# INF 510 Fall 2019 Final Project

1. **The names of team member(s)**

   Qianyao Wu

2. **How to run your code (what command-line switches they are, what happens when you invoke the code, etc.)**

   **The "data_grabing.py" works for grabing data, it requires:**

   - sqlite3, requests, csv, time, argparse, sys and BeautifulSoup

   The command-line '-source=remote' does 4 parts of work: grabing all data from online resources, data cleaning, exporting to csv files, modeling and storing it to database.

   The command-line '-source=test' does the same thing with the remote one. The only differnce is this grab a part of the data.

   The command-line '-source=local' processes data from csv files, models and stores it to database.

   (The "geocoding.csv" is the fourth resource for data visualization that cannot get from online resourses, so please keep it under any circumstance.)

   **Five files works for data analysis and visualization: "award_produced.py", "genre.py", "produced_by_country.py", "rate.py" and "top_rated_by_country.py". They require:**

   - sqlite3, pandas, plotly, widgets, folium and math

   (The "custom.geo.json" is working for data visualization, so please keep it under any circumstance.)

   To run this project, please install all packages above("environment.yml" is submitted), and then clone the repo at https://github.com/Qianyao818/inf510_project (https://github.com/Qianyao818/inf510_project) and execute this notebook

3. **Any major “gotchas” to the code (i.e. things that don’t work, go slowly, could be improved, etc.)**

   Because of rules of API providers, grabing all data from online resources takes about 25 minutes.

   If the data is from '-source=test', the visualization of 'Number of Produced Movies' not works well because the data from 1994 to 2017 are missing. Other plots works well. But all generated results are different from the findings I worte since the data is differnt.

4. **Anything else you feel is relevant to the grading of your project.**

   The code and data is updated after milestone 2. If using the data from milestone 2 to run this notebook, it will not work.

   I tried to run this notebook in another computer after installing same packages I used. But some plots just could not be shown with no error information. I have no idea what is going on. So I print the result of my notebook to a pdf called 'result.pdf' and upload to github. Please contact me if more information is needed.

5. **What did you set out to study? (i.e. what was the point of your project? This should be close to your Milestone 1 assignment, but if you switched gears or changed things, note it here.)**

   The point of this project is to find some interesting relationships among top-rated movies, Oscars award movies and production countries. I am curious about things like which country produces most movies each year, does all Oscars award movies get high rates, is there countries that produced few movies but most of them are fabulous and etc.

6. **What did you Discover/what were your conclusions (i.e. what were your findings? Were your original assumptions confirmed, etc.?)**

   From 1988 to 2017, India is the Top 1 movie produced country who produced more than 24.5k movies. USA is the No.2 with about 12k movies and China is the No.3 with over 700 movies.

   The 250 top-rated movies are from 24 countries. 91 are produced by USA, 28 are from Japan, 25 are UK produced and 17 are from France. This result shows that high number of produced movies does not mean this country's movies are fabulous.

   For the Oscars winners and nominees, high number of produced movies does not show high production ability of this country. USA won most 'Best Picture' award. France and Italy won most 'Foreign Language Film' award. And Ocars winners and nominees are not as popular as expected. For the 'Best Picture', the distribution of the rates are clustering around 7.4. For the 'Foreign Language Film', the overall performance is worse than the 'Best Picture' though its cluster appears around 7.4.

   The last part is about the genres of top-rated movies and the Oscars movies. The majority of top-rated and the Oscars movies are of Drama. Action, Adventure, Animation, Comedy and Crime are common type of these movies. None of them are Horror and TV movies.

7. **What difficulties did you have in completing the project?**

   The data modeling part is quite hard. One reason is that data comes from different resources so that the name of country is different, including common known names, abbreviation and official names. It takes a while to connect them. The other reason is there are repeated information so that how to build a effective database is important. And it does need more thoughts.

The other one comes up when visualizing the data. As the visualization is brand new for me, many attampts are done to realize what I want. Most of time was spent on understanding documents of packages.

8. **What skills did you wish you had while you were doing the project?**

   Skilled in using SQL and pandas.

9. **What would you do “next” to expand or augment the project?**

   I would like to finding more factors to figure out what contributes to a high-rated movie or a Oscar award movie.

