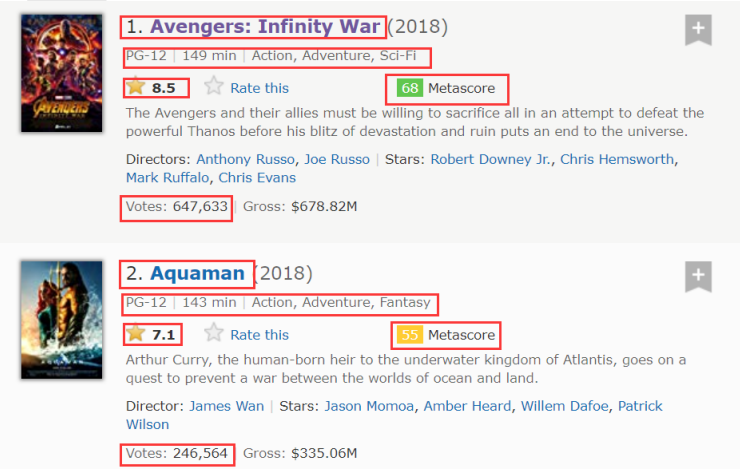
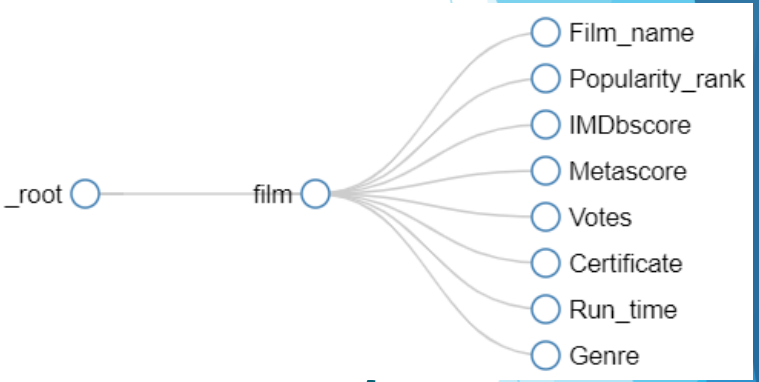
Predict the IMDb score of films released in 2018

Stat 418 project

Zhou, Yichen

Attaining Data – Web Scraper

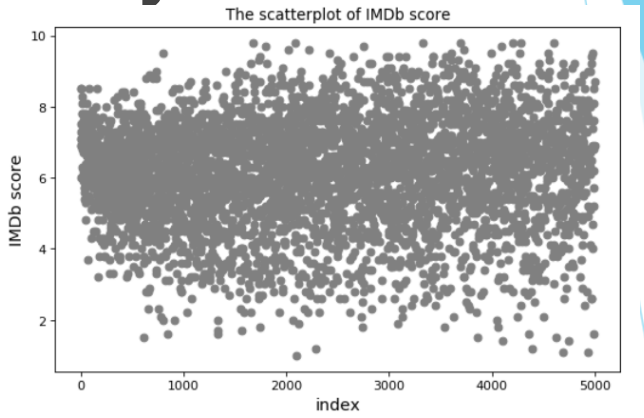


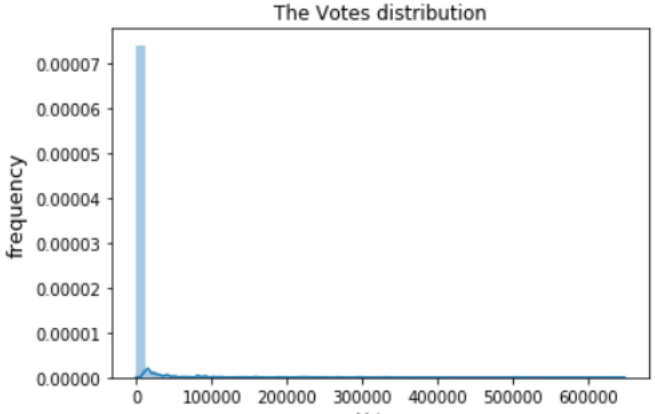


	web-scraper-order	web-scraper-start-url	Film_name	Popularity_rank	IMDbscore	Metascore	Votes	Certificate	Run_time	Genre
0	1557261010-11301	https://www.imdb.com/search/title?title_type=f...	Avengers: Infinity War	1	8.5	68.0	647469.0	PG-12	149 min	Action, Adventure, Sci-Fi
1	1557261010-11302	https://www.imdb.com/search/title?title_type=f...	Aquaman	2	7.1	55.0	246474.0	PG-12	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	PG-13	98 min	Adventure, Drama

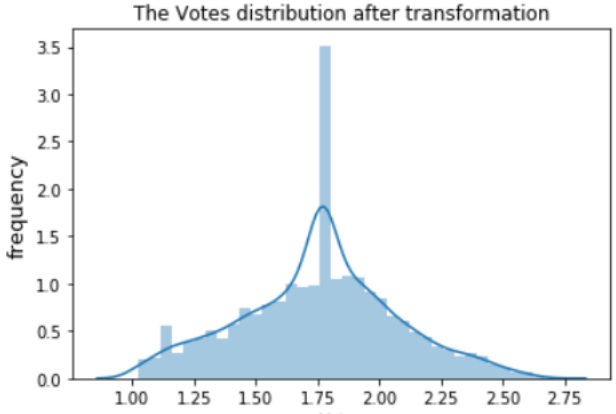
Exploratory data analysis

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 10 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
Certificate             1367 non-null object
Run_time                4284 non-null object
Genre                   4989 non-null object
dtypes: float64(3), int64(1), object(6)
memory usage: 390.7+ KB
```





$\ln(y+1)$



Further Research

- Input: film information data(Popularity rank, Meta score, Votes, Certificate, Run time, Genre)
- Output: estimation of the IMDb score of the film.

- Complete the EDA： Genre variable, plots to show relationships between some variables
- Model Fitting: KRR , XGBoost.
- Model Evaluating: Loss function = MSE, R-Squared score