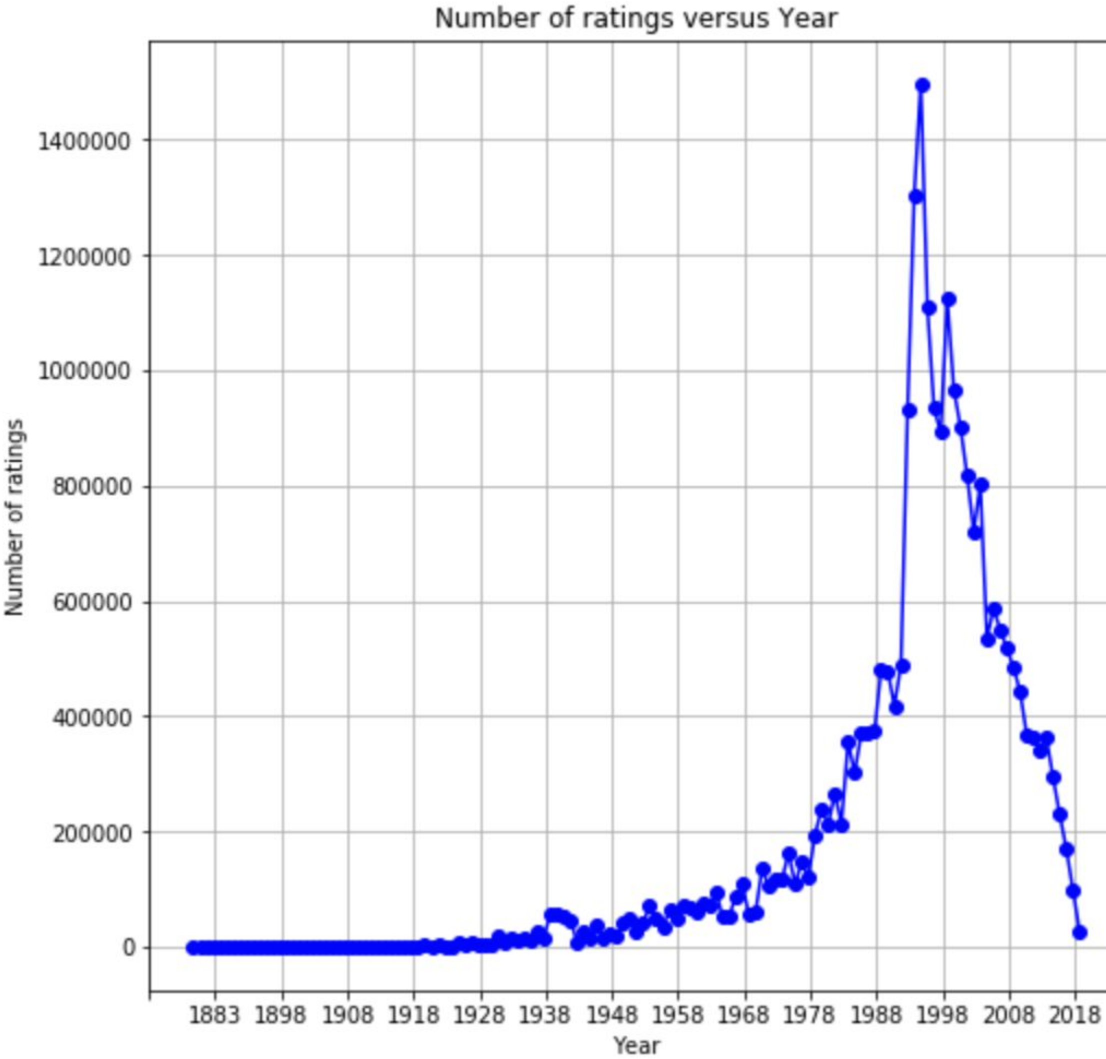
Findings

The number of movies released seems to peak around 2014

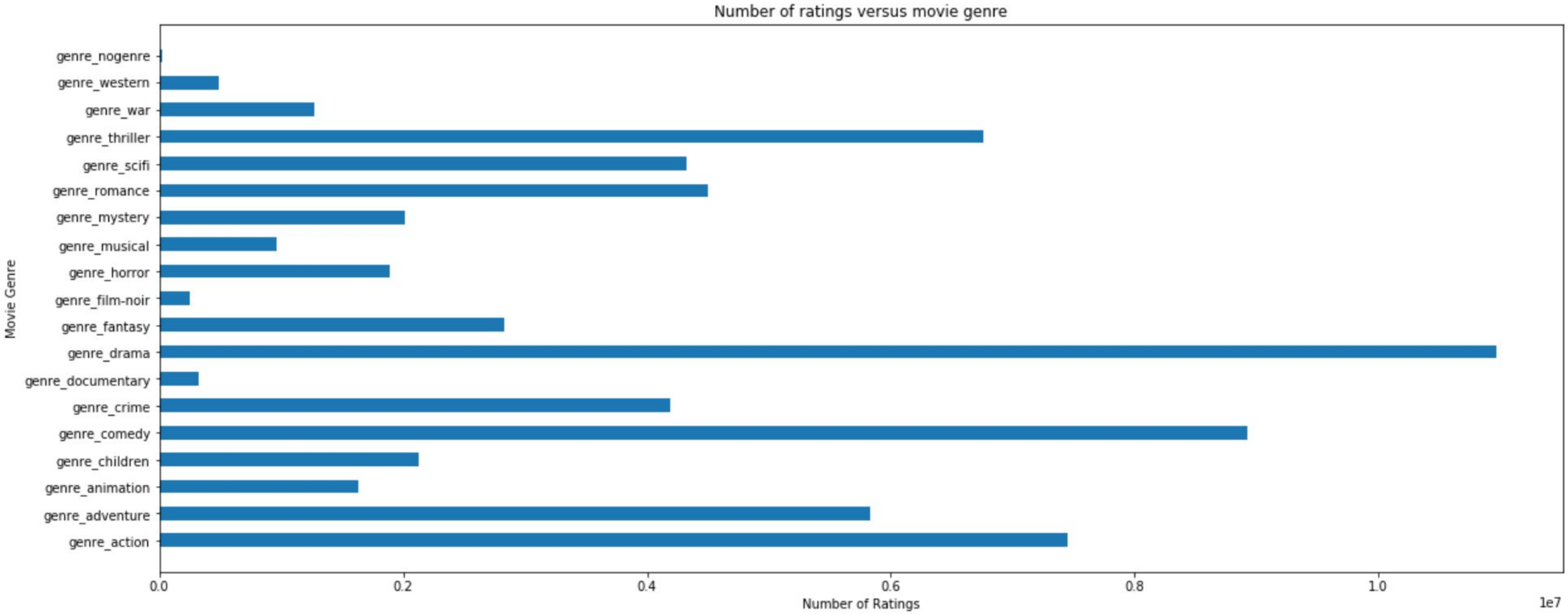


Findings

The number of ratings peak at 1995 even though there is more movies released in 2014

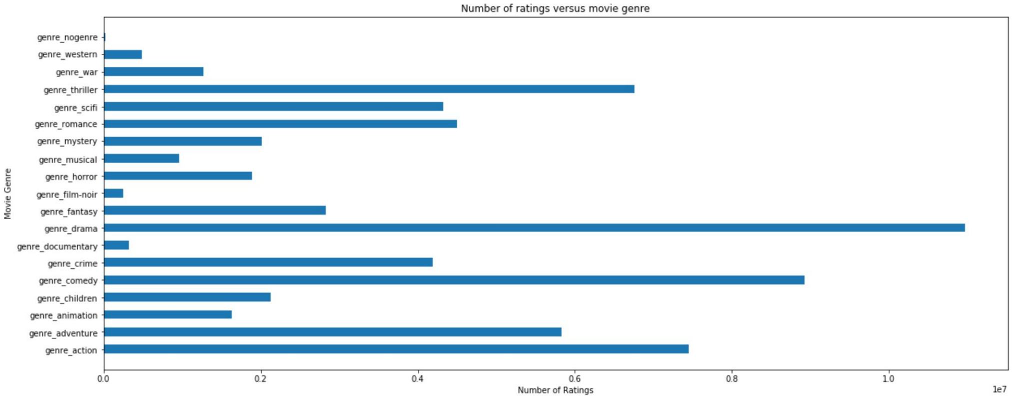


Findings


Number of ratings versus movie genre
Movie Genre
genre_nogenre
genre_western
genre_war
genre_thriller
genre_scifi
genre_romance
genre_mystery
genre_musical
genre_horror
genre_film-noir
genre_fantasy
genre_drama
genre_documentary
genre_crime
genre_comedy
genre_children
genre_animation
genre_adventure
genre_action
0.0
0.2
0.4
0.6
0.8
1.0
1e7
Number of Ratings