In [1]:

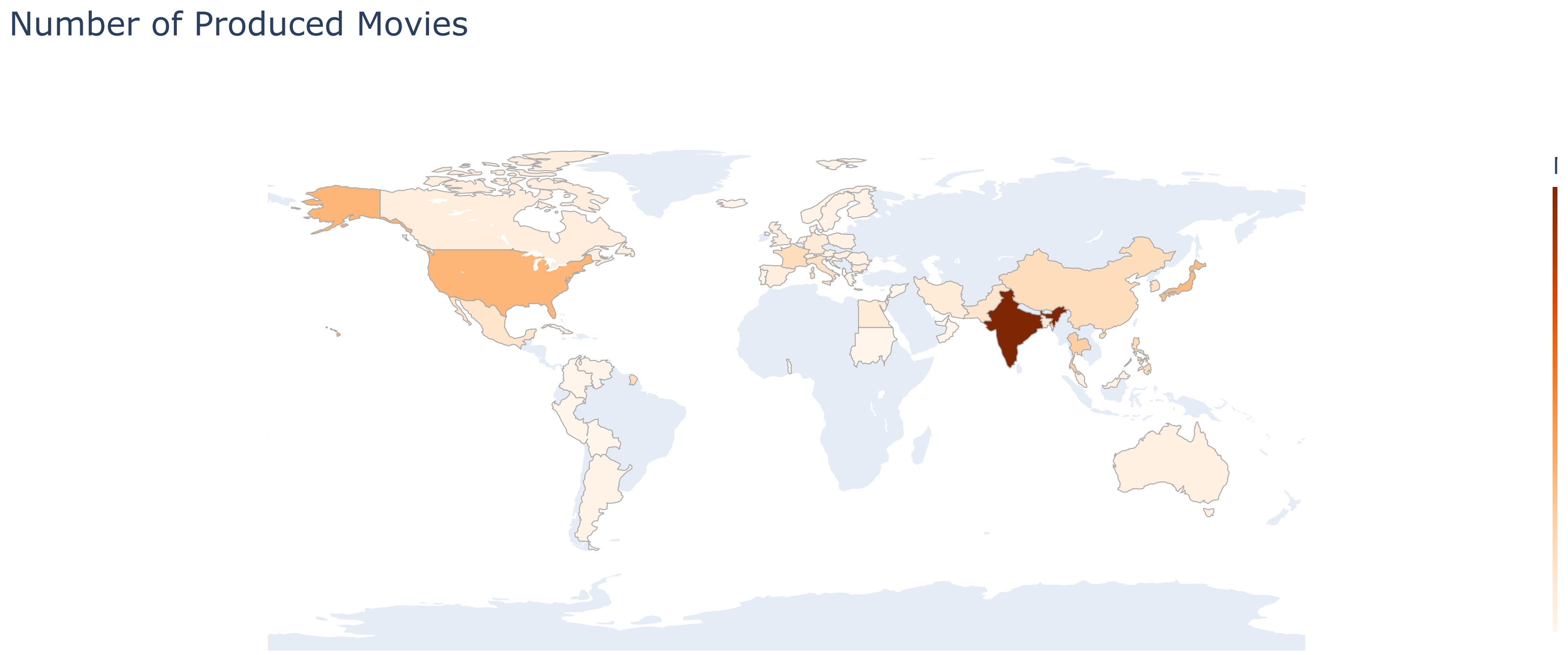
```
from src import top_rated_by_country as t
from src import produced_by_country as p
from src import award_produced as a
from src import genre as g
from src import rate as r
```

First, I would like to know how many movies each country produced from 1988 to 2017. So the following function draw a choropleth map about prodcued movies of each country. If the check box is checked, the map shows the total number of produced movies from 1988 to 2017. The slider is used to show map about number of produced movie each year.

In [2]:

```
#Draw a choropleth map about prodcued movies of each country.
produced_movie_by_country = p.produced_by_country()
produced_movie_by_country
```

