Directors and Genres

Can we find trends between directors and genres?

Logistics

Please adhere to the following guidelines for all submissions:

- one submission per team
- notebooks should be submitted as PDF and as raw (`.ipynb`) version
- all notebooks should be executed so they contain relevant visualizations, and other results
- try to make it as easy as possible for the TFs to get all relevant information about your work
- do not submit big data sets, please provide a readme file with a link instead
- the final report should also be submitted as pdf

Movie Data:

This data is from the imdb 5000 movie data pulled from Kaggle.

EDA continued:

In this part, we continue our EDA by asking if there is any sort of relationship we can note between directors and genres. This could be really important in our ultimate model because examining whether or not specific directors that we are familiar with have a preference for directing movies of a specific genre can give us a good start when we examine movie posters, which typically might contain the director's name on it.

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn
from collections import Counter
%matplotlib inline
```

```
/Users/Ted/anaconda2/lib/python2.7/site-packages/matplotlib/font_manage
r.py:273: UserWarning: Matplotlib is building the font cache using fc-l
ist. This may take a moment.
  warnings.warn('Matplotlib is building the font cache using fc-list. T
his may take a moment.')
```

In [2]:
```
#import
df = pd.read_csv("movie_metadata.csv")
```

In [3]:
```
df.head()
```

Out[3]:

	color	director_name	num_critic_for_reviews	duration	director_facebook_likes	actor_
0	Color	James Cameron	723.0	178.0	0.0	855.0
1	Color	Gore Verbinski	302.0	169.0	563.0	1000.(
2	Color	Sam Mendes	602.0	148.0	0.0	161.0
3	Color	Christopher Nolan	813.0	164.0	22000.0	23000
4	NaN	Doug Walker	NaN	NaN	131.0	NaN

5 rows × 28 columns

In [4]:
```
#process genres by splitting them up; at this time, we are not yet looki
ng at groupings b/c too many permutations
genre = df["genres"]
for i in xrange(len(df["genres"])):
    df["genres"].ix[i,] = df["genres"].ix[i,].split("|")
```

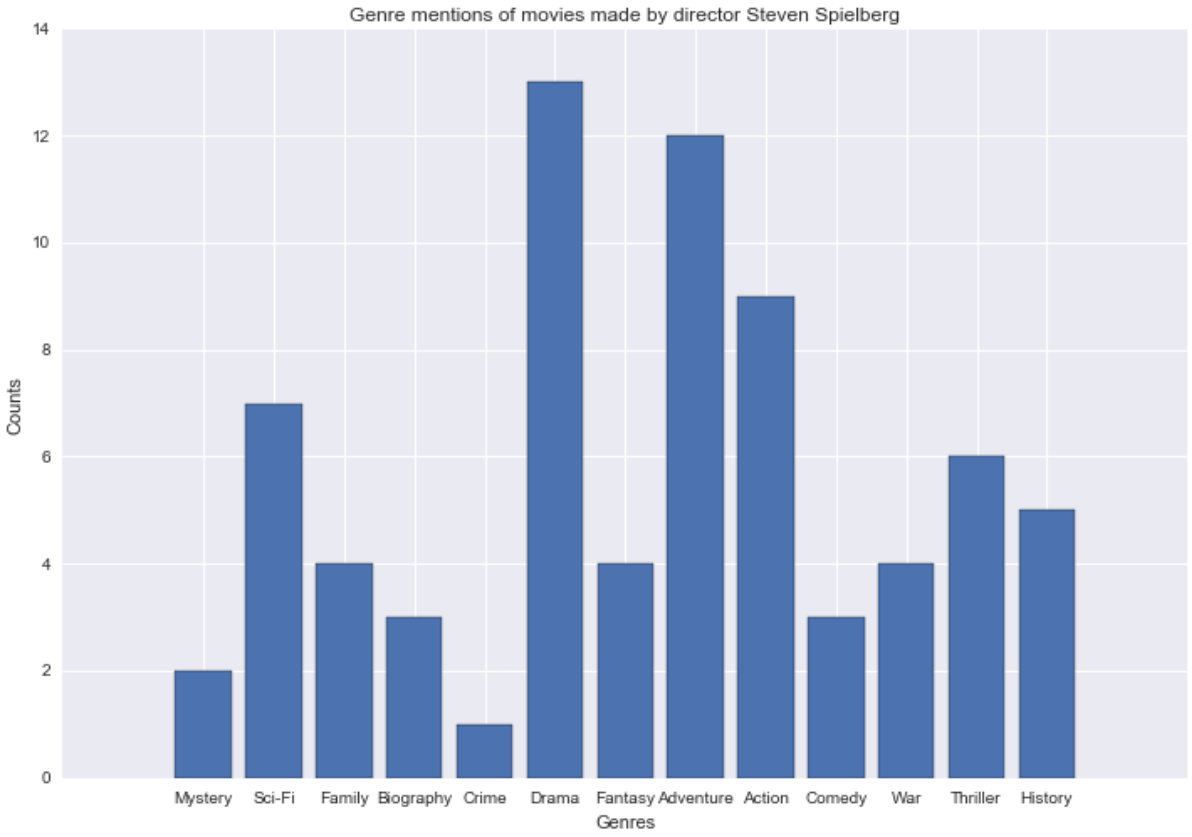
```
/Users/Ted/anaconda2/lib/python2.7/site-packages/pandas/core/indexing.p
y:549: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/pandas-d
ocs/stable/indexing.html#indexing-view-versus-copy
  self.obj[item_labels[indexer[info_axis]]] = value
```

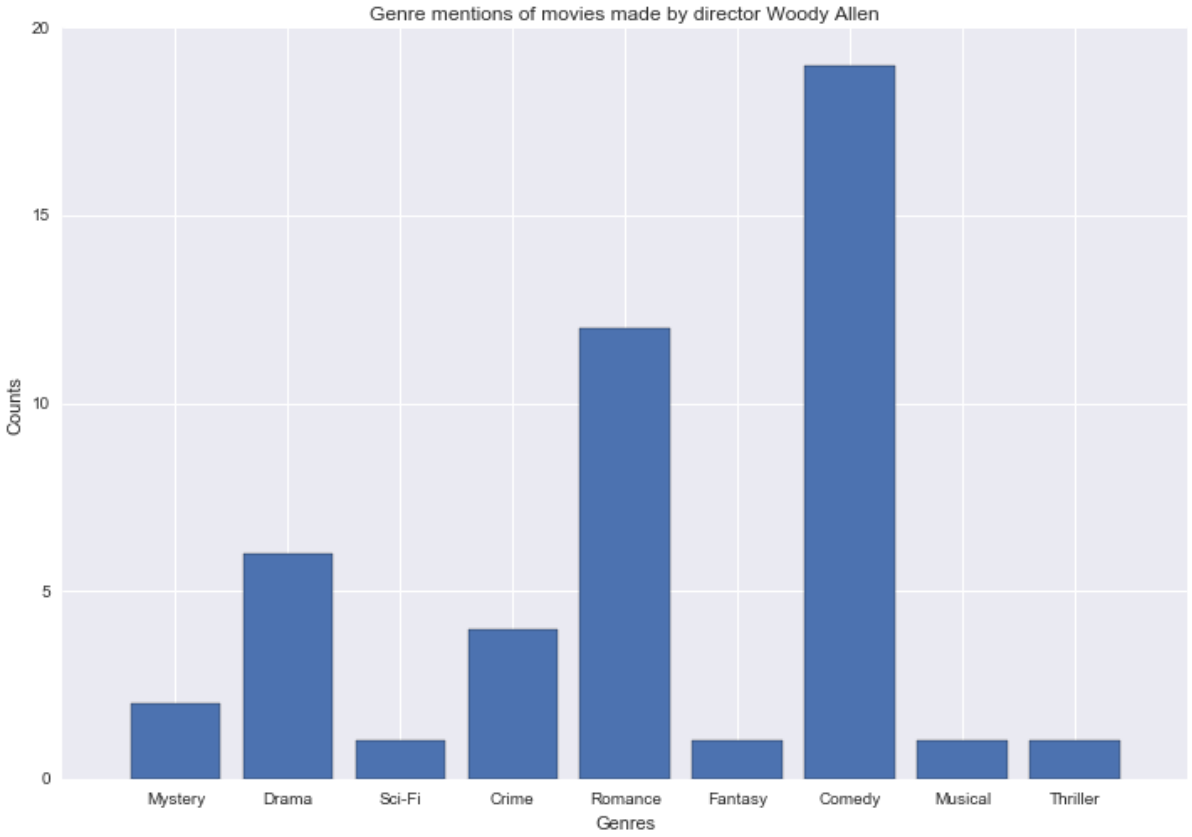
In [5]:
```
#call: function with a directors name, e.g. "plot_director("James Camero
n)"
#return: a plot with the number of films they've made that qualify as a
 certain genre
def plot_director(name):
    curr_df = df[df["director_name"] == name]
    genres = curr_df["genres"]
    genres_new = [item for sublist in list(genres) for item in sublist]
    dictionary = plt.figure(figsize=(12, 8))
    d = Counter(genres_new)
    plt.bar(range(len(d)), d.values(), align = "center")
    plt.xticks(range(len(d)), d.keys())
    plt.tick_params(axis="x", width = 20.0)
    plt.xlabel("Genres")
    plt.ylabel("Counts")
    plt.title("Genre mentions of movies made by director " + name)
