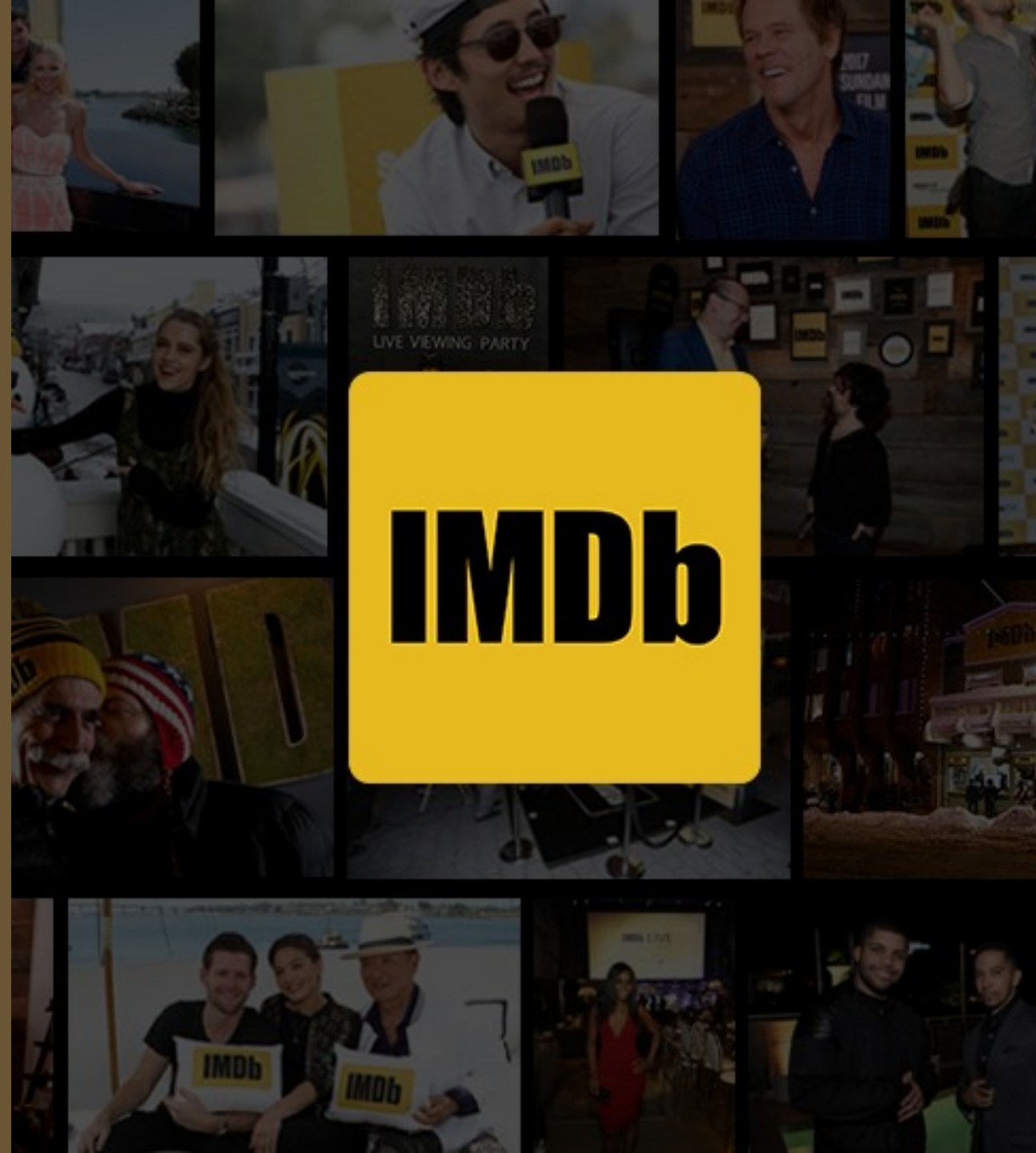
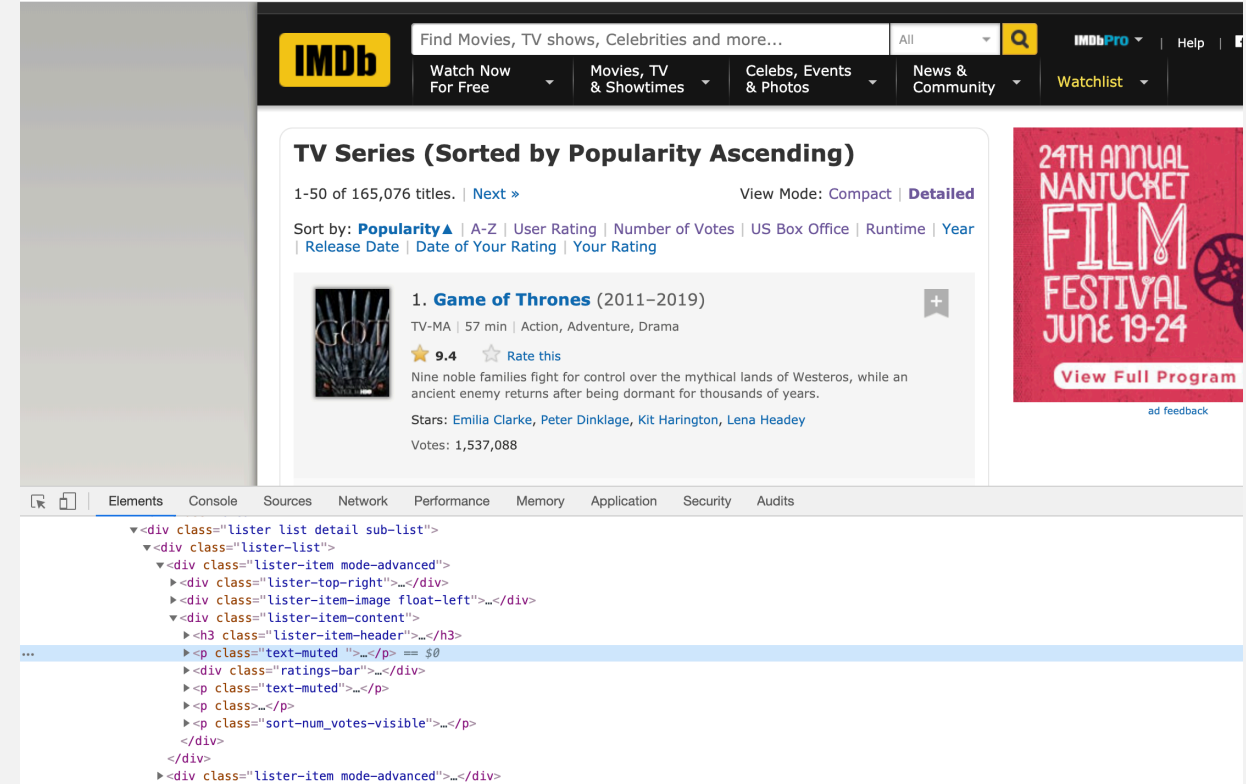


PREDICTING IMDB MOVIE GROSS INCOME USING MULTIPLE LINEAR REGRESSION

Presented by Shu Jiang



WEB SCRAPING



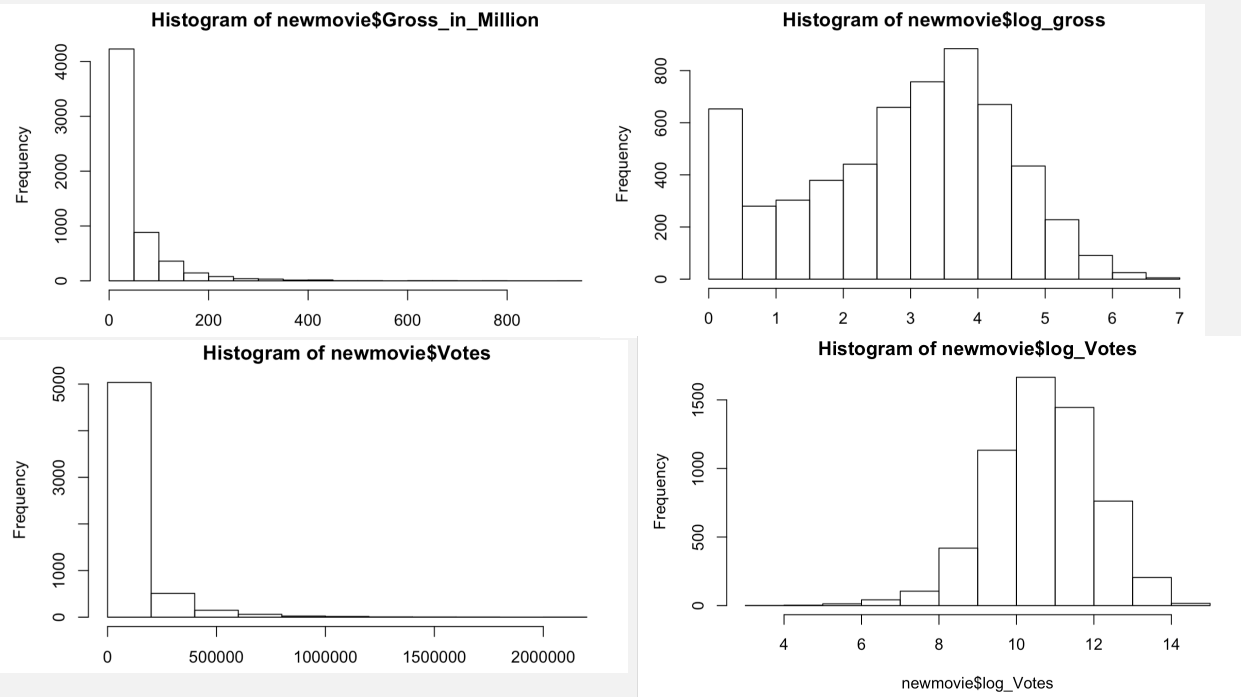
The screenshot shows the IMDb website's 'TV Series (Sorted by Popularity Ascending)' page. The top navigation bar includes the IMDb logo, a search bar, and links for 'Watch Now For Free', 'Movies, TV & Showtimes', 'Celebs, Events & Photos', 'News & Community', and a 'Watchlist'. The main content area displays a list of TV series, with 'Game of Thrones' (2011–2019) as the first entry. It shows a TV-MA rating, 57 min runtime, and genres: Action, Adventure, Drama. The series has a 9.4 star rating and a brief description. Below the description, it lists stars (Emilia Clarke, Peter Dinklage, Kit Harington, Lena Headey) and the number of votes (1,537,088). On the right side, there is a red banner for the '24TH ANNUAL NANTUCKET FILM FESTIVAL' running from June 19-24, with a 'View Full Program' button. At the bottom, the browser's developer tools are open, showing the 'Elements' tab with the HTML structure of the TV series list item.

```
<div class="list-item mode-advanced">
  <div class="list-top-right"></div>
  <div class="list-item-image float-left"></div>
  <div class="list-item-content">
    <h3 class="list-item-header"></h3>
    <p class="text-muted"></p>
    <div class="ratings-bar"></div>
    <p class="text-muted"></p>
    <p class="sort-num_votes-visible"></p>
  </div>
</div>
<div class="list-item mode-advanced"></div>
```