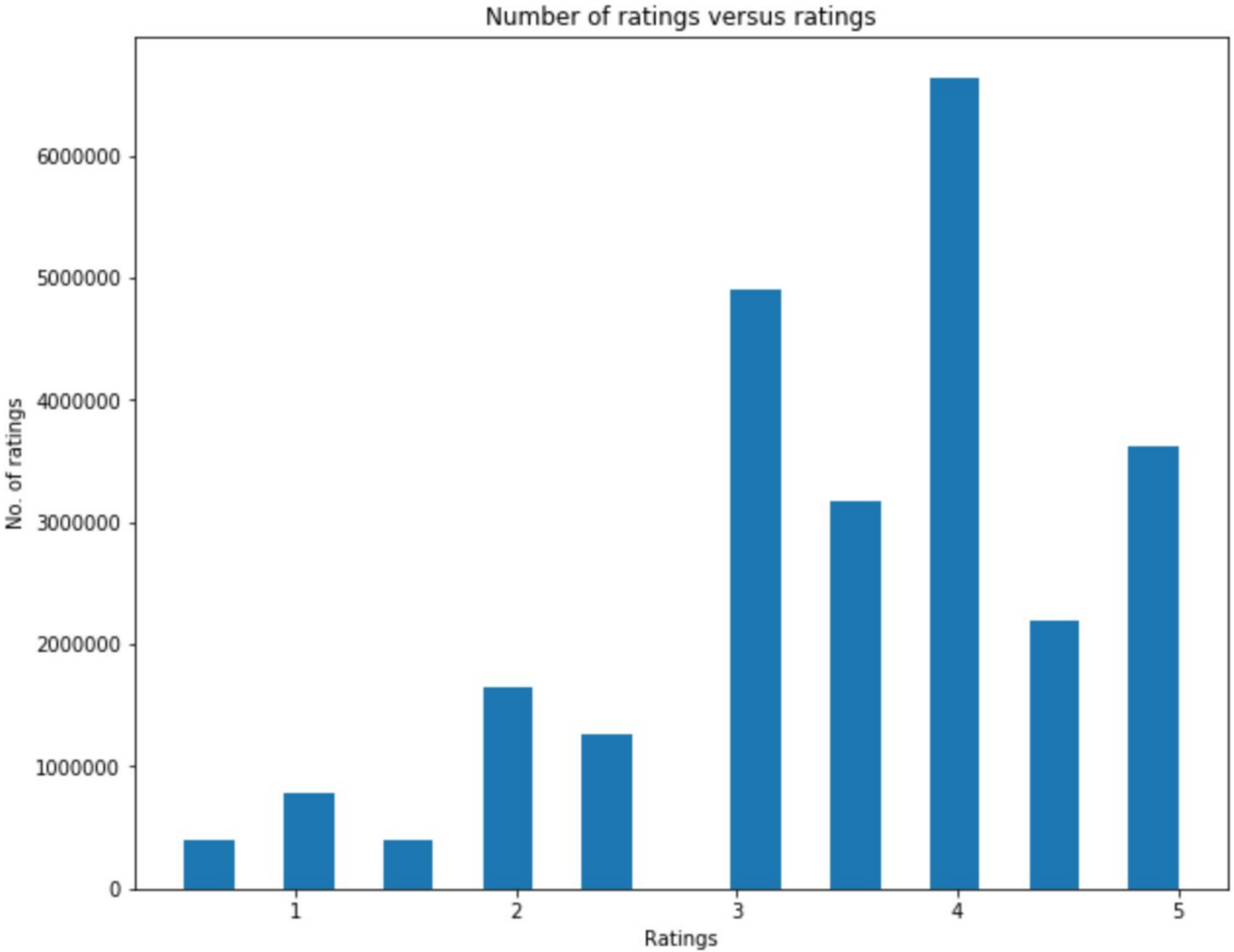
Findings

The number of ratings for thriller, drama, comedy and action are the highest

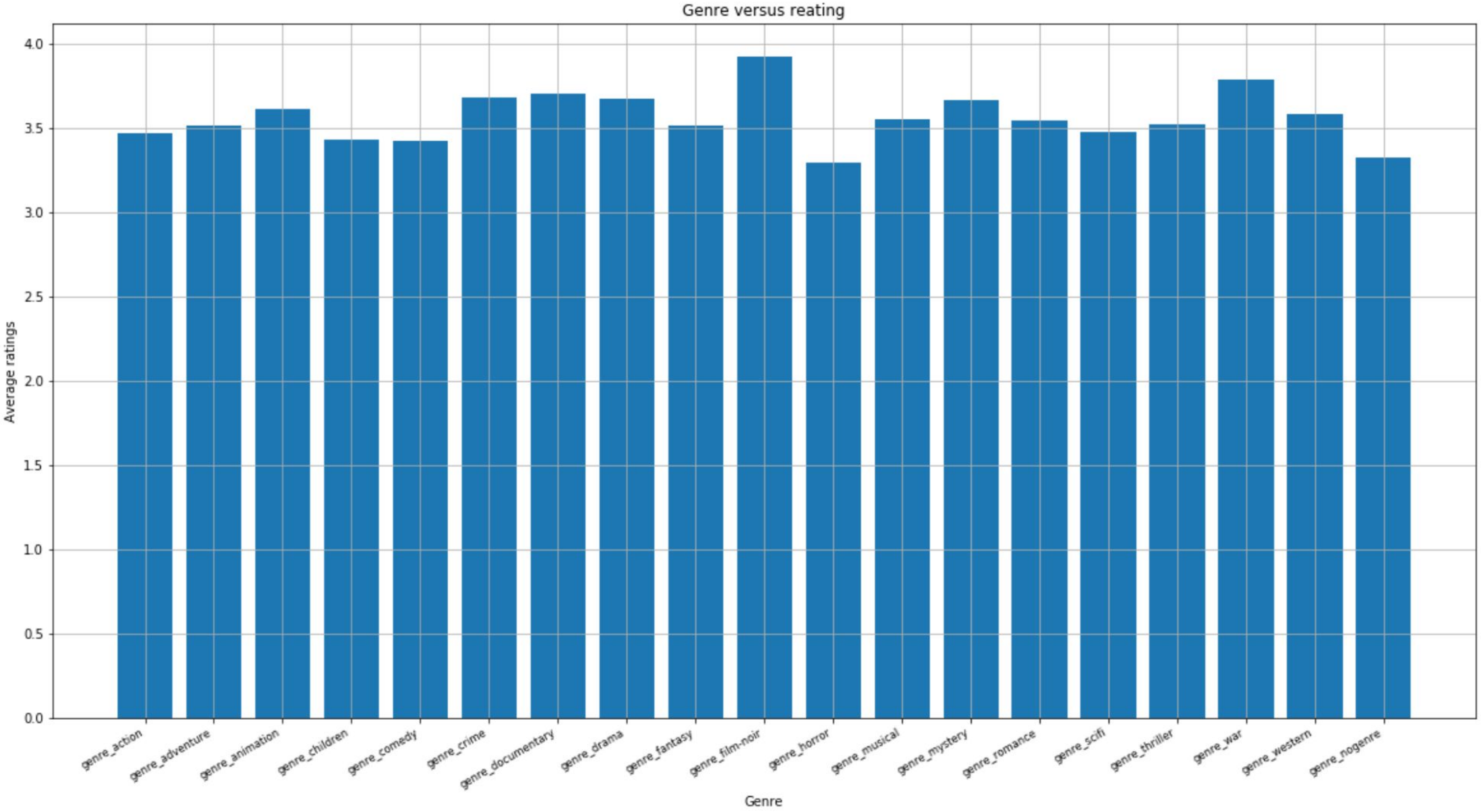


Findings

Most users seem to make an average rating of 4

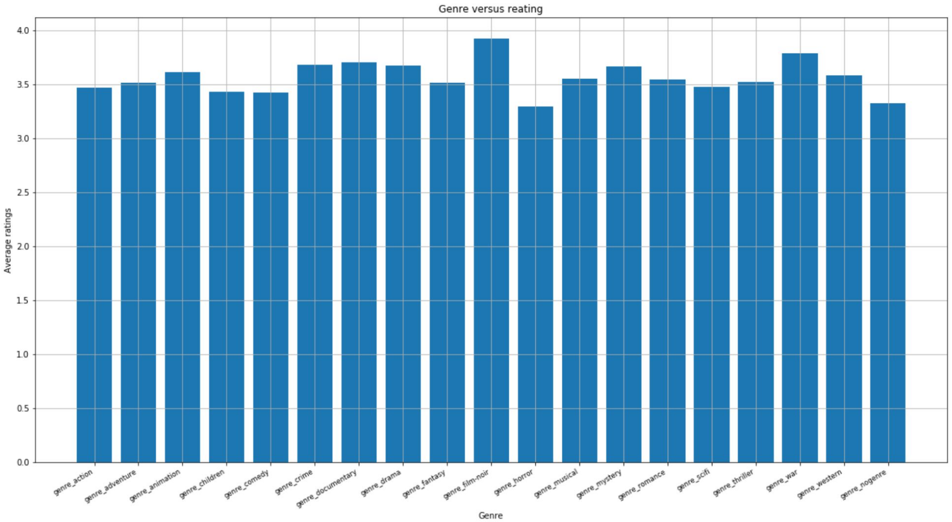


Findings


Genre versus reating
Average ratings
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
genre_action
genre_adventure
genre_animation
genre_children
genre_comedy
genre_crime
genre_documentary
genre_drama
genre_fantasy
genre_film-noir
genre_horror
genre_musical
genre_mystery
genre_romance
genre_scifi
genre_thriller
genre_war
genre_western
genre_nogenre
Genre

Findings

The ratings for each genre were about the same. The genre Film-noir receive the highest average rating and the horror genre receive the lowest rating



Acknowledgements

Thanks!

References

No applicable references

MovieLens Data Analysis

In [50]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

In [51]:

```
movies = pd.read_csv('movies.csv')
ratings = pd.read_csv('ratings.csv')
```

In [52]:

```
movies.head()
```

Out[52]:

	movieId	title	genres
0	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy
1	2	Jumanji (1995)	Adventure|Children|Fantasy
2	3	Grumpier Old Men (1995)	Comedy|Romance
3	4	Waiting to Exhale (1995)	Comedy|Drama|Romance
4	5	Father of the Bride Part II (1995)	Comedy

In [53]:

```
ratings.head()
```

Out[53]:

	userId	movieId	rating	timestamp
0	1	296	5.0	1147880044
1	1	306	3.5	1147868817