    plt.show()
```

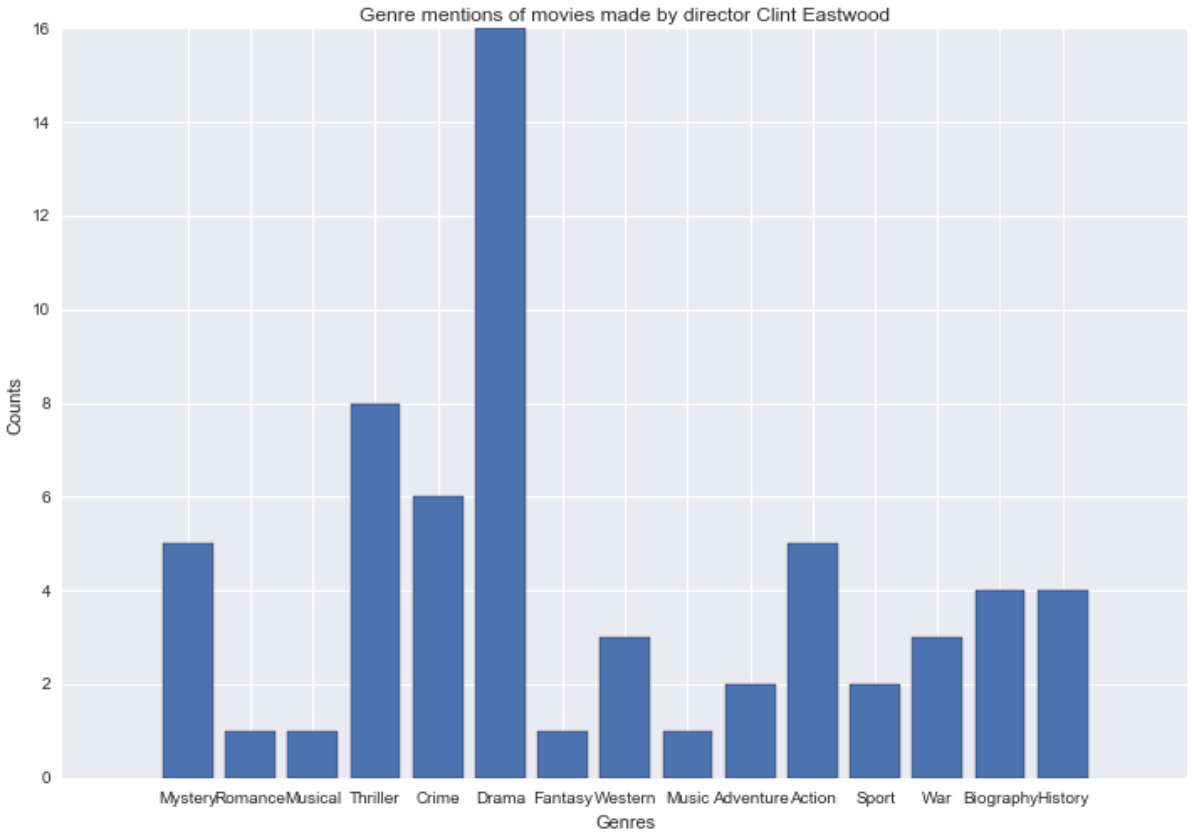
In [6]:
```
#return the top 10 directors that show up the most in this dataset
import operator
sorted(Counter(df["director_name"]).items(), key =
operator.itemgetter(1))[::-1][1:11]
```

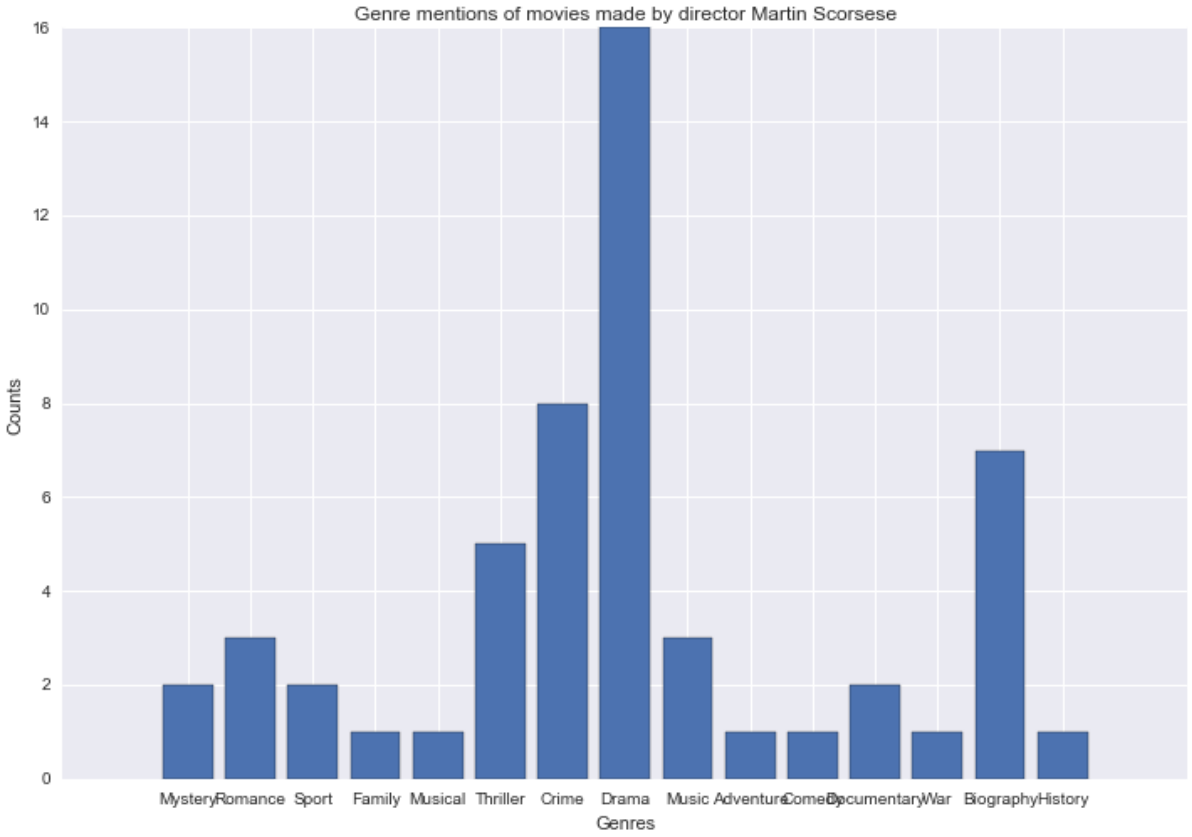
Out[6]:
```
[('Steven Spielberg', 26),
 ('Woody Allen', 22),
 ('Clint Eastwood', 20),
 ('Martin Scorsese', 20),
 ('Ridley Scott', 17),
 ('Steven Soderbergh', 16),
 ('Spike Lee', 16),
 ('Tim Burton', 16),
 ('Renny Harlin', 15),
 ('Oliver Stone', 14)]
```

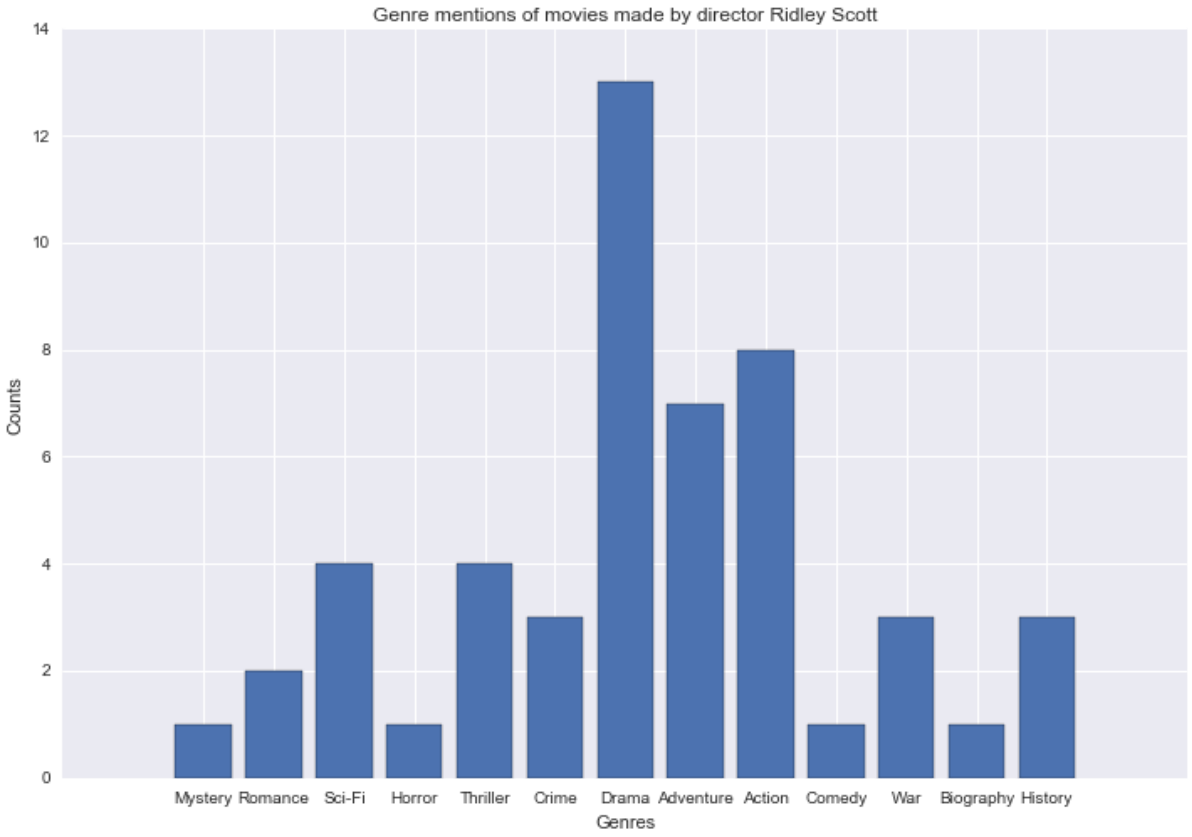
In [7]:
```
top = ["Steven Spielberg", "Woody Allen", "Clint Eastwood", "Martin Scor
sese", "Ridley Scott", "Steven Soderbergh", "Spike Lee", "Tim Burton",
"Renny Harlin", "Oliver Stone"]
for d in top:
    plot_director(d)
```

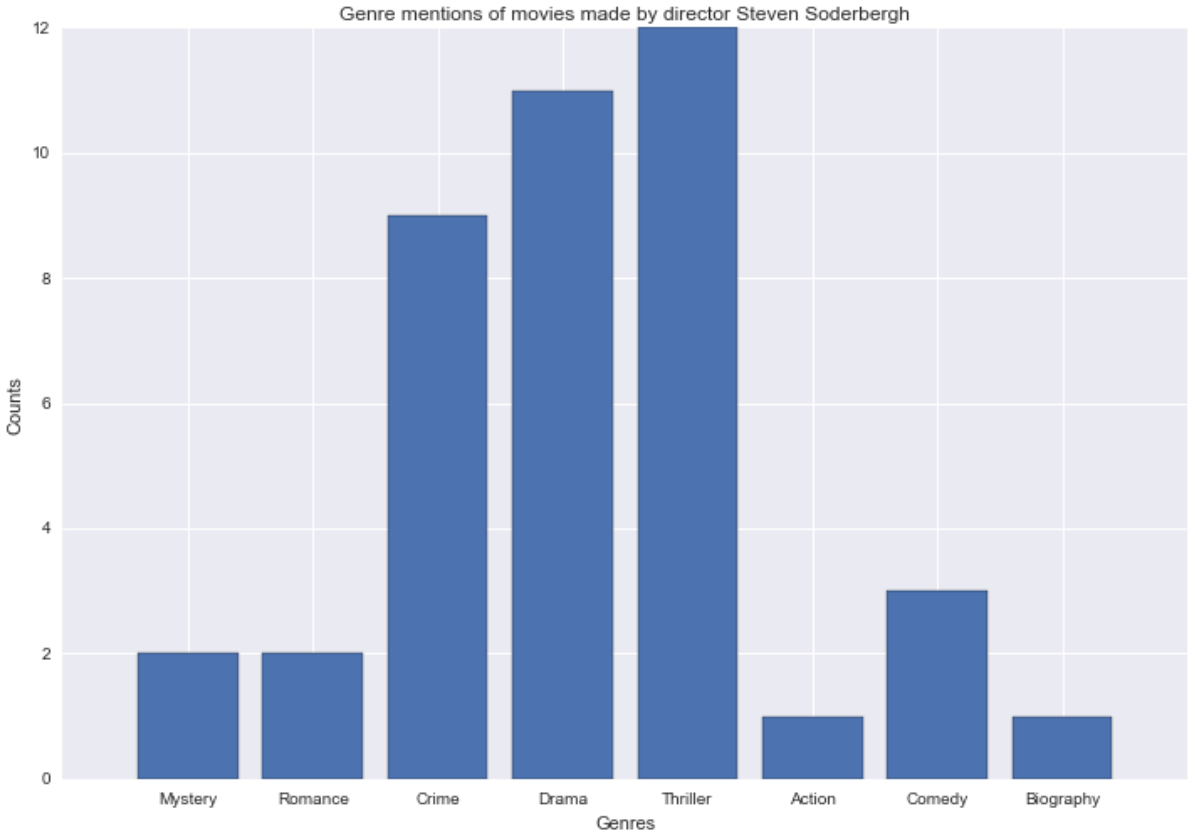

Genre mentions of movies made by director Steven Spielberg
Counts
0
2
4
6
8
10
12
14
Mystery
Sci-Fi
Family
Biography
Crime
Drama
Fantasy
Adventure
Action
Comedy
War
Thriller
History
Genres

