

# Predict the IMDb score of films

Stat 418 project

Zhou, Yichen



# Attaining Data – Web Scraper



## 1. Avengers: Infinity War (2018)

PG-12 | 149 min | Action, Adventure, Sci-Fi

★ 8.5

☆ Rate this

68

Metascore

The Avengers and their allies must be willing to sacrifice all in an attempt to defeat the powerful Thanos before his blitz of devastation and ruin puts an end to the universe.

Directors: Anthony Russo, Joe Russo | Stars: Robert Downey Jr., Chris Hemsworth, Mark Ruffalo, Chris Evans

Votes: 647,633 Gross: \$678.82M



## 2. Aquaman (2018)

PG-12 | 143 min | Action, Adventure, Fantasy

★ 7.1

☆ Rate this

55

Metascore

Arthur Curry, the human-born heir to the underwater kingdom of Atlantis, goes on a quest to prevent a war between the worlds of ocean and land.

Director: James Wan | Stars: Jason Momoa, Amber Heard, Willem Dafoe, Patrick Wilson

Votes: 246,564 Gross: \$335.06M

\_root

film

Film\_name

Popularity\_rank

IMDbscore

Metascore

Votes

Certificate

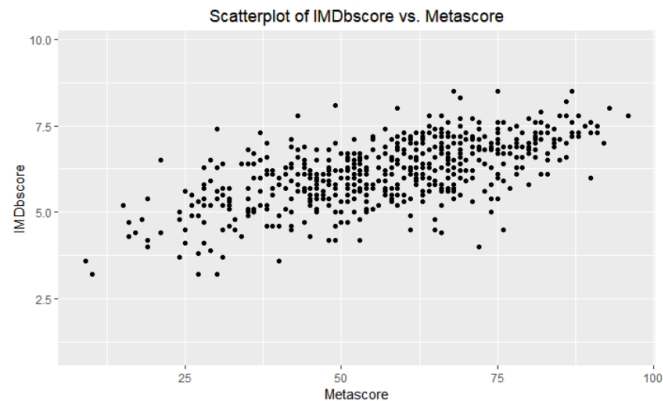
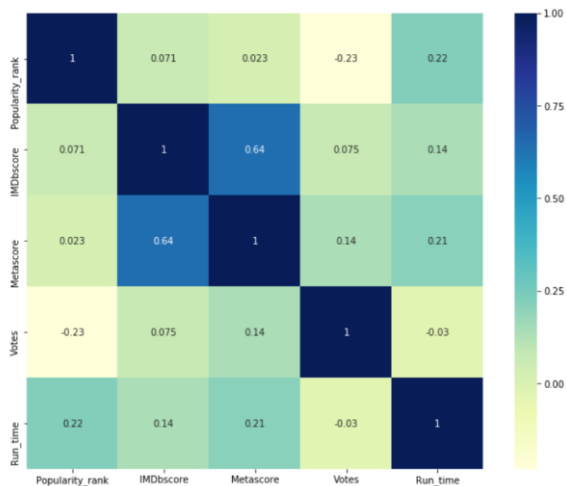
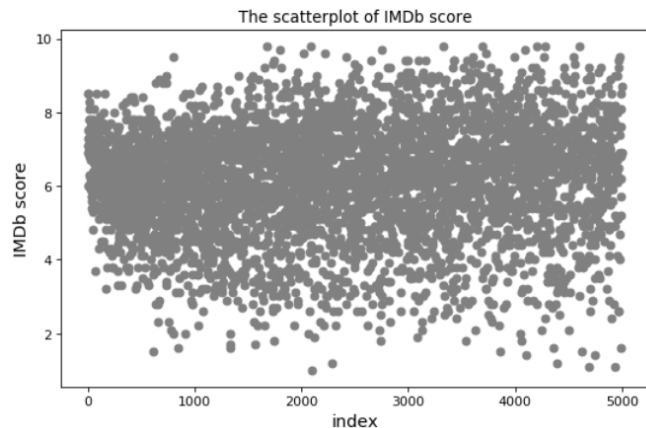
Run\_time

Genre

web-scraper-order	web-scraper-start-url	Film_name	Popularity_rank	IMDbscore	Metascore	Votes	Run_time	Genre
0	1557261010-11301 https://www.imdb.com/search/title?title_type=f...	Avengers: Infinity War	1	8.5	68.0	647469.0	149 min	Action, Adventure, Sci-Fi
1	1557261010-11302 https://www.imdb.com/search/title?title_type=f...	Aquaman	2	7.1	55.0	246474.0	143 min	Action, Adventure, Fantasy
2	1557261010-11303 https://www.imdb.com/search/title?title_type=f...	Arctic	3	6.9	71.0	13690.0	98 min	Adventure, Drama

# Exploratory data analysis

```
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 5000 entries, 0 to 4999  
Data columns (total 9 columns):  
web-scraper-order    5000 non-null object  
web-scraper-start-url 5000 non-null object  
Film_name            5000 non-null object  
Popularity_rank       5000 non-null int64  
IMDbscore            4459 non-null float64  
Metascore            563 non-null float64  
Votes               4459 non-null float64  
Run_time             4284 non-null object  
Genre                4989 non-null object  
dtypes: float64(3), int64(1), object(5)  
memory usage: 351.6+ KB
```



# Feature Engineering

## Genre

Action, Adventure, Sci-Fi

Action, Adventure, Fantasy

Adventure, Drama

Animation, Action, Adventure

Action, Adventure, Comedy

Action, Adventure, Sci-Fi

```
for (i in 1:length(data[, 'Genre'])) {  
  a <- as.vector.factor(data[i, 'Genre'])  
  b <- unlist(strsplit(a, split=","))  
  data[i, 'Genre'] <- b[1]  
}
```

## Genre

Action

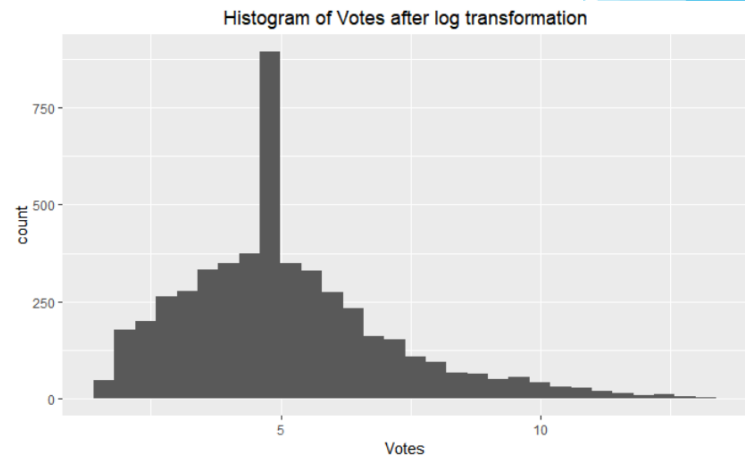
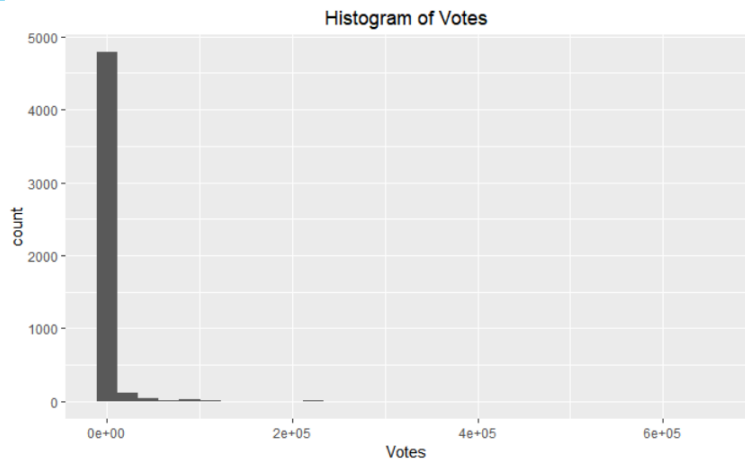
Action

Adventure

Animation

Action

Action



# Model – Linear Regression, XGBoost

## Linear Regression

```
Call:
lm(formula = IMDBscore ~ Popularity_rank + Metascore + log(Votes) +
    Run_time + Genre, data = data)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-5.8620 -0.6727  0.0934  0.7895  4.0798
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.366e+00  2.230e-01  15.092 < 2e-16 ***
Popularity_rank  3.750e-05  1.662e-05   2.256  0.024116 *
Metascore      2.762e-02  3.212e-03   8.601 < 2e-16 ***
log(Votes)     -5.127e-02  1.248e-02  -4.108  4.06e-05 ***
Run_time       1.067e-02  8.888e-04  12.000 < 2e-16 ***
GenreAdventure  2.290e-01  9.411e-02   2.433  0.014997 *
GenreAnimation  3.995e-01  1.131e-01   3.531  0.000418 ***
GenreBiography  9.251e-01  1.271e-01   7.277  3.94e-13 ***
GenreComedy     1.959e-01  6.008e-02   3.261  0.001116 **
GenreDrama      5.958e-01  5.696e-02  10.459 < 2e-16 ***
GenreHorror     -4.002e-01  7.173e-02  -5.580  2.54e-08 ***
GenreRomance    4.255e-01  1.390e-01   3.061  0.002220 **
```

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

```
Residual standard error: 1.303 on 4988 degrees of freedom
Multiple R-squared:  0.1209,    Adjusted R-squared:  0.119
F-statistic: 62.36 on 11 and 4988 DF,    p-value: < 2.2e-16
```

"The RMSE of the model is 1.303423"

## XGBoost (nround=10)

```
[1] train-rmse:4.182815
[2] train-rmse:3.061179
[3] train-rmse:2.313911
[4] train-rmse:1.829632
[5] train-rmse:1.530007
[6] train-rmse:1.348144
[7] train-rmse:1.247435
[8] train-rmse:1.187237
[9] train-rmse:1.154672
[10] train-rmse:1.136153
```

	Length	Class	Mode
handle	1	xgb.Booster.handle	externalptr
raw	32406	-none-	raw
niter	1	-none-	numeric
evaluation_log	2	data.table	list
call	16	-none-	call
params	4	-none-	list
callbacks	2	-none-	list
feature_names	5	-none-	character
nfeatures	1	-none-	numeric

# Deployment

<https://zhouyichen961104.shinyapps.io/stat-418-project/>

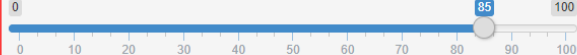
## IMDb Score prediction

### Way 1 to input data

Popularity\_Rank

50

Metascore



Votes

60000

Run\_time

120

Genre

Adventure

Model

- ☐ Linear Regression  
☒ XGBoost

IMDbscore

7.16

### Way 2 to input data

Input .csv File

Browse... No file selected

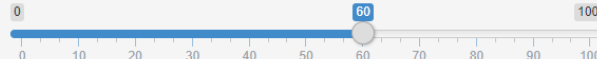
## IMDb Score prediction

### Way 1 to input data

Popularity\_Rank

500

Metascore



Votes

600

Run\_time

90

Genre

Action

Model

- ☒ Linear Regression  
☐ XGBoost

IMDbscore

5.83

6.02

4.38

8.16

9.21

3.95

5.44

6.30

7.57

7.82

4.11

5.20

5.86

8.61

6.65

4.99

7.08

7.71

8.86

3.80

### Way 2 to input data

Input .csv File

Browse... examp.csv

Upload complete