2	1	307	5.0	1147868828
3	1	665	5.0	1147878820
4	1	899	3.5	1147868510

In [54]:

```
df = pd.merge(movies, ratings, how='inner')
```

In [55]:
```
df.head()
```

Out[55]:

	movieId	title	genres	userId	rating	timestamp
0	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy	2	3.5	1141415820
1	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy	3	4.0	1439472215
2	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy	4	3.0	1573944252
3	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy	5	4.0	858625949
4	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy	8	4.0	890492517

In [56]:
```
df.isnull().sum()
```

Out[56]:
```
movieId      0
title        0
genres       0
userId       0
rating       0
timestamp    0
dtype: int64
```

In [57]:
```
df['year'] = df.title.str.extract("\((\d{4})\)", expand=True)
```

In [58]:
```
df.head()
```

Out[58]:

	movieId	title	genres	userId	rating	timestamp
0	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy	2	3.5	1141415820
1	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy	3	4.0	1439472215
2	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy	4	3.0	1573944252
3	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy	5	4.0	858625949
4	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy	8	4.0	890492517

In [59]:
```
movies_versus_year = df[['movieId', 'year']].drop_duplicates().grou
pby('year').agg('count')
```

In [60]:
```
import matplotlib.pyplot as plt
```

In [61]:
```
movies_versus_year.sample(5)
```

Out[61]:

	movieId
year	
1998	772
1930	102
1933	175
1979	443
2014	2346

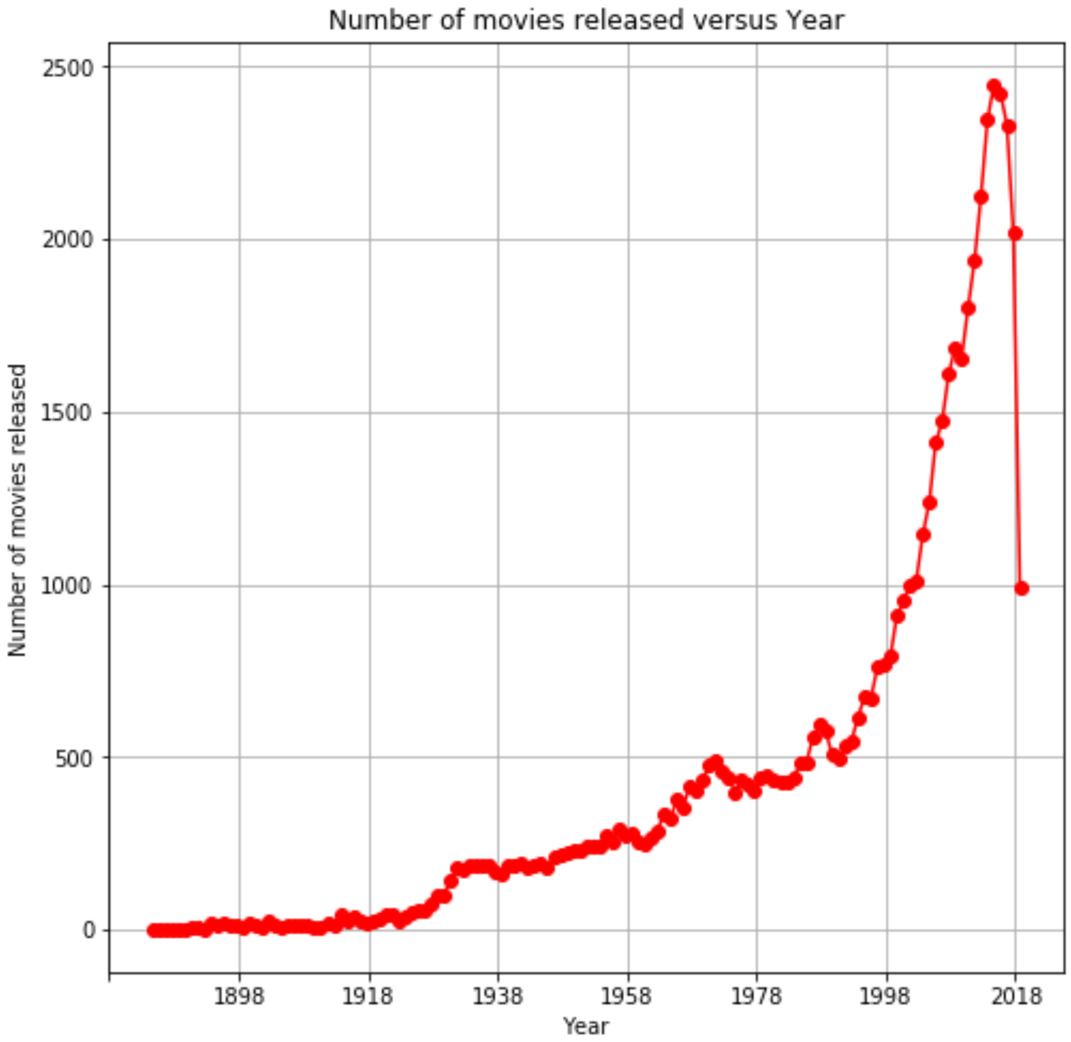
In [62]:
```
movies_versus_year.index
```

Out[62]:
```
Index(['1874', '1878', '1880', '1883', '1887', '1888', '1890', '18
91', '1892',
       '1894',
       ...
       '2010', '2011', '2012', '2013', '2014', '2015', '2016', '20
17', '2018',
       '2019'],
      dtype='object', name='year', length=135)
```

In [63]:

```
fig, ax1 = plt.subplots(figsize=(8,8))

ax1.plot(movies_versus_year.index, movies_versus_year, "r-o")
ax1.grid(None)
start, end = ax1.get_xlim()
ax1.xaxis.set_ticks(np.arange(start, end, 20))
ax1.set_xlabel('Year')
ax1.set_ylabel('Number of movies released');
plt.title('Number of movies released versus Year')
plt.show()
```

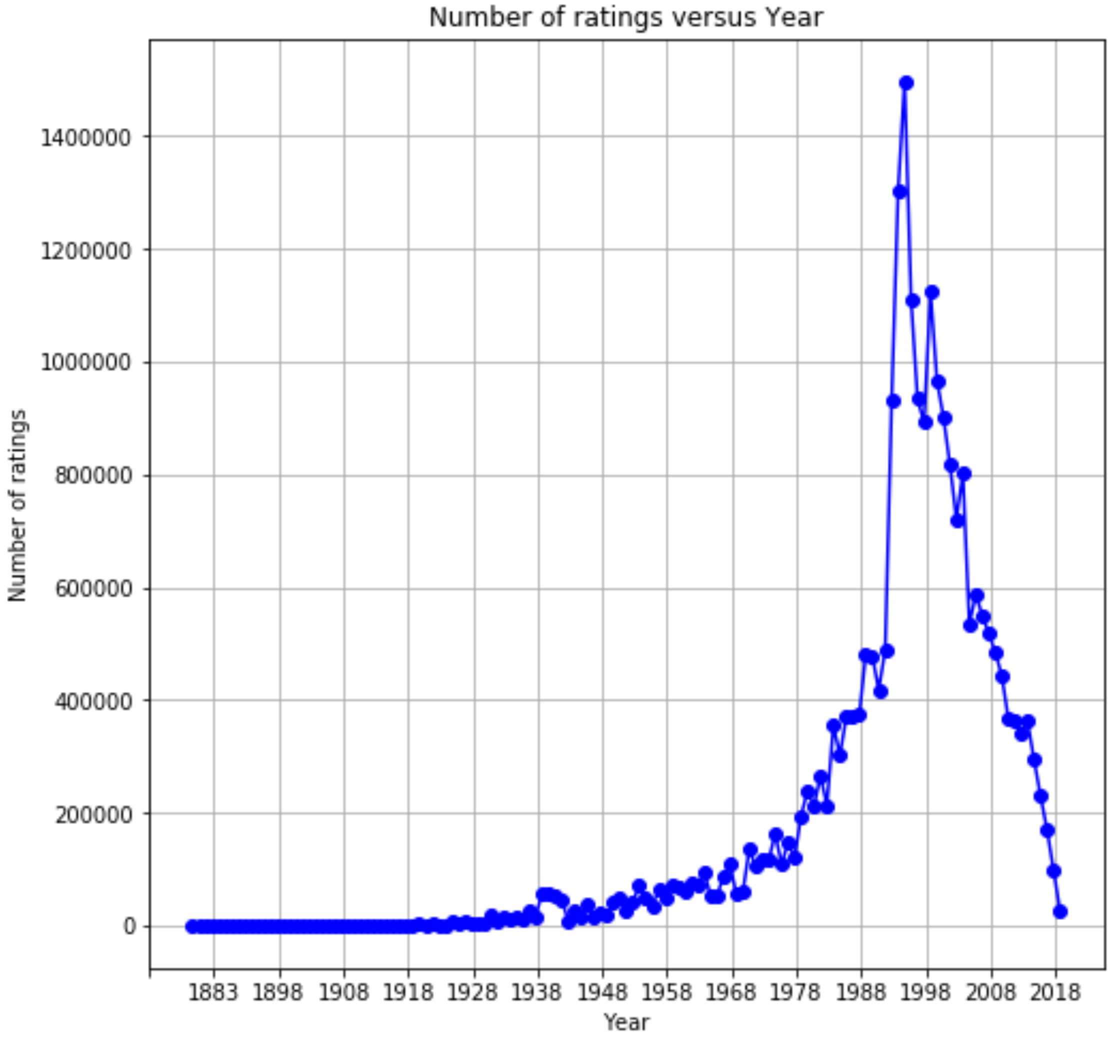


The number of movies released seem to peak around 2014

```
In [64]: ratings_versus_year = df[['rating', 'year']].groupby('year').agg('c
         ount')

         fig, ax2 = plt.subplots(figsize=(8,8))

         ax2.plot(movies_versus_year.index, ratings_versus_year, "b-o")
         ax2.grid(None)
         start, end = ax2.get_xlim()
         ax2.xaxis.set_ticks(np.arange(start, end, 10))
         ax2.set_xlabel('Year')
         ax2.set_ylabel('Number of ratings');
         plt.title('Number of ratings versus Year')
         plt.show()
```



Even though there are more movies in 2014, the total number ratings peak at 1995.

There are 19 movie genres

- Action
- Adventure
- Animation
- Children's
- Comedy
- Crime
- Documentary
- Drama
- Fantasy
- Film-Noir
- Horror
- Musical
- Mystery
- Romance
- Sci-Fi
- Thriller
- War
- Western
- (no genres listed)

In [65]:

```
df['genre_action'] = df['genres'].apply(lambda x:1 if 'Action' in x else 0)
df['genre_adventure'] = df['genres'].apply(lambda x:1 if 'Adventure' in x else 0)
df['genre_animation'] = df['genres'].apply(lambda x:1 if 'Animation' in x else 0)
df['genre_children'] = df['genres'].apply(lambda x:1 if 'Children' in x else 0)
df['genre_comedy'] = df['genres'].apply(lambda x:1 if 'Comedy' in x else 0)

df['genre_crime'] = df['genres'].apply(lambda x:1 if 'Crime' in x else 0)
df['genre_documentary'] = df['genres'].apply(lambda x:1 if 'Documentary' in x else 0)
df['genre_drama'] = df['genres'].apply(lambda x:1 if 'Drama' in x else 0)
df['genre_fantasy'] = df['genres'].apply(lambda x:1 if 'Fantasy' in x else 0)
df['genre_film-noir'] = df['genres'].apply(lambda x:1 if 'Film-Noir' in x else 0)

df['genre_horror'] = df['genres'].apply(lambda x:1 if 'Horror' in x else 0)
df['genre_musical'] = df['genres'].apply(lambda x:1 if 'Musical' in x else 0)
df['genre_mystery'] = df['genres'].apply(lambda x:1 if 'Mystery' in x else 0)
df['genre_romance'] = df['genres'].apply(lambda x:1 if 'Romance' in x else 0)
df['genre_scifi'] = df['genres'].apply(lambda x:1 if 'Sci-Fi' in x else 0)

df['genre_thriller'] = df['genres'].apply(lambda x:1 if 'Thriller' in x else 0)
df['genre_war'] = df['genres'].apply(lambda x:1 if 'War' in x else 0)
df['genre_western'] = df['genres'].apply(lambda x:1 if 'Western' in x else 0)
df['genre_nogenre'] = df['genres'].apply(lambda x:1 if '(no genres listed)' in x else 0)
```

In [66]:
```
df.head()
```

Out[66]:

	movieId	title	genres	userId	rating	timestamp
0	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy	2	3.5	1141415820
1	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy	3	4.0	1439472215
2	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy	4	3.0	1573944252
3	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy	5	4.0	858625949
4	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy	8	4.0	890492517

5 rows × 26 columns

In [67]:
```
genres_df = df.iloc[:,7:]
genres_df.head()
```

Out[67]:

	genre_action	genre_adventure	genre_animation	genre_children	genre_comedy	genre_c
0	0	1	1	1	1	
1	0	1	1	1	1	
2	0	1	1	1	1	
3	0	1	1	1	1	
4	0	1	1	1	1	

In [68]:
```
genres_df.sum(axis=0)
```

Out[68]:
```
genre_action           7446918
genre_adventure        5832424
genre_animation        1630987
genre_children         2124258
genre_comedy           8926230
genre_crime            4190259
genre_documentary       322449
genre_drama           10962833
genre_fantasy          2831585
genre_film-noir         247227
genre_horror           1892183
genre_musical           964252
genre_mystery          2010995
genre_romance          4497291
genre_scifi            4325740
genre_thriller         6763272
genre_war              1267346
genre_western           483731
genre_nogenre            26627
dtype: int64
```

In [69]:
```
genres_df.columns
```

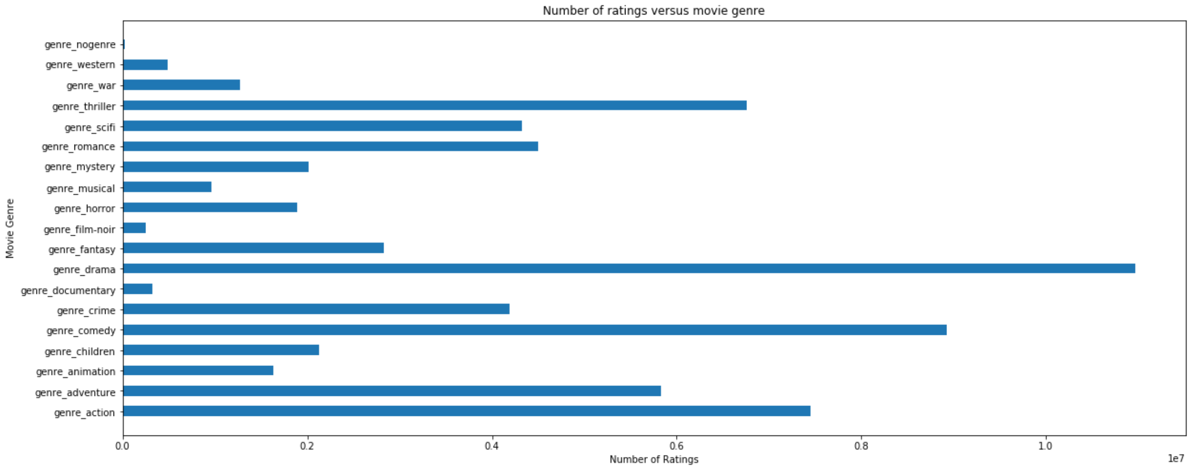
Out[69]:
```
Index(['genre_action', 'genre_adventure', 'genre_animation', 'genr
e_children',
       'genre_comedy', 'genre_crime', 'genre_documentary', 'genre_
drama',
       'genre_fantasy', 'genre_film-noir', 'genre_horror', 'genre_
musical',
       'genre_mystery', 'genre_romance', 'genre_scifi', 'genre_thr
iller',
       'genre_war', 'genre_western', 'genre_nogenre'],
      dtype='object')
```

In [70]:

```
fig, ax3 = plt.subplots(figsize=(20,8))

#ax3.plot(genres_df.columns, genres_df.sum(axis=0), "b-o")
ax3.barh(genres_df.columns, genres_df.sum(axis=0),
                     align='center',
                     height=0.5,
                     tick_label=genres_df.columns)

#ax3.grid(None)
#start, end = ax3.get_xlim()
#ax3.xaxis.set_ticks(np.arange(start, end, 10))
ax3.set_xlabel('Number of Ratings')
ax3.set_ylabel('Movie Genre');
plt.title('Number of ratings versus movie genre')
plt.show()
```



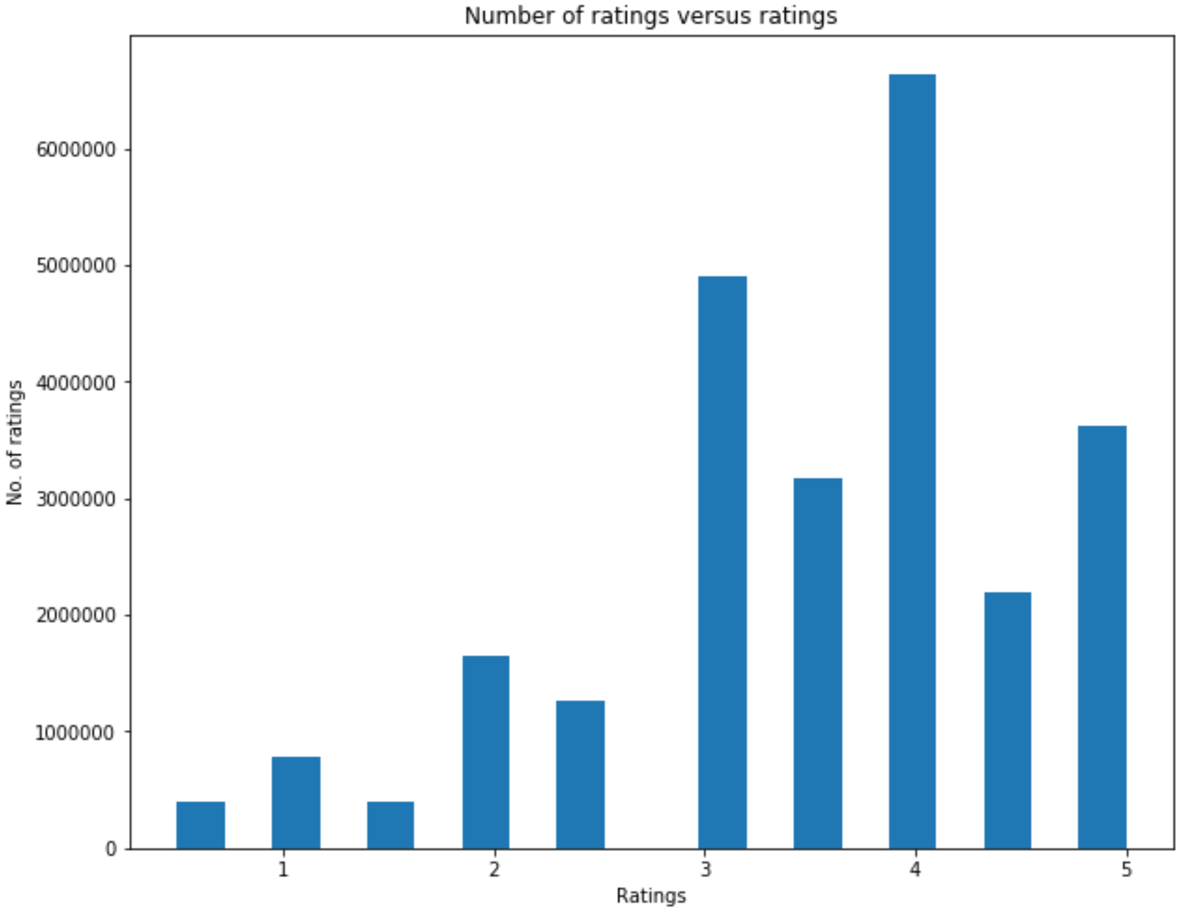
There seem to be more ratings on movie genres such as drama, comedy and action. people watching these genres rate more

In [71]:

```
fig, ax4 = plt.subplots(figsize=(10,8))

#ax3.plot(genres_df.columns, genres_df.sum(axis=0), "b-o")
ax4.hist(df['rating'],bins=20)

#ax3.grid(None)
#start, end = ax3.get_xlim()
#ax3.xaxis.set_ticks(np.arange(start, end, 10))
ax4.set_xlabel('Ratings')
ax4.set_ylabel('No. of ratings');
plt.title('Number of ratings versus ratings')
plt.show()
```



Seems like majority of users give a 4 out of 5 rating for movies

In [ ]:

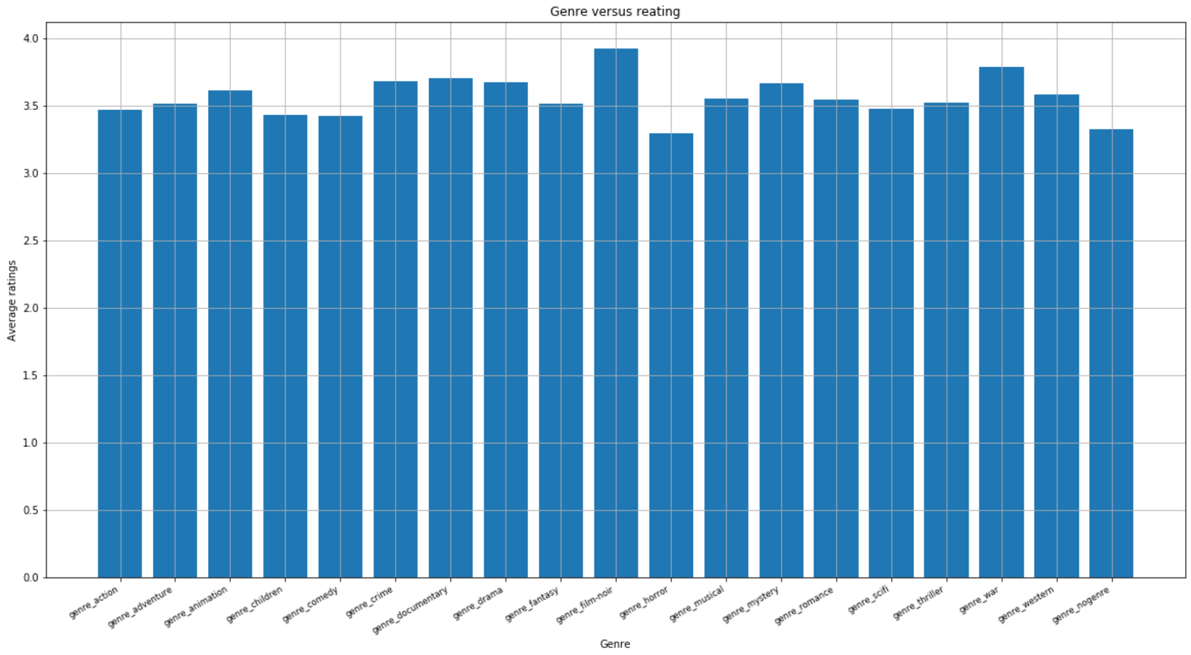
```
for index in range(19):
    #print(index)
    #df.iloc[:,index+7] = df.iloc[:,4] * df.iloc[:,index+7]
    df.iloc[:,index+7] = genres_df.iloc[:,index].mul(df.iloc[:,4],
fill_value=0)
```

```
In [85]: df.iloc[:,7:].sum(axis=0) / genres_df.sum(axis=0)
```

```
Out[85]: genre_action          3.466592
         genre_adventure       3.517445
         genre_animation       3.614946
         genre_children        3.432507
         genre_comedy          3.423993
         genre_crime           3.685044
         genre_documentary     3.705281
         genre_drama           3.677185
         genre_fantasy         3.511589
         genre_film-noir       3.925728
         genre_horror          3.293563
         genre_musical         3.554716
         genre_mystery         3.670169
         genre_romance         3.542712
         genre_scifi           3.478143
         genre_thriller        3.522964
         genre_war             3.791466
         genre_western         3.585755
         genre_nogenre         3.326379
         dtype: float64
```

In [91]:
```
fig, ax5 = plt.subplots(figsize=(20,10))

ax5.bar(genres_df.columns, df.iloc[:,7:].sum(axis=0) / genres_df.sum(axis=0))
ax5.grid(None)
#start, end = ax5.get_xlim()
#ax5.xaxis.set_ticks(np.arange(start, end, 10))
ax5.set_xlabel('Genre')
ax5.set_ylabel('Average ratings');
plt.setp(ax5.get_xticklabels(), rotation=30, horizontalalignment='right', fontsize='small')
plt.title('Genre versus reating')
plt.show()
```



The ratings for each genre were about the same. The genre Film-noir receive the highest average rating and the horror genre receive the lowest rating

In [ ]: