Movie Data Analysis

-- Utilizing the Data Collected From IMDb and Kaggle

Zhuoxuan Li 116010123
Shunan Jiang 115010026
Ran Hu 116010078

Introduction

Things we want to know about *Avengers: Endgame*

Release Date: April 24, 2019
Directors: Anthony Russo, Joe Russo
Stars: Brie Larson, Scarlett Johansson, Karen Gillan …
Production Co: Marvel Studios
Runtime: 181 min
Genres: Action | Adventure | Fantasy

IMDb Rating: ?

Gross: ?

Similar Movies: ?

How to predict the rating and make recommendations?

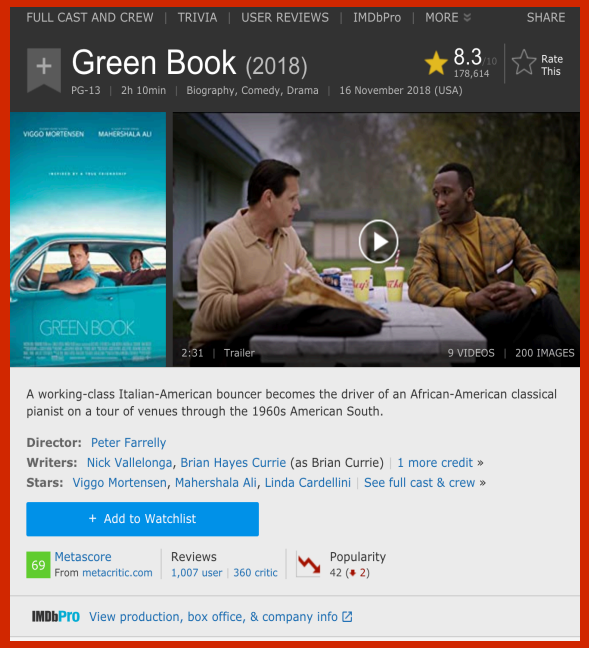
100, 000 movies

visualization, regression, clustering

Multiple Sources

Top 1000 Actors and Actresses
a list of 1000 people

https://www.imdb.com/list/ls058011111



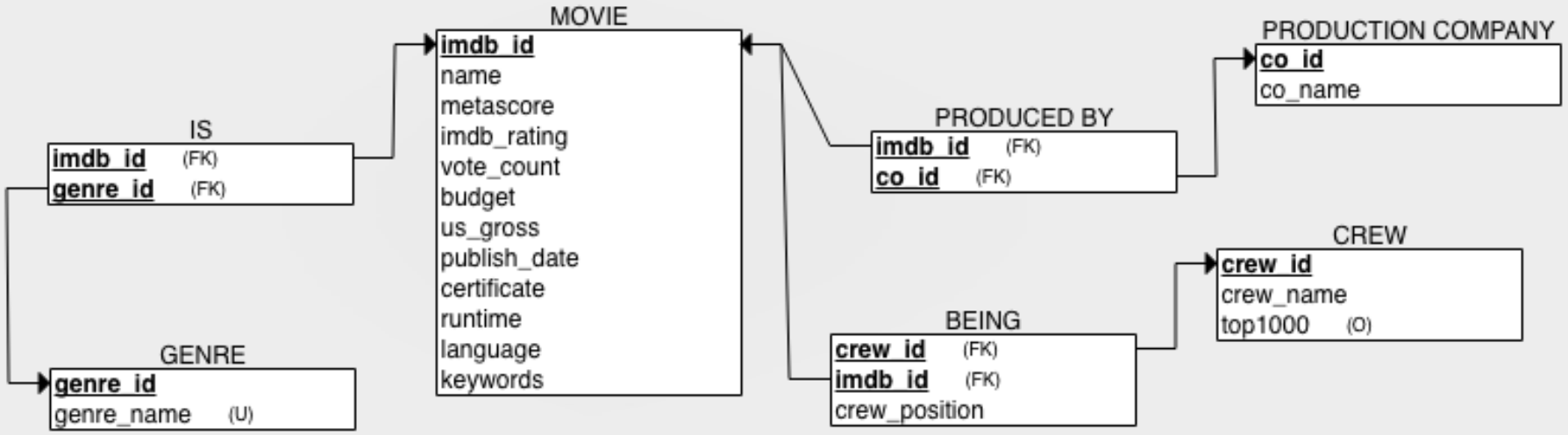
https://www.imdb.com/title/tt6966692/?ref_=nv_sr_1?ref_=nv_sr_1