□ 1989 - 2017 Year: 1989



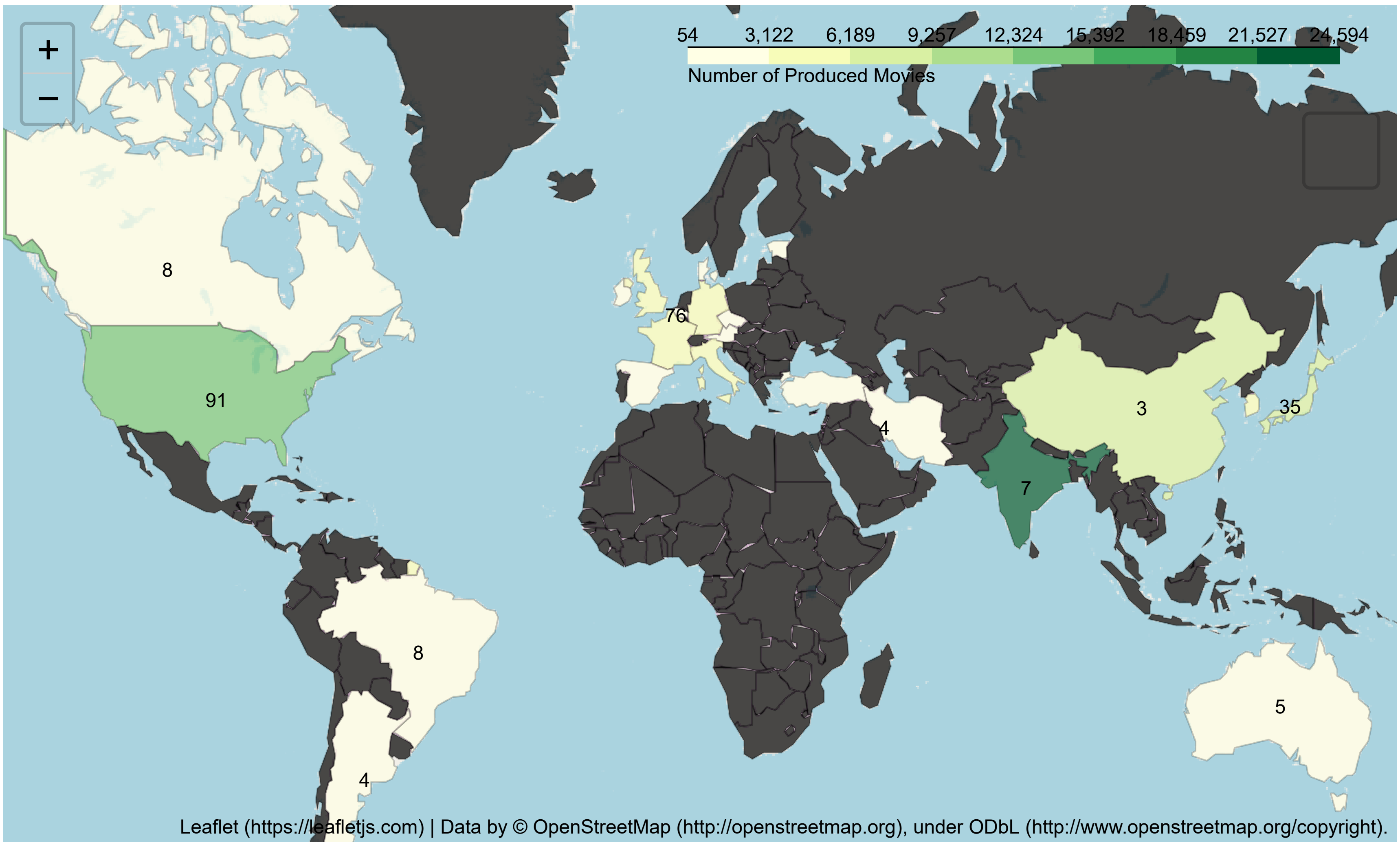
This map shows that from 1988 to 2017, India is the Top 1 movie produced country who produced more than 24.5k movies. USA is the No.2 with about 12k movies and China is the No.3 with over 700 movies. And India, USA and China are always the most productive countries through this long period of time.

The second thing I would like to know is is there any relationship bewteen top-rated movies and countries that produced a great amount of movies. I use a map and a scatter plot to visualize this information.

In [3]:

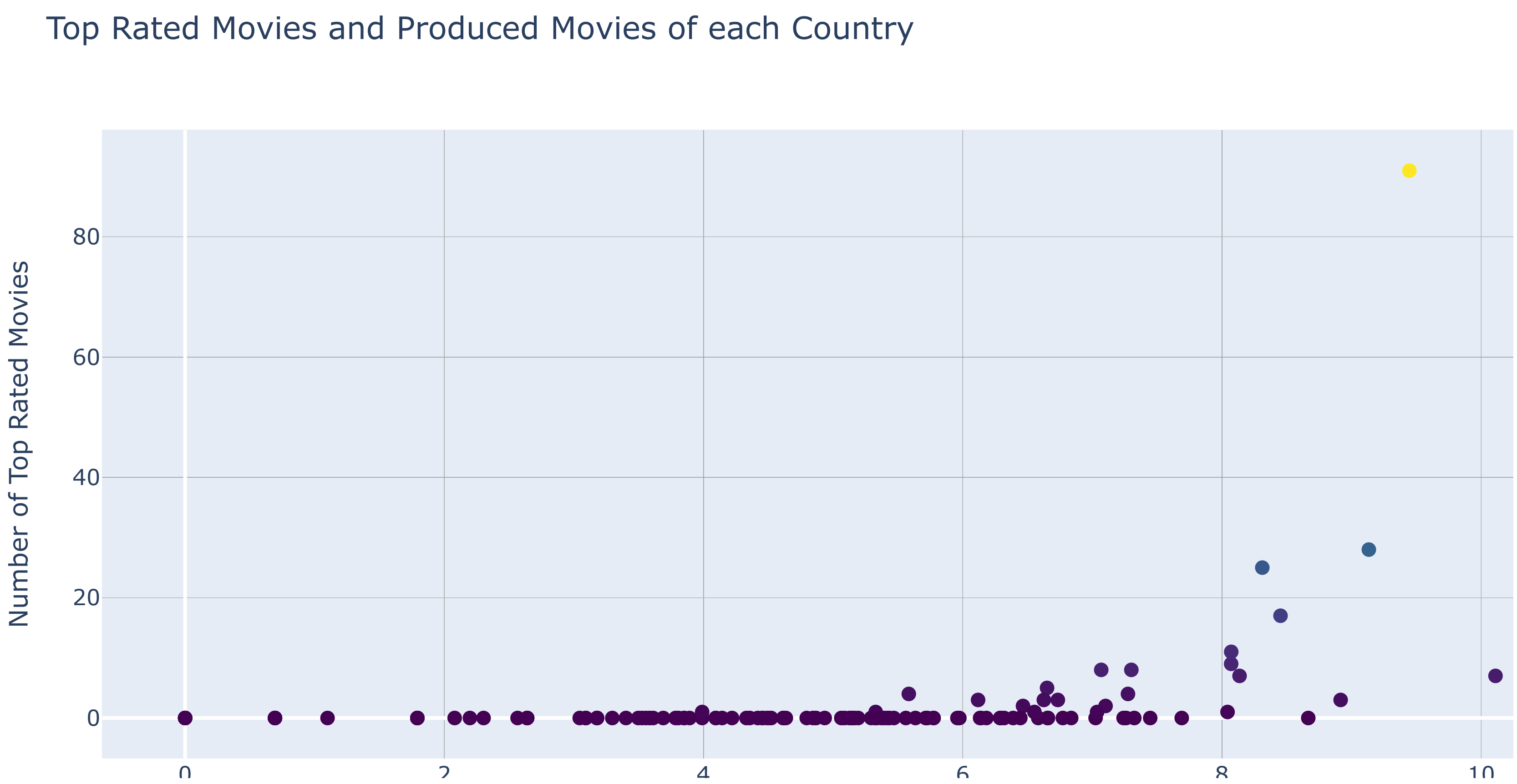
```
# Draw a choropleth map shows total number of produced movies and the distribution of top-rated movies.
# With the scale of the map changes, detailed information is shown.
top_country_map = t.top_rated_by_country_map()
top_country_map
```

Out[3]:


+
−
54
3,122
6,189
9,257
12,324
15,392
18,459
21,527
24,594
Number of Produced Movies
8
91
76
4
3
35
7
8
5
4
Leaflet (https://leafletjs.com) | Data by © OpenStreetMap (http://openstreetmap.org), under ODbL (http://www.openstreetmap.org/copyright).

In [4]:

```
# Generate a scatter plot to see the relationship between total number of produced movies and number of top-rated movies.
# Log function is applied to better visualize this information.
top_country_scatter = t.top_rated_by_country_scatter()
top_country_scatter
```

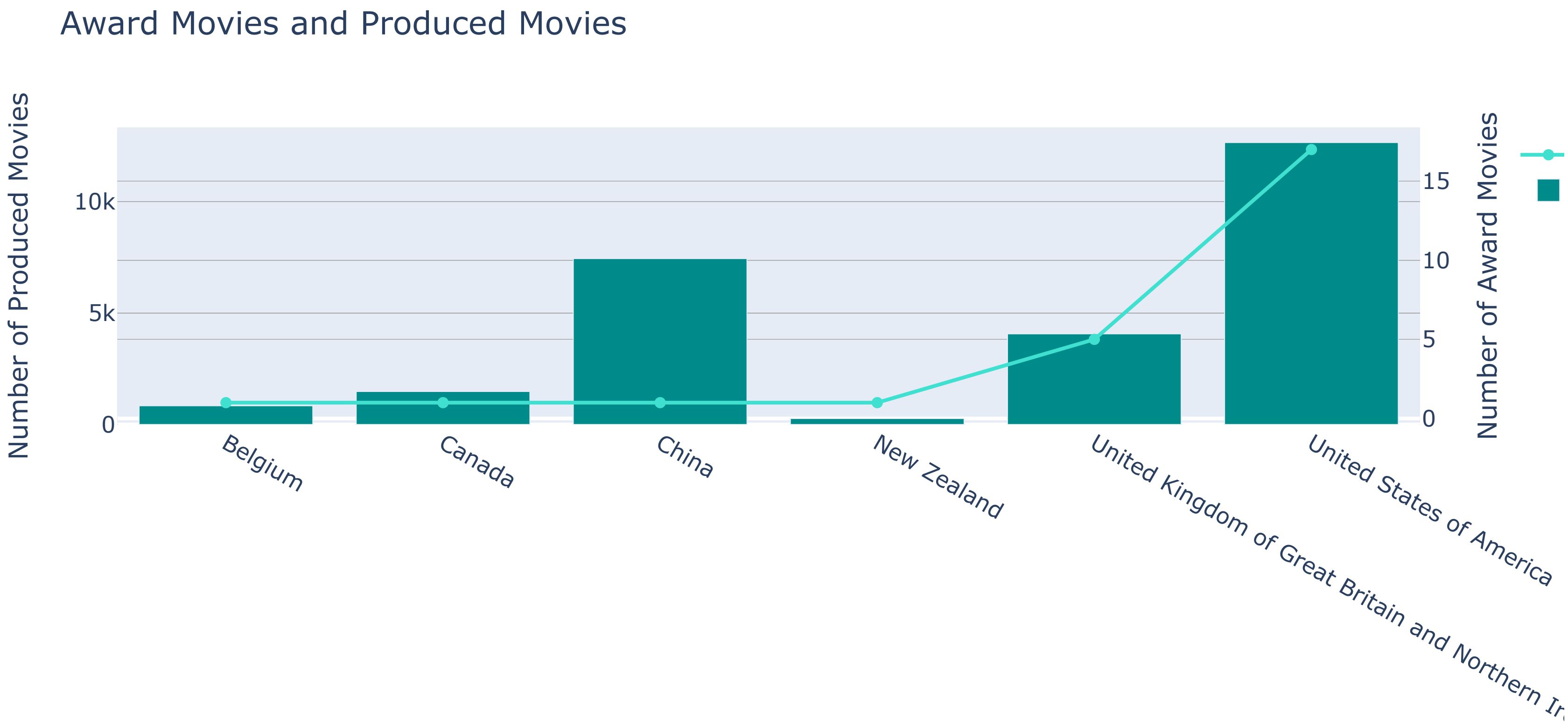


The 250 top-rated movies are from 24 countries. 91 are produced by USA, 28 are from Japan, 25 are UK produced and 17 are from France. This result shows that high number of produced movies does not mean this country's movies are fabulous.

After that, I would like to know something about the Oscars winners and nominees. The first is which country has more Oscars winners and nominees and it there a relationship with number of produced movies of this country. The bar and line chart is used to visualize.

In [5]:
```
# Generate a bar chart and a line chart to show the relationship between the Oscars movies and countries.
# Four types of the Oscars movies are shown by the combination of drop-down menus.
award_produced = a.award_produced()
award_produced
```

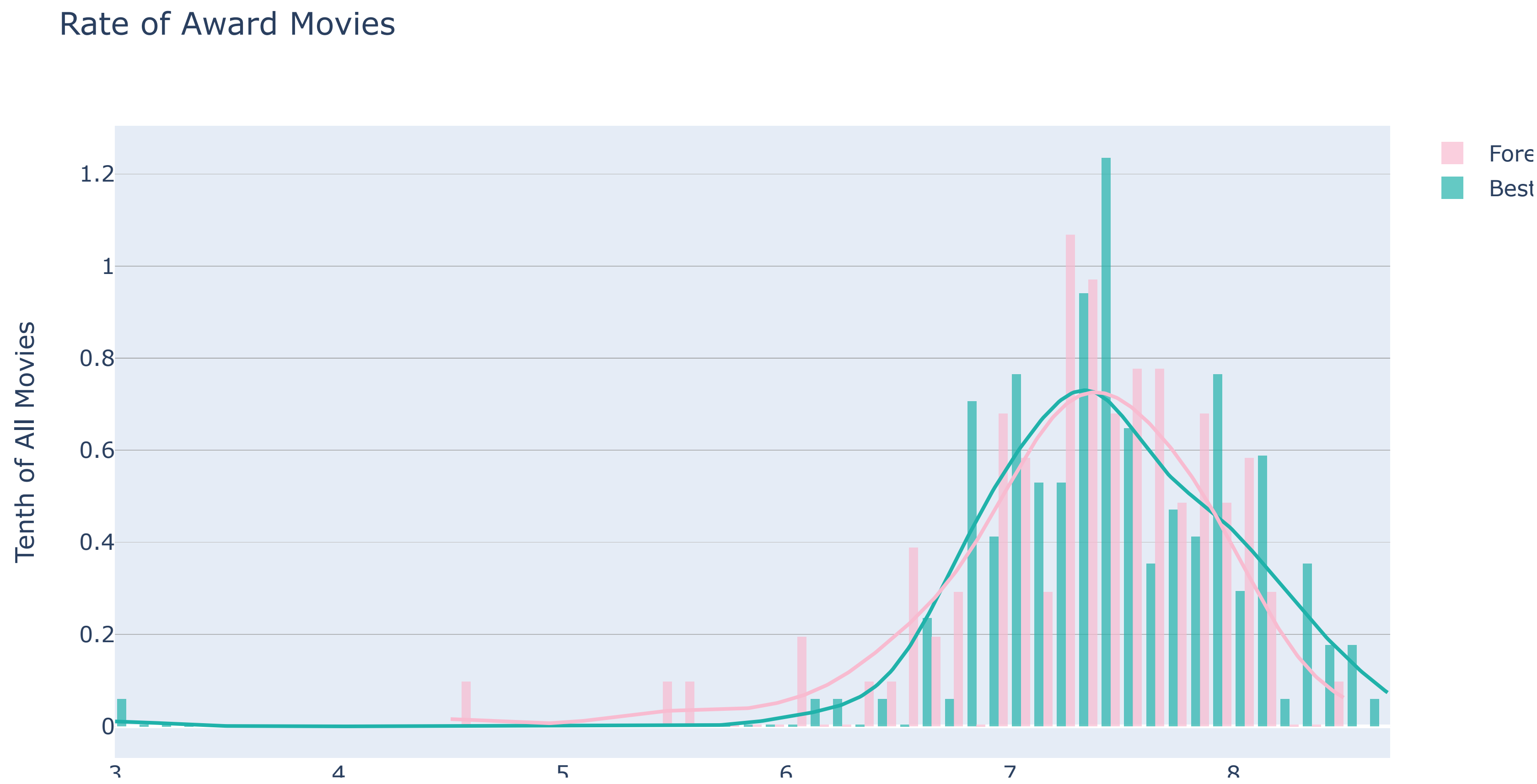
Award name: Best Picture  Award type: winner



For the Oscars winners and nominees, high number of produced movies does not show high production ability of this country. USA won most 'Best Picture' award. France and Italy won most 'Foreign Language Film' award.

I am also curious about are all the Oscars movies have high rates. So I use a distplot to show the distribution of 'Foreign Language Film' and 'Best Picture'.

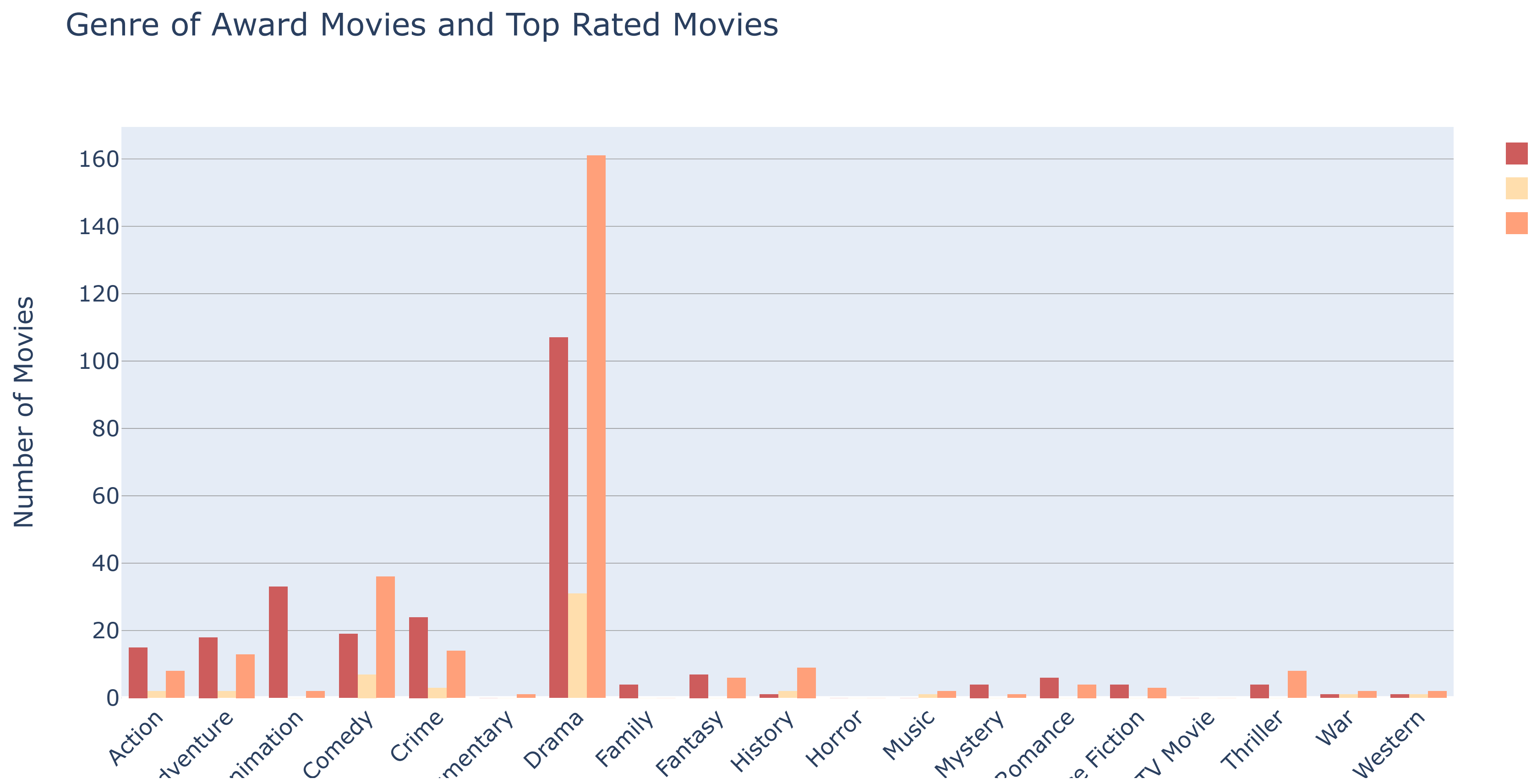
In [6]:
```
# Generate a distplot to show the distribution of rates in 'Foreign Language Film' and 'Best Picture'.
rate = r.rate()
rate
```



As shown in the plot, Ocars winners and nominees are not as popular as expected. For the 'Best Picture', the distribution of the rates are clustering around 7.4. For the 'Foreign Language Film', the overall performance is worse than the 'Best Picture' though its cluster appears around 7.4.

The genres of top-rated movies and the Oscars movies are also interesting. The genre type reflects the taste of audiences. Histogram is used to show this information.

In [7]:
```
# Generate a histogram plot to show the genre type of top-rated movies, Oscars winners and Oscars nominees.
genre = g.genre()
genre
```



The plot shows that the majority of top-rated and the Oscars movies are of Drama. Action, Adventure, Animation, Comedy and Crime are common type of these movies. None of them are Horror and TV movies.

In [ ]: