

IMDB Top 250 Rating Films Analysis

Extract Data

Python is used to extra the data of [IMDB Top 250 rating films](#). Using Chrome, it is possible to see the source code of this page. I used Urllib and Beautiful Soup to fetch the data.













IMDb Charts

Top 250

As voted by regular IMDb users

Showing 250 Titles

Sort by: IMDb Rating  

Rank & Title		IMDb Rating	Your Rating	
	1. The Shawshank Redemption (1994)	★ 9.2	☆	
	2. The Godfather (1972)	★ 9.2	☆	
	3. The Godfather: Part II (1974)	★ 9.0	☆	
	4. The Dark Knight (2008)	★ 8.9	☆	
	5. Pulp Fiction (1994)	★ 8.9	☆	
	6. Schindler's List (1993)	★ 8.9	☆	

Extract Data and URL from IMDB Charts

With Urllib, we can get the full html code with the following python code. `Page.read()` returns the full html code exactly as what we get by using “view source code” of chrome.

```
import urllib
```

```
url = "http://www.imdb.com/chart/top?ref_=nv_ch_250_4"
```

```
page = urllib.request.urlopen(url)
```

BeautifulSoup can be used to parse the HTML code we fetched from the website. Below is the HTML code for The Godfather. We can see that film names are stored in tag <a> in tag <td> with class attributes "titleColumn".

```
<td class="titleColumn">
  <span name="ir" data-value="9.171">2.</span>
  <a href="/title/tt0068646/?ref_=chttp_tt_2"
title="Francis Ford Coppola (dir.), Marlon Brando, Al Pacino" >The Godfather</a>
  <span name="rd" data-value="1972-03-24" class="secondaryInfo">(1972)</span>
</td>
<td class="ratingColumn imdbRating">
  <strong name="nv" data-value="994809" title="9.2 based on 994,809 votes">9.2</strong>
</td>
```

We can get the title of movies with the following code. Link.contents returns the film title. The url of godfather can be get by calling link["href"]

```
from bs4 import BeautifulSoup

movies = soup.find_all(class_ = "titleColumn")

for movie in movies:

    for link in movie.find_all("a",href = True):

        name = link.contents

        url = link["href"]
```

In similar ways, we can get the releasing date of these movies. And to get further information about total number of ratings, reviews and critics, we need to access the webpage of each movie.

Extract Rating and Genre for each Movie

With url we fetched from previous step, it is possible to get number of ratings from movie page. The source code of Godfather is shown below. We can see that total number of rating is storage in tag span with itemprop = "ratingCount".

```
<div class="star-box-details" itemtype="http://schema.org/AggregateRating" itemscope itemprop="aggregateRating">
  Ratings:
  <strong><span itemprop="ratingValue">9.2</span></strong><span class="mellow"><span itemprop="bestRating">10</span></span> from <a
href="ratings?ref_tt_ov_rt"
title="994,819 IMDb users have given a weighted average vote of 9.2/10" > <span itemprop="ratingCount">994,819</span> users
</a>&nbsp;</div>
```

So we can get the total number of rating by the following python code. In similar ways we can get genre information and number of review and critics.

```
url1 = line.rstrip()

page = urllib.request.urlopen(base + url1)

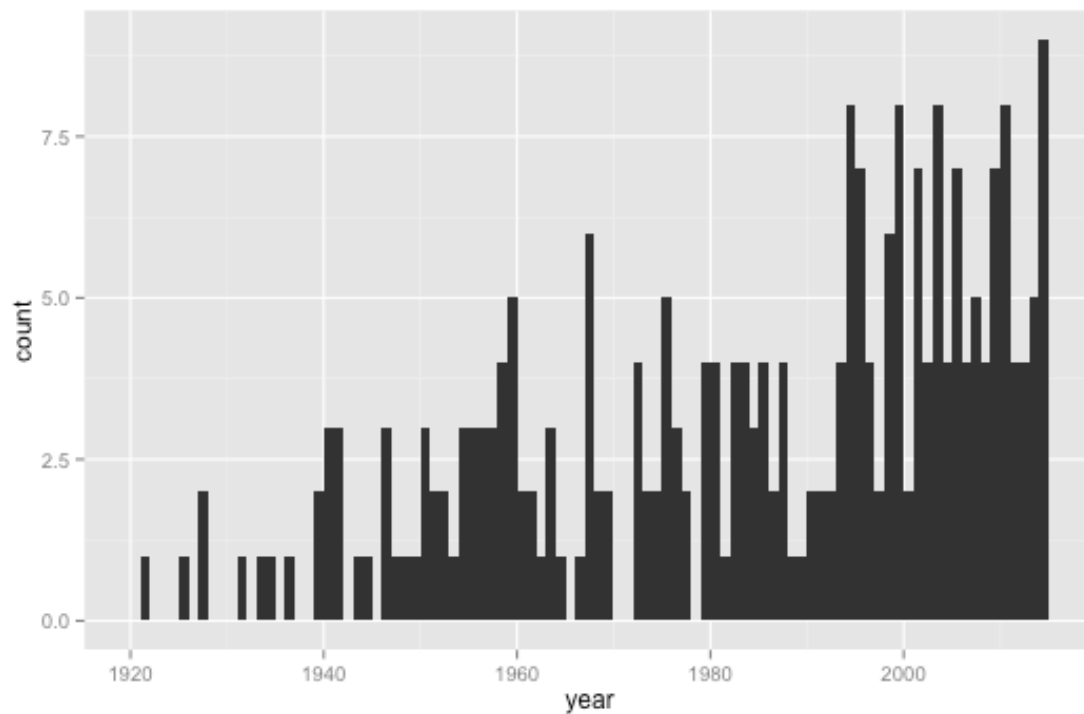
soup = BeautifulSoup(page.read())

nRate = soup.find_all("span", {"itemprop" : "ratingCount"})
```

Data Visualization and Analysis

Histogram of Year

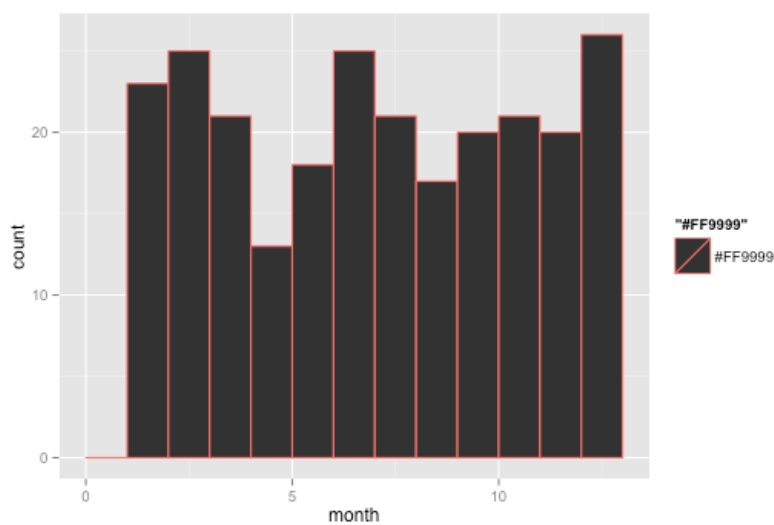
The histogram of total number of films by years is shown below. We can see that there tends to be fewer high rating movies that dated before 1940. Moreover, there seems to be two peaks, one is around 1960s and another is around 2005.



Histogram of Month

The histogram of total number of top rating movies by month is shown in the graph below.

We can see that the total film counts by month can be considered as following a uniform distribution. Which means the releasing month of film does not have significant effect on the rating of movies.



Correlation Matrix

The correlation matrix of rating, total number of rating, total number of reviews, total number of critics and the releasing year is shown in the table below. From that we can see ratings are positively correlated with number of rating, review and critics but negatively related with releasing year.

	Rating	Total Rating	Total Review	Total Critic	Year
Rating	1	0.627	0.5358	0.1303	-0.0135
Total Rating	0.627	1	0.867	0.5534	0.459
Total Review	0.5358	0.867	1	0.5347	0.3544
Total Critic	0.1303	0.5534	0.5347	1	0.5012
Year	-0.0135	0.459	0.3544	0.5012	1

Bar plot of Genre

The bar plot of total number of movies by genre is shown in the bar plot below. We can see that the genre with highest number of high rating movies is drama. The genre with fewest high rating movies is musical.