kaggle™

IMDB Movies Dataset
Over 14,000 movies from IMDB

https://www.kaggle.com/orgesleka/imdbmovies

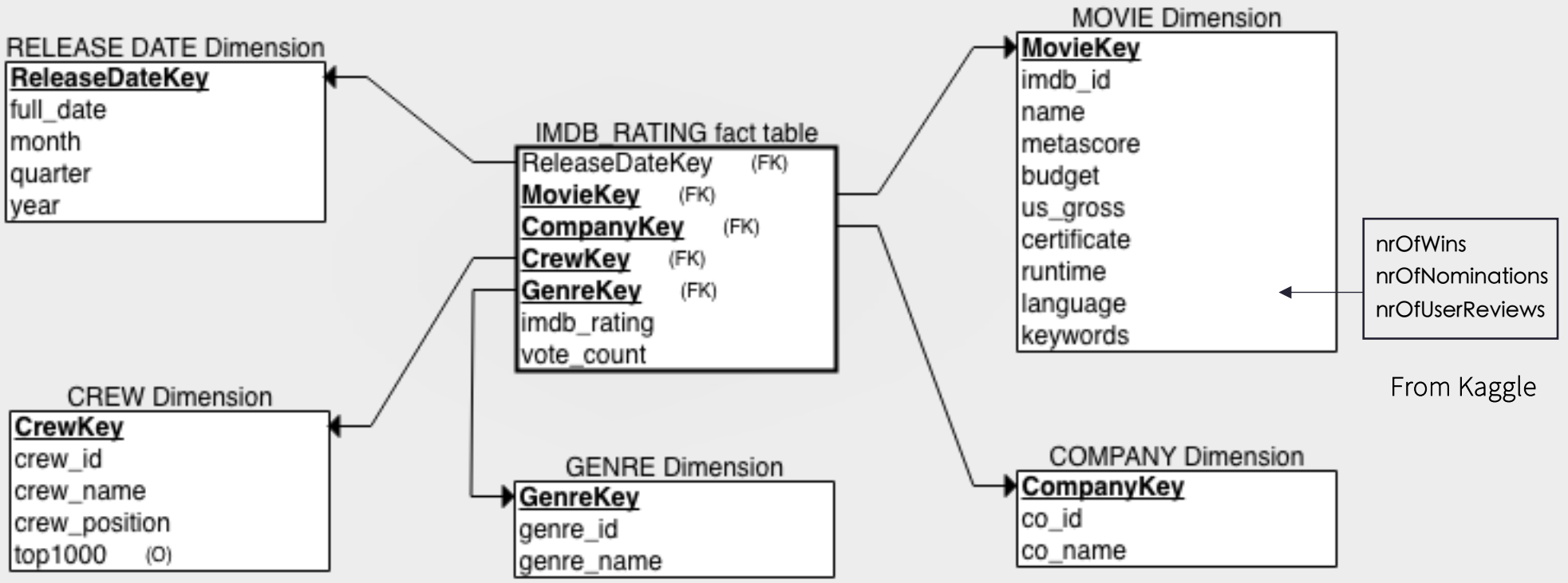
Relational modeling



Relational schema: IMDb Movie Database (Source: IMDb)

Our operational database

Dimensional modeling



IMDb Movie dimensional model for the subject **imdb_rating** (Source: IMDb & Kaggle)

Future improvement: an analytical database

Web Crawler

More than 100k movie information
to make our analysis **reliable**.

Challenges In Web Crawling

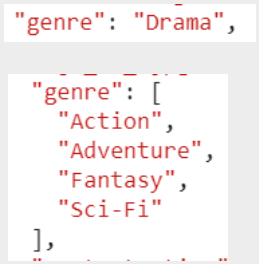
1. Speed of Crawling

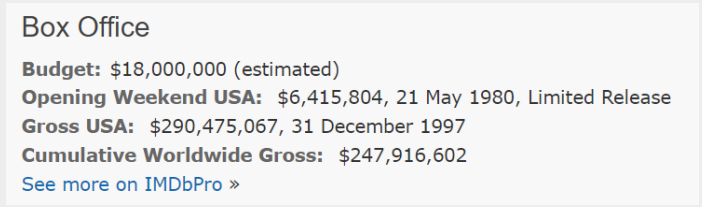
1s per row in average = More than 29h In total

2. Unstable Internet Connection

3. Handle Different Data Type / Absent Info

(None, str or list)





Full Box Office Info v.s. No Box Office Info

4. Formatting Information

E.g. Different Monetary Unit (CAD or $)

E.g. Containing Comma v.s. No Comma in numbers

Solutions & Key Points

1. Accelerate our crawling process

VPN + 4G Network

Fetching Most Info from embedded json instead of html content

2. Build Robust Program

Use “Try…Except” to handle errors

Build program which can be restart at anytime

Handle different data type (string or list in json)

Handle absent data

Record fail information in file for traceback

Solutions & Key Points

3. Use Headers to Change Accepted Language into English

4. Simple Data formatting

E.g.

Delete comma in numbers

Combine Multiple Content for save (Language)

5. Fix Genre Column

Discover Mistakes

Recover(Select affected rows and Fetch Again)

Data Analysis

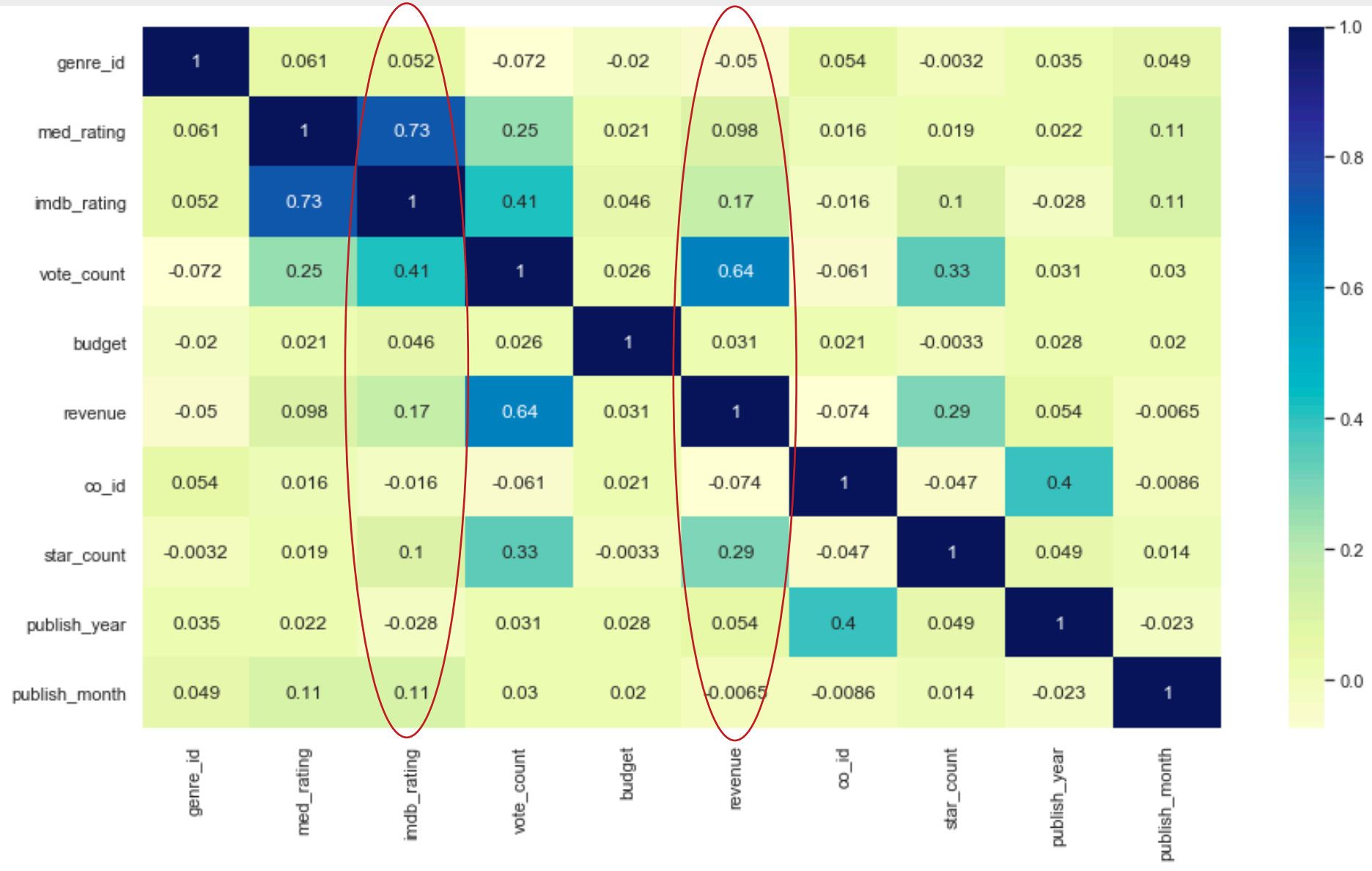
1. Overall Statistics
2. Correlation Analysis
3. Dependency

Overall Statistics

Star_count is an attribute created by ourself.
It means the number of superstars in a movie.

	Med_Rating	Imdb_Rating	Vote_Count	Budget	Revenue	**Star_Count**	Year	Month
Mean	56.05702	**6.524455**	9.74E+04	5.92E+07	4.36E+07	1.806676	2007.275	6.921293
std	17.54054	0.94115	1.49E+05	5.35E+08	6.97E+07	1.377649	9.383668	3.434003
Min	1	1.5	4.70E+01	1.50E+03	3.88E+02	0	1980	1
50%	56	6.6	4.37E+04	2.10E+07	1.75E+07	2	2011	**7**
Max	100	9.3	2.08E+06	3.00E+10	9.37E+08	6	2019	12

Correlation Analysis

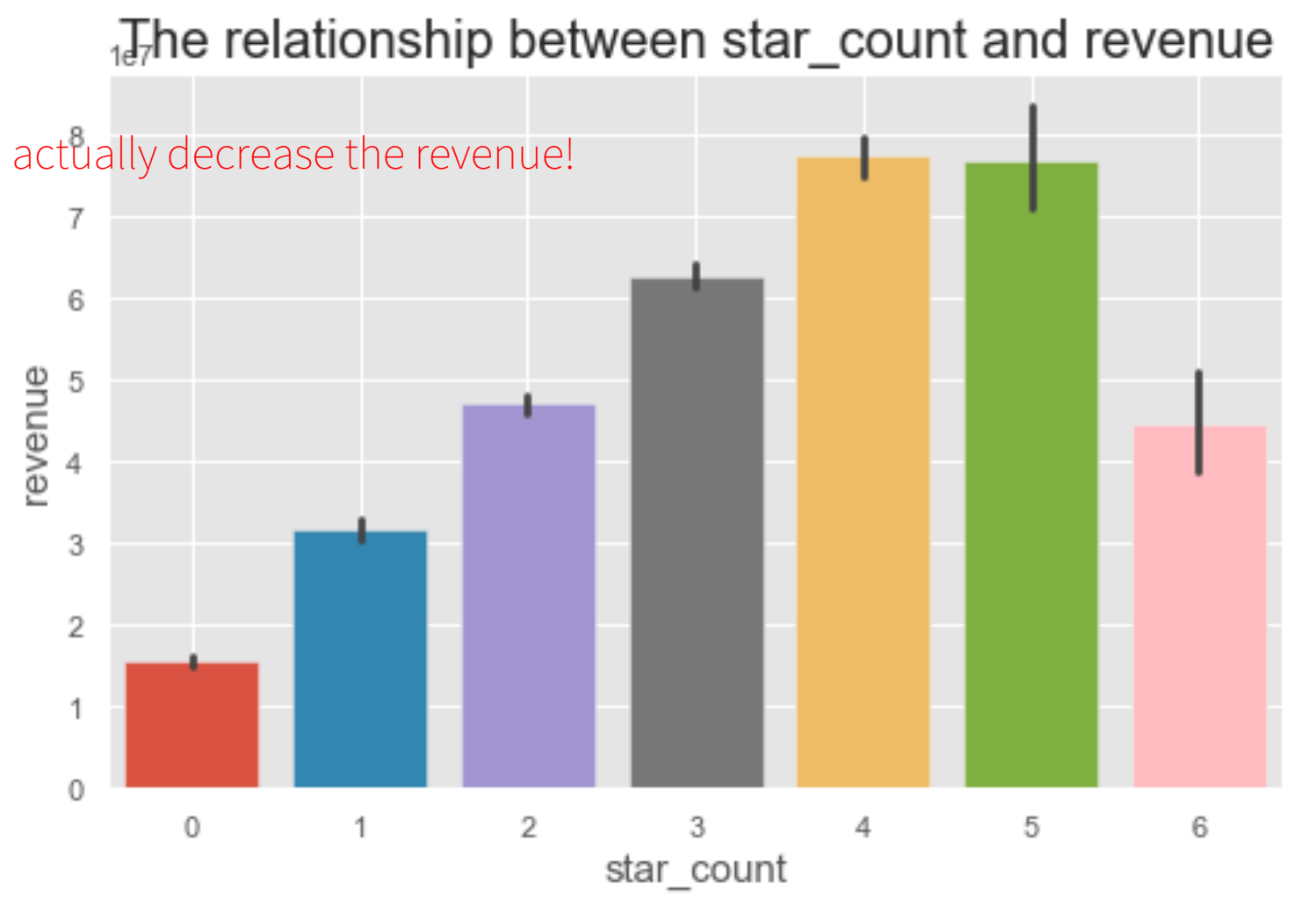


Imdb_rating:
1. med_rating
2. vote_count
3. revenue

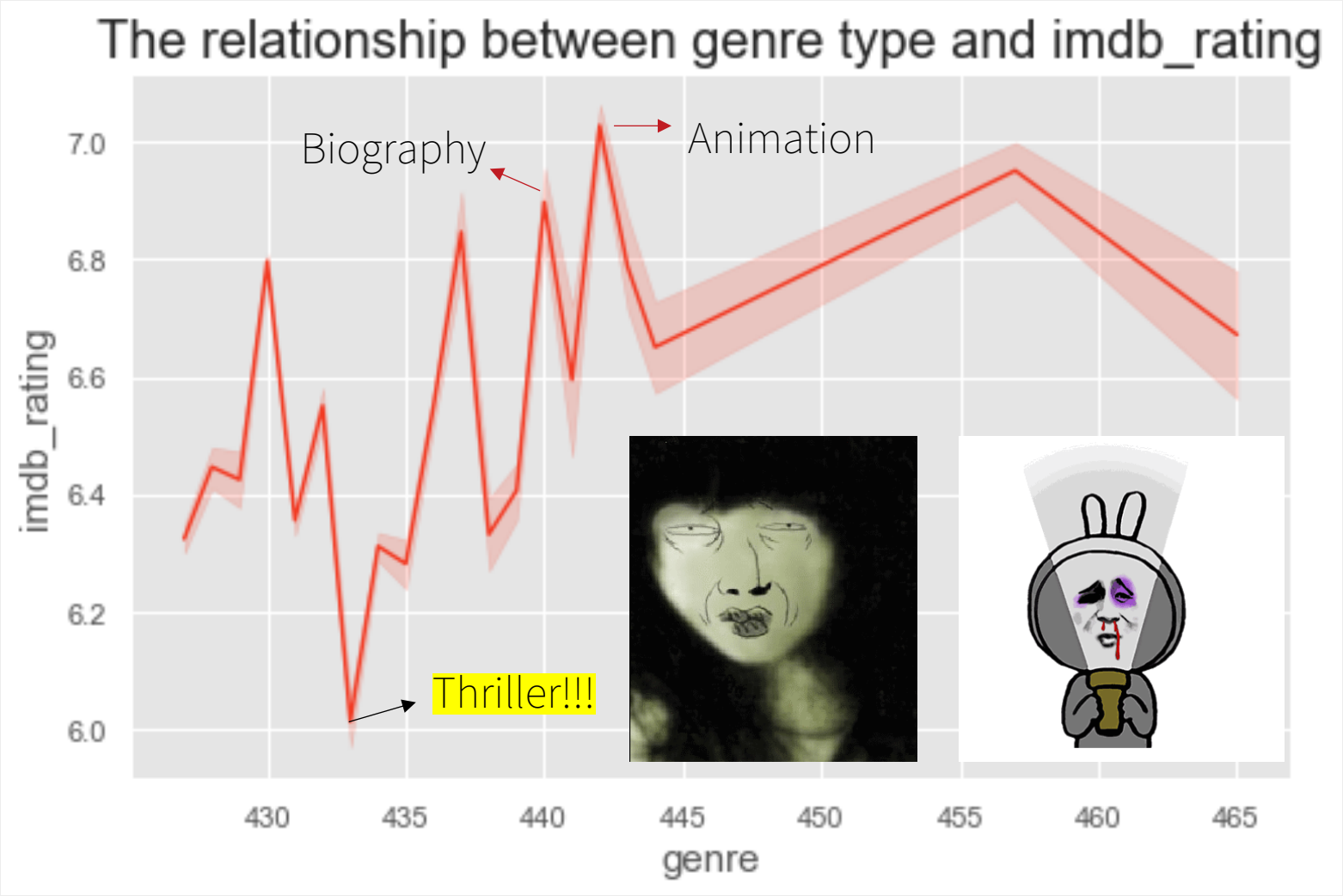
Revenue:
1. vote_count
2. imdb_rating
3. star_count

Diving Deeper (I) – Dependency – star_count

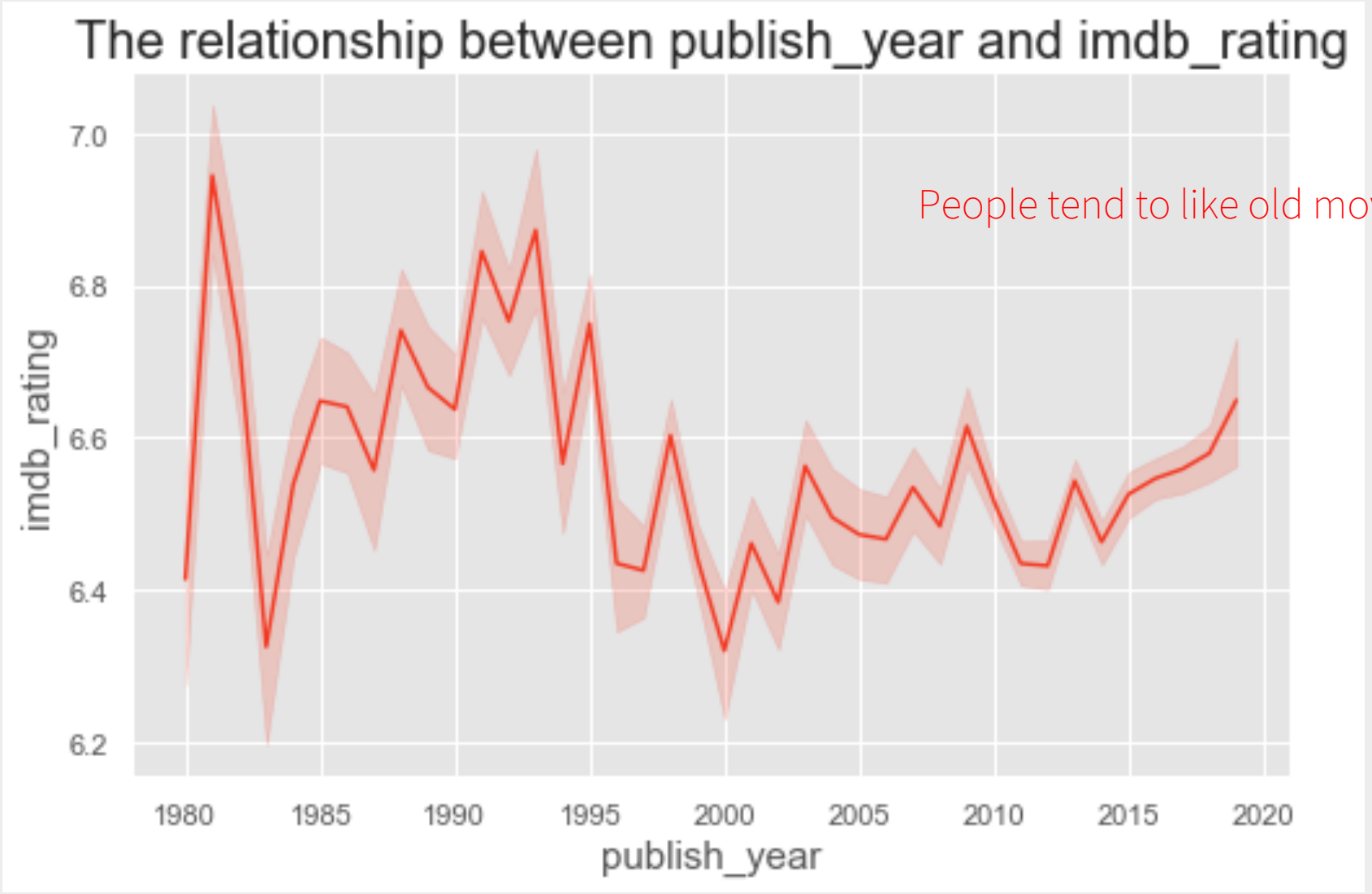
Too many stars will actually decrease the revenue!



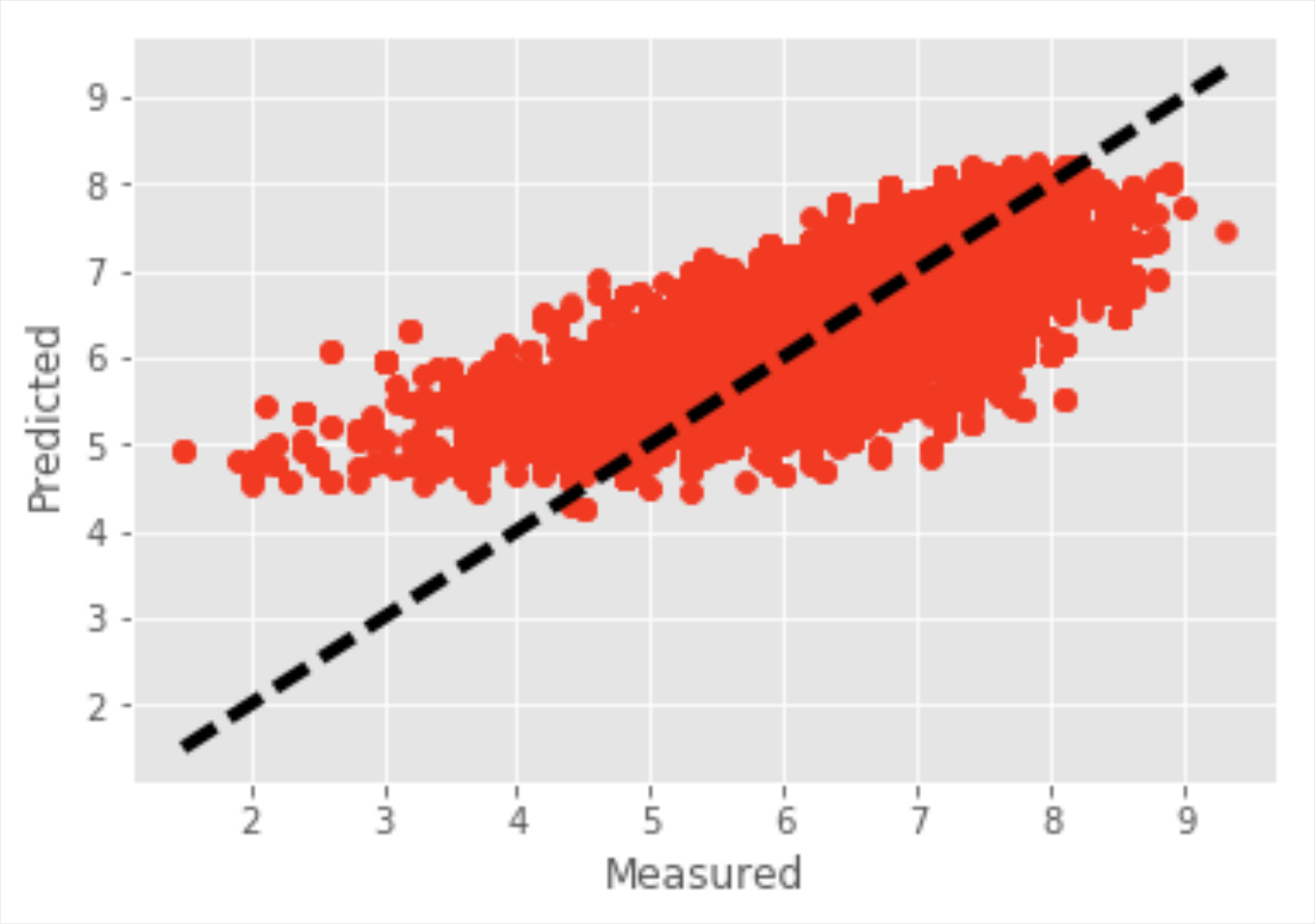
Diving Deeper (II) – Dependency - Genre



Diving Deeper (III) – Dependency - Year



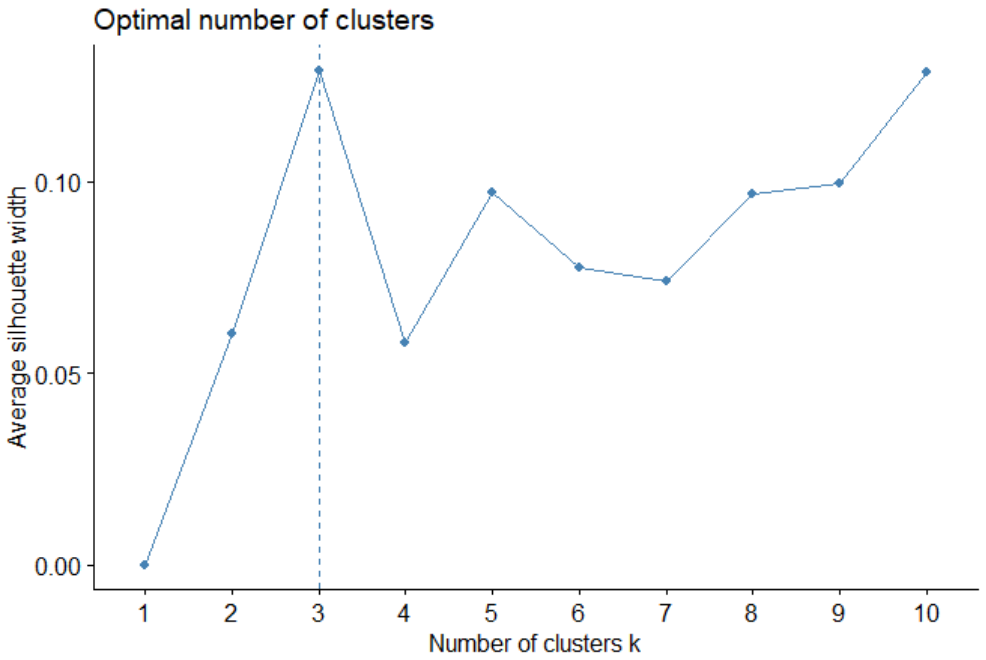
Predict Rating Scores – Multiple Regression

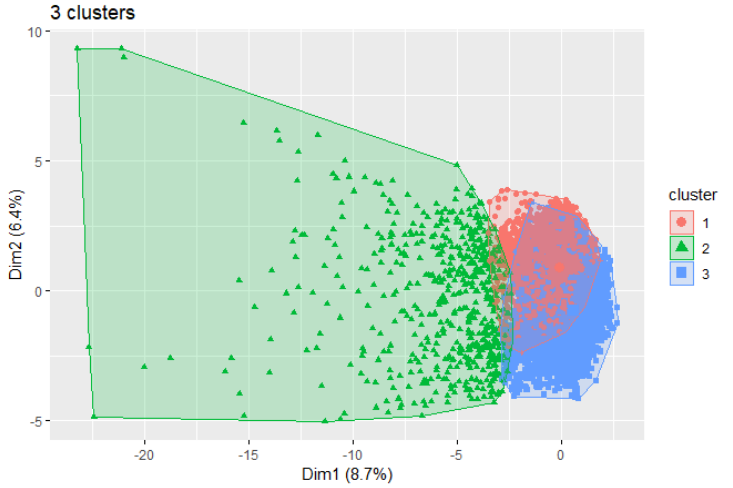


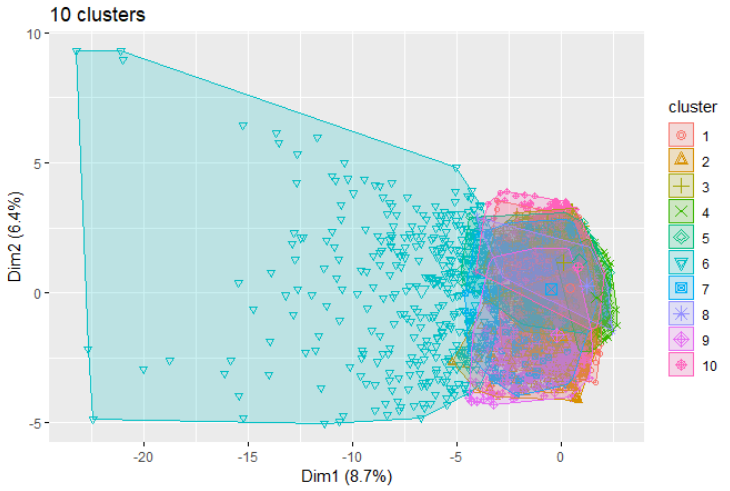
Predicted Score: 7.59

K-means clustering – find similar movies

among 13,000+ movies (source: Kaggle)

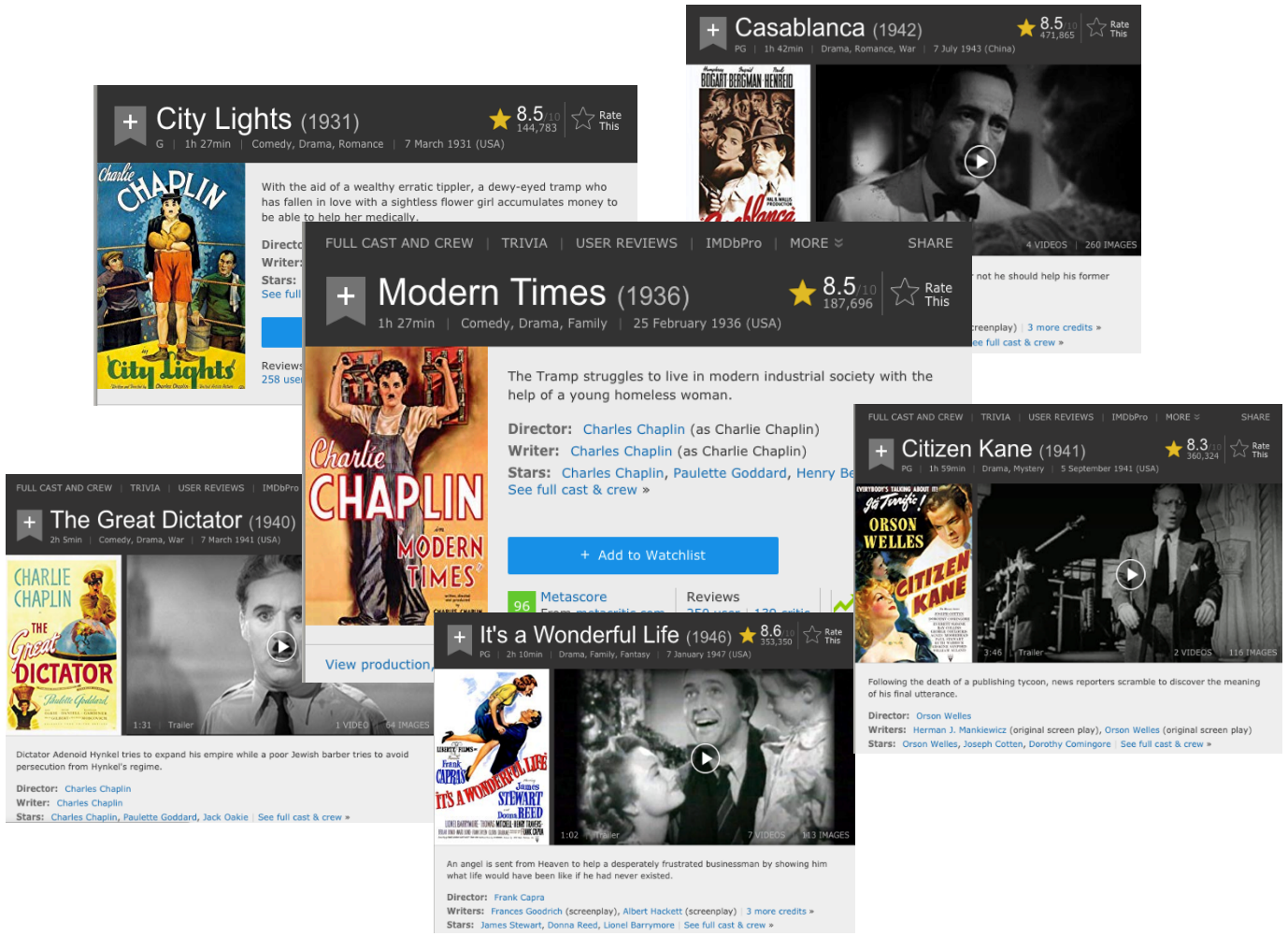




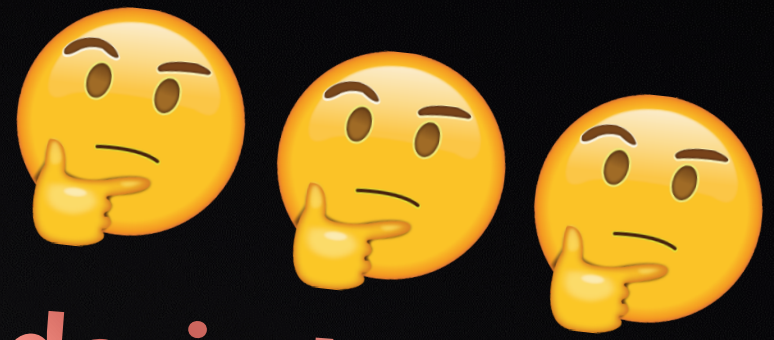


See results of 10-means clustering in clustering_data&result.csv

A recommended movie list for *Modern Times* (from clustering results)

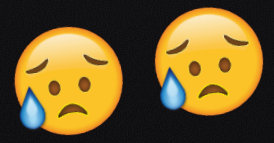


Imdb_id	Name	Imdb_Rating	Vote_Count	Publish_Date
tt0122050	Xin qi long zhu 新七龙珠	4.2	1001	2000/6/13
tt1328865	Ci ling 刺陵	3.8	1048	2009/12/30
tt1478291	Xun zhao Cheng Long 寻找成龙	4.1	1014	2009/7/3
tt1717715	Xi you ji: Da nao tian gong 西游记：大闹天宫	4.9	4338	2014/1/30
tt1847713	Xue di zi 血滴子	4.8	1202	2012/12/20
tt2460488	Tian ji: Fu chun shan ju tu 天机：富春山居图	2.4	1641	2013/6/9
tt2644714	Bu er shen tan 不二神探	4.7	2443	2013/6/21
tt3465456	Bai fa mo nu zhuan zhi ming yue tian guo 白发魔女传之明月天国	5.4	1401	2014/4/25
tt3585004	Zhong Kui fu mo: Xue yao mo ling 钟馗伏魔：雪妖魔灵	5.4	1101	2015/2/19
tt4819498	Jiu ceng yao ta 九层妖塔	5.2	1149	2015/9/30
tt5273624	Journey to the West: Demon Chapter 西游2：伏妖篇	5.4	2286	2017/1/28
tt5481184	Feng shen bang 封神榜	4.6	1757	2016/7/28



12 Mandarin Movies

Who’s Rating **BELOW 5.5**
But vote by **MORE THAN 1k** people
FROM 1980-2019, ON IMDB



Have you ever tried One of them?

Scan for google doc version →