	Name	Year	Level	Duration	Genre	Score	Director	Votes	Gross_in_Million
0	Aladdin	2019	PG	128 min	Adventure Comedy Family	7.4	Guy Ritchie	44991	142.697174
1	Godzilla: King of the Monsters	2019	PG-13	131 min	Action Adventure Fantasy	6.9	Michael Dougherty	21581	
2	Rocketman	2019	R	121 min	Biography Drama Music	7.7	Dexter Fletcher	13985	
3	Avengers: Endgame	2019	PG-13	181 min	Action Adventure Sci-Fi	8.8	Anthony Russo	430042	807.688784
4	John Wick: Chapter 3 - Parabellum	2019	R	131 min	Action Crime Thriller	8.0	Chad Stahelski	70336	114.652812
5	Booksmart	2019	R	102 min	Comedy	7.5	Olivia Wilde	13480	11.038184
6	Once Upon a Time ... in Hollywood	2019	NaN	159 min	Comedy Drama	9.8	Quentin Tarantino	5034	
7	Terminator: Dark Fate	2019	NaN		Action Adventure Sci-Fi		Tim Miller		
8	The Perfection	2018	TV-MA	90 min	Drama Horror Thriller	6.2	Richard Shepard	10362	
9	Brightburn	2019	R	90 min	Horror Sci-Fi	6.5	David Yarovesky	10610	11.897552
10	Rim of the World	2019	TV-14	98 min	Action Adventure Comedy	5.1	McG	8275	
11	Captain Marvel	2019	PG-13	123 min	Action Adventure Sci-Fi	7.1	Anna Boden	286653	426.011433
12	Ma	2019	R	99 min	Horror Thriller	6.1	Tate Taylor	2530	
13	Rambo: Last Blood	2019	NaN		Action Adventure Thriller		Adrian Grunberg		
14	The Goldfinch	2019	R		Drama		John Crowley		
15	Us	2019	R	116 min	Horror Mystery Thriller	7.2	Jordan Peele	96627	174.89178

- USING THE LIBRARY BEAUTIFULSOUP TO INSPECT WEBPAGE AND EXTRACT RELATED DATA CONCERNING EACH MOVIE

Log transformation: gross_in_million and votes

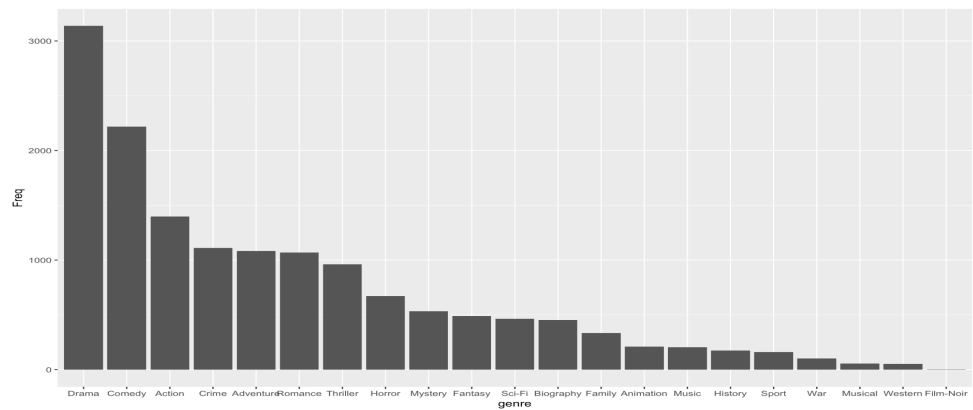


Change numerical variable
“runtime” and “Year” to categorical
variable

```
table(newmovie$final_Duration)
```

Long	Normal	Short
1222	2430	2157

Reclassify and sort the frequency of genre:
Top 3: Drama, Comedy and Action



```
> table(newmovie$new_Year)
```

before2000	first_ten	second_ten
2125	1865	1819

```
Call:
lm(formula = log_gross ~ new_Year + Level + final_Duration +
    Score + log_Votes + Animation + Comedy + Drama + Sport, data = test)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-3.4551 -0.4522  0.1284  0.5933  2.2949
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.74908    0.47222   -1.586   0.11332
new_Yearfirst_ten -0.14723    0.11624   -1.267   0.20590
new_Yearsecond_ten  0.14097    0.10660    1.322   0.18668
LevelPG-13     -0.49155    0.14848   -3.311   0.00100 **
LevelR        -1.08863    0.14020   -7.765 4.87e-14 ***
final_DurationNormal -0.52736    0.10121   -5.211 2.79e-07 ***
final_DurationShort -0.64907    0.12606   -5.149 3.81e-07 ***
Score         -0.32906    0.05800   -5.673 2.41e-08 ***
log_Votes      0.67671    0.04184   16.173 < 2e-16 ***
Animation      0.15732    0.25273    0.622  0.53391
Comedy         0.20744    0.09840    2.108  0.03553 *
Drama         -0.28625    0.09847   -2.907  0.00382 **
Sport         0.11147    0.25163    0.443  0.65797
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.9081 on 487 degrees of freedom
Multiple R-squared:  0.5629,    Adjusted R-squared:  0.5522
F-statistic: 52.27 on 12 and 487 DF,  p-value: < 2.2e-16
```

Linear model on
test data set
 R^2 squared: 0.5629

Linear model on
training data set
 R^2 squared: 0.6461

```
summary(mbic)

##
## Call:
## lm(formula = log_gross ~ new_Year + Level + final_Duration +
##     Score + log_Votes + Animation + Drama, data = datanml)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.46763 -0.44222  0.08769  0.50939  2.18998
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.79353    0.26020   -6.893 8.24e-12 ***
## new_Yearfirst_ten -0.23326    0.05762   -4.048 5.44e-05 ***
## new_Yearsecond_ten -0.18214    0.05506   -3.308 0.000964 ***
## LevelPG-13     -0.25615    0.07151   -3.582 0.000352 ***
## LevelR        -0.93855    0.06767  -13.870 < 2e-16 ***
## final_DurationNormal -0.29996    0.05091   -5.893 4.75e-09 ***
## final_DurationShort -0.41889    0.06264   -6.688 3.27e-11 ***
## Score         -0.30418    0.03136   -9.699 < 2e-16 ***
## log_Votes      0.74709    0.02192   34.079 < 2e-16 ***
```

56

```
## Animation      0.59917    0.13159    4.553 5.74e-06 ***
## Drama         -0.30303    0.04803   -6.309 3.76e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7763 on 1403 degrees of freedom
## Multiple R-squared:  0.6461, Adjusted R-squared:  0.6435
## F-statistic: 256.1 on 10 and 1403 DF,  p-value: < 2.2e-16
```

FINAL MODEL AND
CROSS VALIDATION

SHINYAPP DEPLOYMENT

movie gross prediction by multiple linear regression

Select log_Votes:

4.9 10 14.5

4.9 5.9 6.9 7.9 8.9 9.9 10.9 11.9 12.9 13.9 14.5

Select Movie Score:

1 6 9.8

1 1.9 2.8 3.7 4.6 5.5 6.4 7.3 8.2 9.1 9.8

Select Year:

☒ before2000

☐ 2000-2010

☐ after 2010

Select movie level:

☒ PG

☐ PG-13

☐ R

Select movie duration:

☒ >120 min

☐ 101-120 min

☐ <=100 min

Select movie genre: Animation

☒ include

☐ not include

Select movie genre: Comedy

☒ include

☐ not include

Select movie genre: Drama

☒ include

☐ not include

Select movie genre: Sport

☒ include

☐ not include

data

32071.05