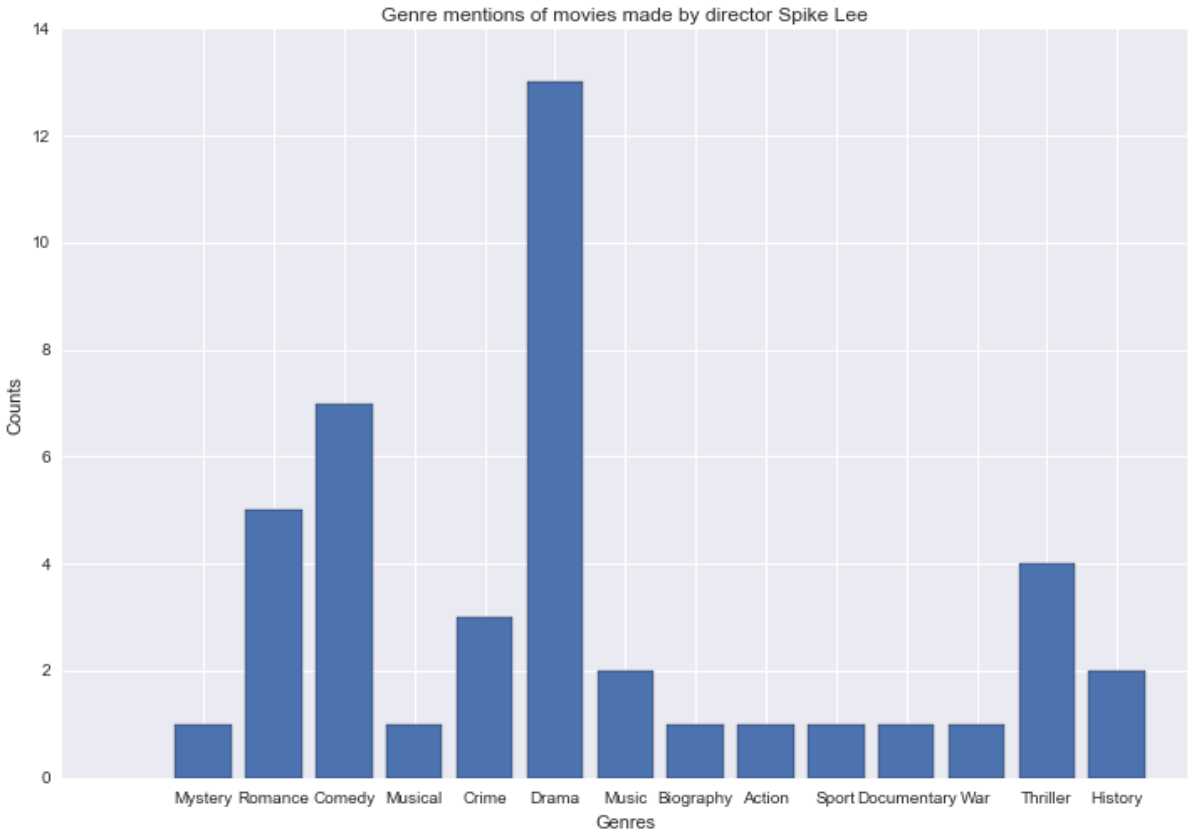

Genre mentions of movies made by director Woody Allen
Counts
0
5
10
15
20
Mystery
Drama
Sci-Fi
Crime
Romance
Fantasy
Comedy
Musical
Thriller
Genres

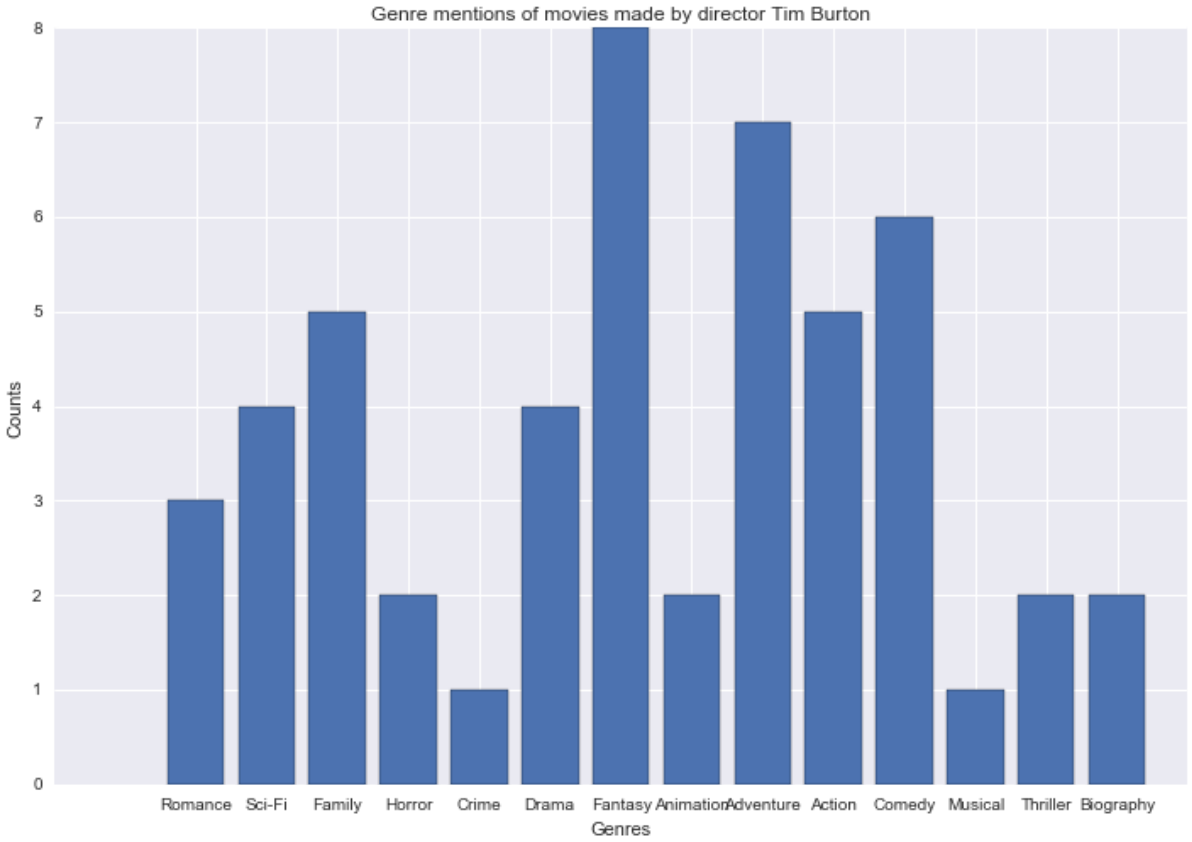

Genre mentions of movies made by director Clint Eastwood
Counts
0
2
4
6
8
10
12
14
16
Mystery
Romance
Musical
Thriller
Crime
Drama
Fantasy
Western
Music
Adventure
Action
Sport
War
Biography
History
Genres

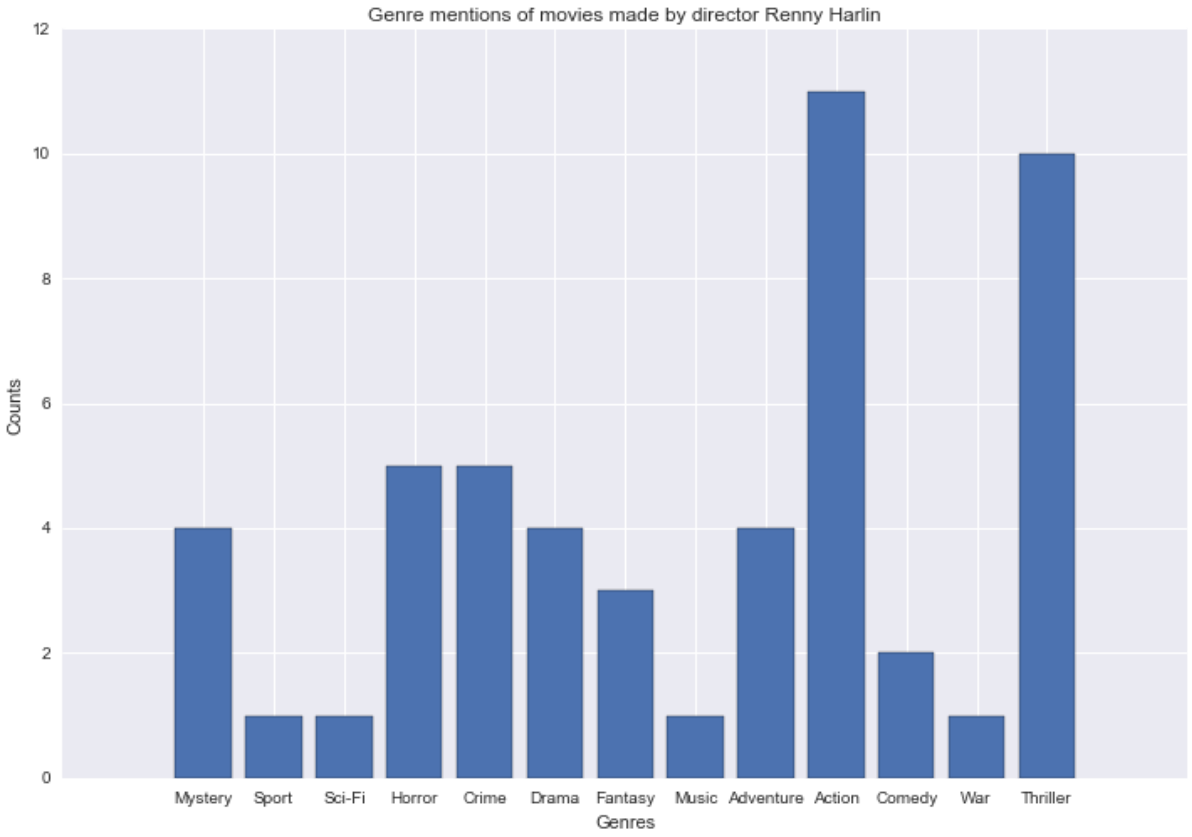

Genre mentions of movies made by director Martin Scorsese
Counts
0
2
4
6
8
10
12
14
16
Mystery
Romance
Sport
Family
Musical
Thriller
Crime
Drama
Music
Adventure
Comedy
Documentary
War
Biography
History
Genres

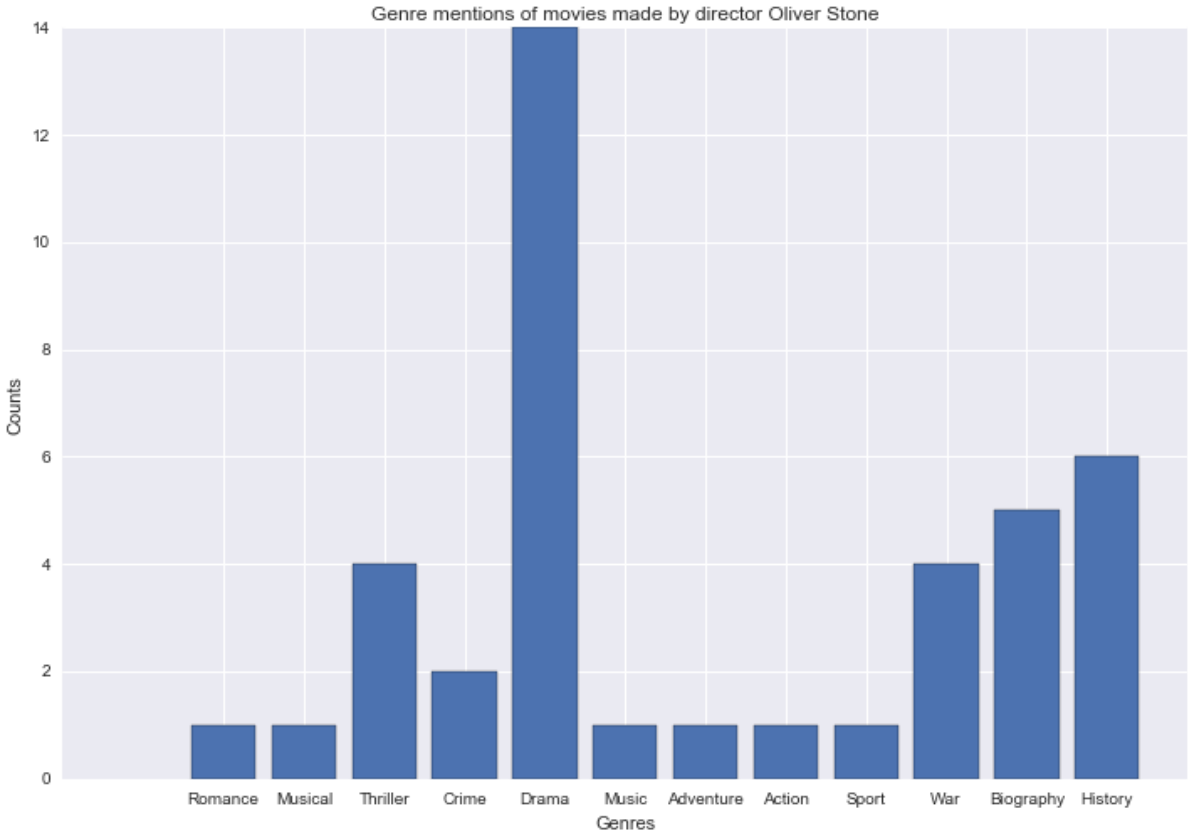

Genre mentions of movies made by director Ridley Scott
Counts
Genres
Mystery
Romance
Sci-Fi
Horror
Thriller
Crime
Drama
Adventure
Action
Comedy
War
Biography
History
0
2
4
6
8
10
12
14


Genre mentions of movies made by director Steven Soderbergh
Counts
Genres
Mystery
Romance
Crime
Drama
Thriller
Action
Comedy
Biography
0
2
4
6
8
10
12


Genre mentions of movies made by director Spike Lee
Counts
0
2
4
6
8
10
12
14
Mystery
Romance
Comedy
Musical
Crime
Drama
Music
Biography
Action
Sport
Documentary
War
Thriller
History
Genres


Genre mentions of movies made by director Tim Burton
Counts
0
1
2
3
4
5
6
7
8
Romance
Sci-Fi
Family
Horror
Crime
Drama
Fantasy
Animation
Adventure
Action
Comedy
Musical
Thriller
Biography
Genres


Genre mentions of movies made by director Renny Harlin
Counts
0
2
4
6
8
10
12
Mystery
Sport
Sci-Fi
Horror
Crime
Drama
Fantasy
Music
Adventure
Action
Comedy
War
Thriller
Genres


Genre mentions of movies made by director Oliver Stone
Counts
0
2
4
6
8
10
12
14
Romance
Musical
Thriller
Crime
Drama
Music
Adventure
Action
Sport
War
Biography
History
Genres

Conclusions:

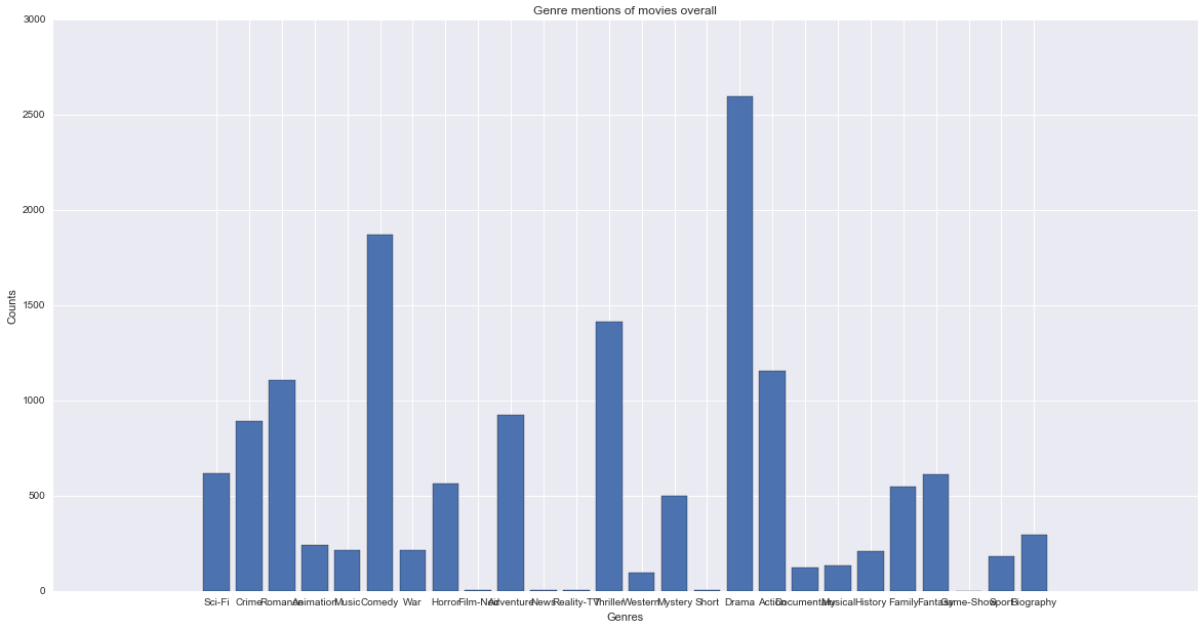
We see here that there are very obviously some genres that specific directors have frequently dabbled in. For example, Woody Allen most frequently directs movies that fall under comedy and/or romance genres, while Clint Eastwood has directed numerous more movies in that are dramas than any other category. Renny Harlin, meanwhile, directs many action and thriller movies.

It may be easy to look at that, but we also must contextualize this information using the knowledge we know about the genres overall. If most directors happen to direct a certain specific genre, then it might look like a calling card to help us identify the genre if we find out the director. But what if it turns out that the entire industry just so happens to really like directing that genre? Then the information's value is likely not as important.

In [15]:
```
# curr_df = df[df["director_name"] == name]
genres = df["genres"]
genres_new = [item for sublist in list(genres) for item in sublist]
sorted(Counter(genres_new).items(), key = operator.itemgetter(1))[::-1]
```

Out[15]:
```
[('Drama', 2594),
 ('Comedy', 1872),
 ('Thriller', 1411),
 ('Action', 1153),
 ('Romance', 1107),
 ('Adventure', 923),
 ('Crime', 889),
 ('Sci-Fi', 616),
 ('Fantasy', 610),
 ('Horror', 565),
 ('Family', 546),
 ('Mystery', 500),
 ('Biography', 293),
 ('Animation', 242),
 ('Music', 214),
 ('War', 213),
 ('History', 207),
 ('Sport', 182),
 ('Musical', 132),
 ('Documentary', 121),
 ('Western', 97),
 ('Film-Noir', 6),
 ('Short', 5),
 ('News', 3),
 ('Reality-TV', 2),
 ('Game-Show', 1)]
```

In [32]:
```
dictionary = plt.figure(figsize=(20, 10))
# d = sorted(Counter(genres_new).items(), key = operator.itemgetter(1))
[::-1]
d = Counter(genres_new)
plt.bar(range(len(d)), d.values(), align = "center")
plt.xticks(range(len(d)), d.keys())
plt.tick_params(axis="x", width = 20.0)
plt.xlabel("Genres")
plt.ylabel("Counts")
plt.title("Genre mentions of movies overall")
plt.show()
#d
```



As we discussed above, this is somewhat true in that the most common films are dramas, while comedy comes in second. However, when dramas/comedies are clearly the most common type of film a director has worked on, it can still be very useful information. Additionally, if a director directs a lot of movies in other categories, then those can also be potentially useful factors for us to consider.