# A PEEK OF GROUP DIFFERENCES FROM MOVIE RATING PREFERENCE

JIALAN ZHU

# PROBLEM STATEMENT

MOVIE

RATING

DIFFERENCE

DOMESTIC

IMDb

INTERNATIONAL

IMDb

豆

# WORK FLOW

WEB SCRAPING (SCRAPY)

IMDb

IMDb + 豆

DATA MANIPULATION

**PYTHON3**

**NUMPY**

**PANDAS**

**RE**

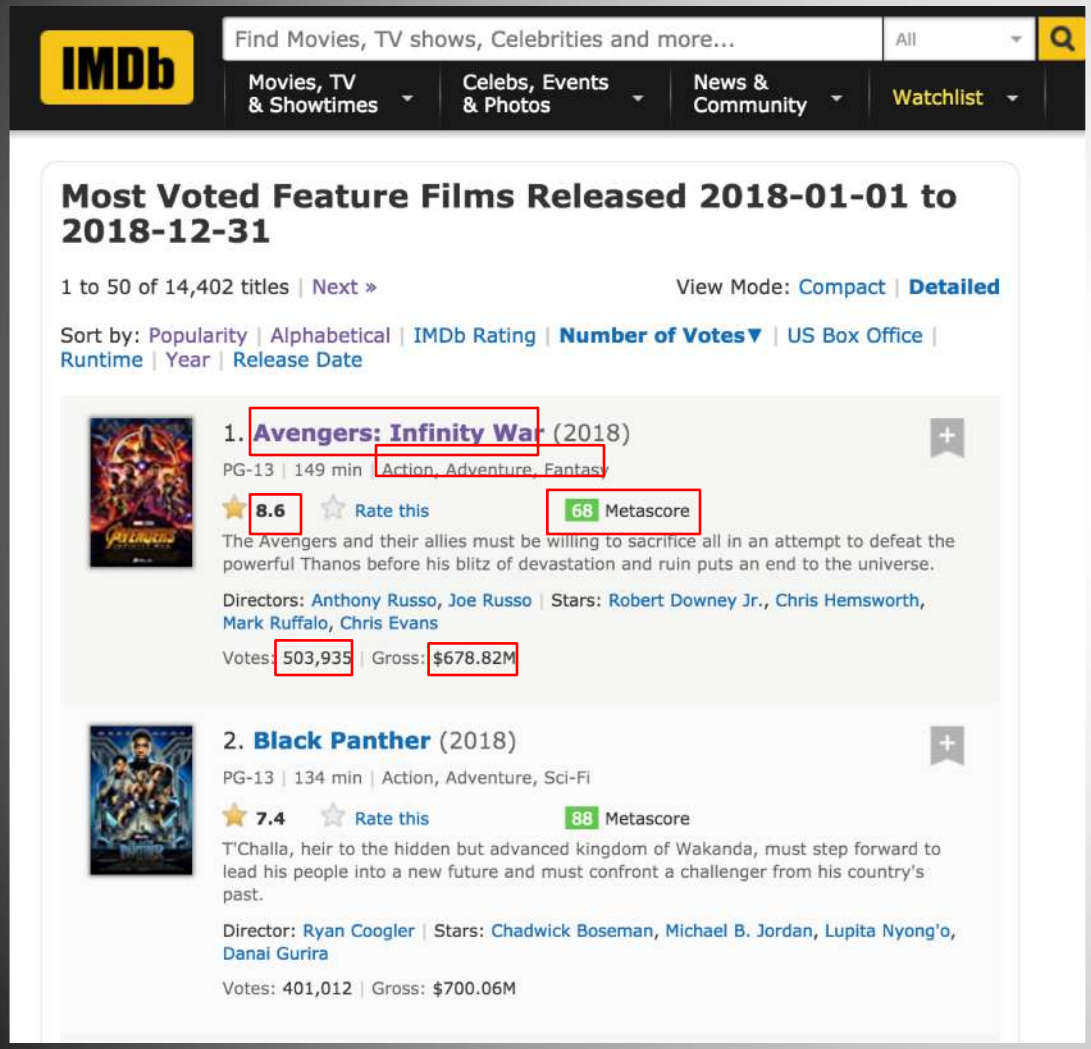
DATA VISUALIZATION

**PYTHON3**

**SEABORN**

**MATPLOTLIB**

# WEB SCRAPING



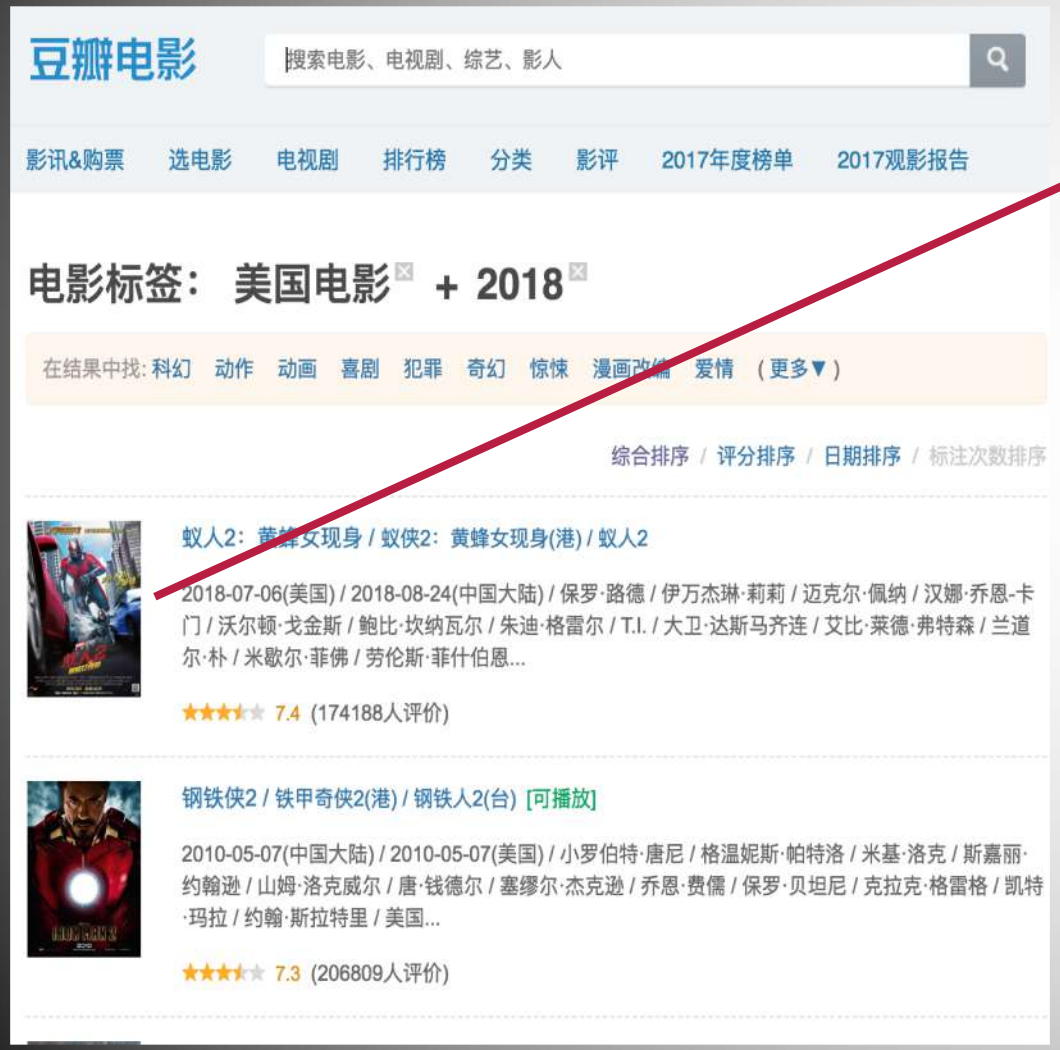

SCRAPING DATA:

1. MOVIE NAME
2. GENRE
3. IMDB RATING
4. METASCORE
5. NUMBER OF VOTES
6. GROSS

**(6301 X 7)**

| | genre | gross | imdb_rating | meta_rating | movie_name | movie_rated | n_of_votes |
|---|---|---|---|---|---|---|---|
| 0 | Drama | NaN | 5.6 | NaN | In My Room | NaN | 96 |
| 1 | Thriller | NaN | 5.4 | NaN | The Wrong Cruise | TV-14 | 96 |
| 2 | Comedy | NaN | 8.0 | NaN | Vadhayiyaan Ji Vadhayiyaan | NaN | 96 |
| 3 | Biography, Comedy, Music | NaN | 7.3 | NaN | Ana e Vitória | NaN | 96 |
| 4 | Drama | NaN | 6.7 | NaN | L'animale | NaN | 96 |
| 5 | Horror, Thriller | NaN | 4.7 | NaN | The Witch Files | NaN | 95 |

# WEB SCRAPING



SCRAPING DATA:

1. MOVIE NAME
2. DOUBAN RATING
3. NUMBER OF VOTES

## 403 ERROR!

headers= {'User-Agent': 'Mozilla/5.0 (X11; Linux x86_64; rv:48.0) Gecko/20100101 Firefox/48.0'}
Request(url, headers=headers)

**(698 X 3)**

| | movie_name_douban | movie_rating_douban | n_of_vote_douban |
|---|---|---|---|
| 0 | 漫威影业：宇宙集结 Marvel Studios: Assembling a Universe | 8.5 | 2999.0 |
| 1 | 极品老妈 第三季 Mom Season 3 | 8.9 | 3595.0 |
| 2 | 甜蜜与卑微 Sweet and Lowdown | 7.8 | 3509.0 |
| 3 | 罗尼和我 Ronny & I | 8.6 | 4212.0 |
| 4 | 主厨的餐桌 第一季 Chef's Table Season 1 | 9.2 | 4986.0 |

# DATA MANIPULATION & VISUALIZATION

**(6301 X 7)**

| | genre | gross | imdb_rating | meta_rating | movie_name | movie_rated | n_of_votes |
|---|---|---|---|---|---|---|---|
| 0 | Drama | NaN | 5.6 | NaN | In My Room | NaN | 96 |
| 1 | Thriller | NaN | 5.4 | NaN | The Wrong Cruise | TV-14 | 96 |
| 2 | Comedy | NaN | 8.0 | NaN | Vadhayiyaan Ji Vadhayiyaan | NaN | 96 |
| 3 | Biography, Comedy, Music | NaN | 7.3 | NaN | Ana e Vitória | NaN | 96 |
| 4 | Drama | NaN | 6.7 | NaN | L'animale | NaN | 96 |
| 5 | Horror, Thriller | NaN | 4.7 | NaN | The Witch Files | NaN | 95 |

1. DROP NA
2. CONVERT RATING SYSTEM
3. CONVERT STING TYPE NUMBER TO NUMERIC
4. REMOVE NON-ALPHABIC CHARACTER IN MOVIE NAMES
5. CREATE COLUMN FOR MATCHING

**(406 X 8)**

| | genre | gross | imdb_rating | meta_rating | movie_name | movie_rated | n_of_votes | movie_name_for_match |
|---|---|---|---|---|---|---|---|---|
| 0 | drama | NaN | 5.6 | NaN | in my room | NaN | 96 | in my room |
| 1 | thriller | NaN | 5.4 | NaN | the wrong cruise | TV-14 | 96 | the wrong cruise |
| 2 | comedy | NaN | 8.0 | NaN | vadhayiyaan ji vadhayiyaan | NaN | 96 | vadhayiyaan ji vadhayiyaan |
| 3 | biography, comedy, music | NaN | 7.3 | NaN | ana e vitória | NaN | 96 | ana e vit ria |
| 4 | drama | NaN | 6.7 | NaN | l'animale | NaN | 96 | l animale |

**(3289 X 8)**

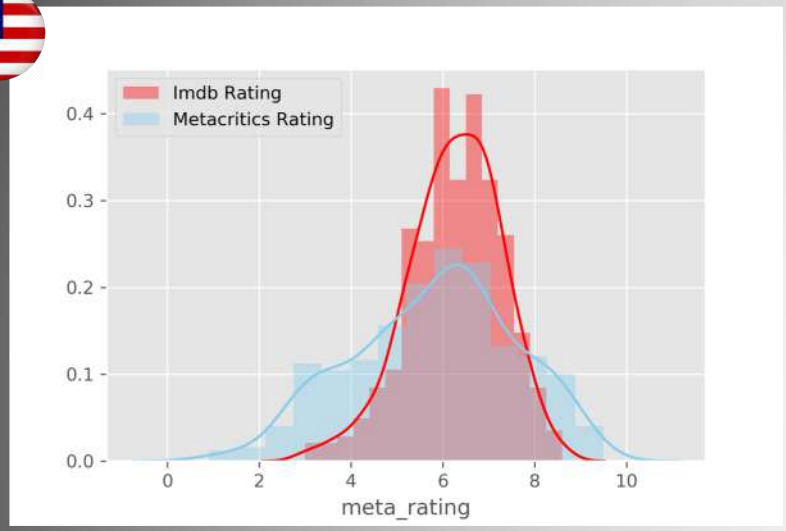
**(599 X 4)**

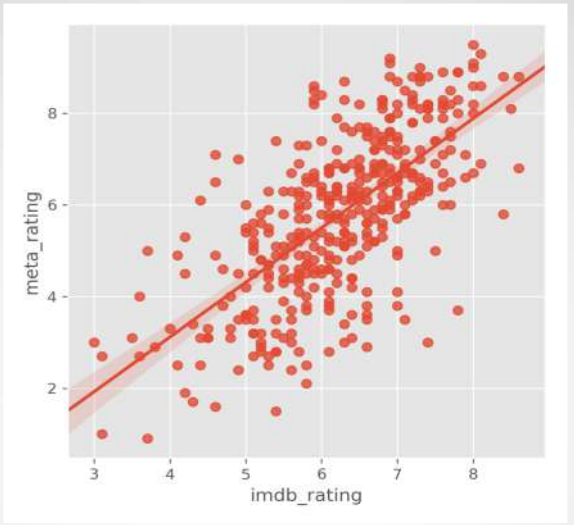
1. INNER MERGE ON MOVIE NAME
2. REMOVE UNNECESSARY COLUM
3. REMVOE DUPLICATE MOVIES

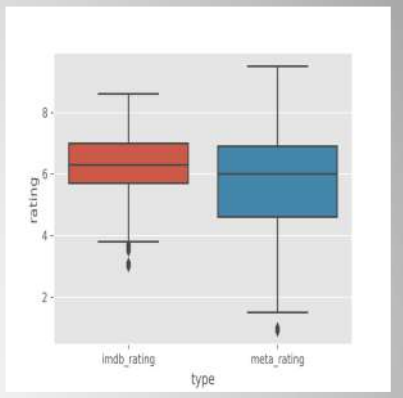
**(35 X 10)**

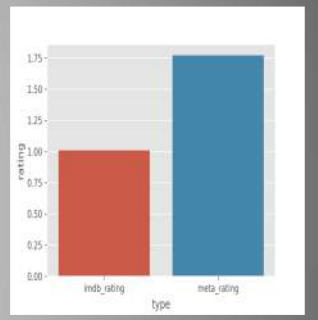
| | genre | gross | imdb_rating | meta_rating | movie_name | movie_rated | n_of_votes | movie_name_for_match | movie_rating_douban | n_of_vote_douban |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | horror, thriller | NaN | 6.9 | NaN | sabrina | NaN | 111 | sabrina | 7.7 | 16205.0 |
| 1 | drama | NaN | 3.6 | NaN | macbeth | NaN | 118 | macbeth | 7.2 | 9710.0 |
| 2 | drama, family | $1.29M | 4.8 | 40.0 | little women | PG-13 | 207 | little women | 8.1 | 11498.0 |
| 3 | action, thriller | NaN | 4.9 | NaN | genius | NaN | 434 | genius | 7.5 | 31606.0 |
| 4 | biography, drama | NaN | 8.0 | 91.0 | the favourite | R | 907 | the favourite | 7.6 | 146.0 |

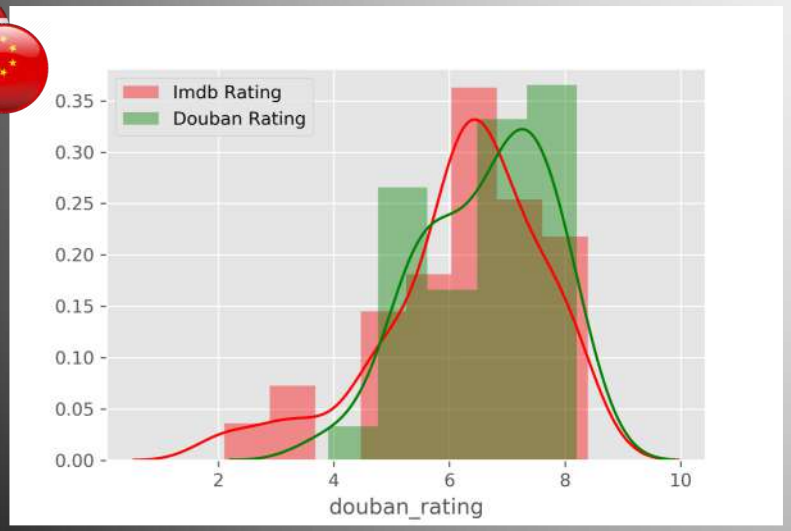
# DATA MANIPULATION & VISUALIZATION

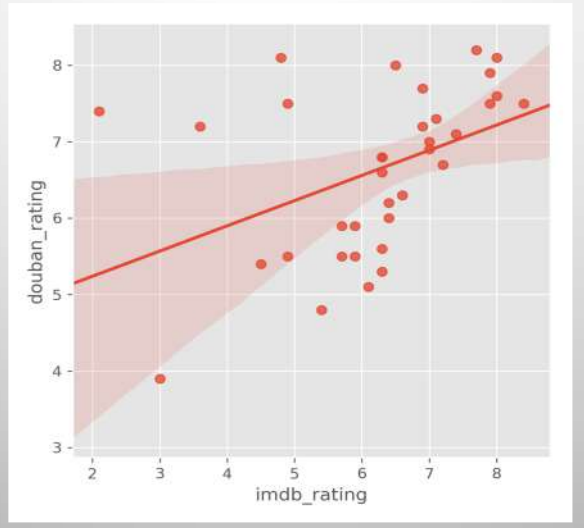


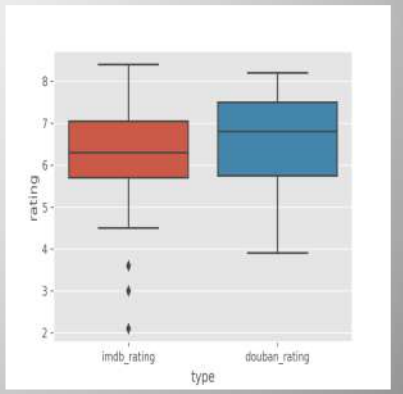



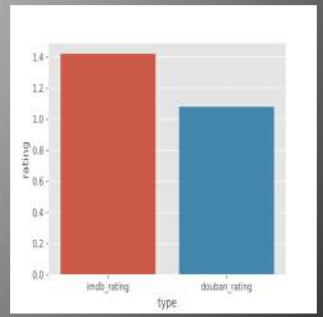



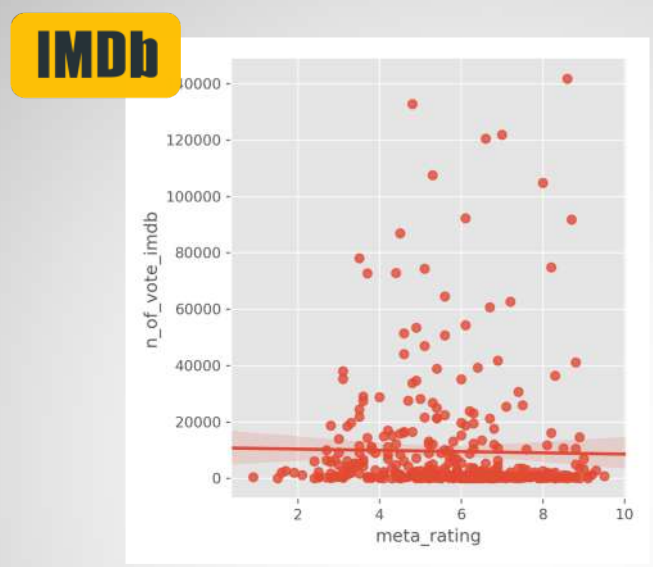



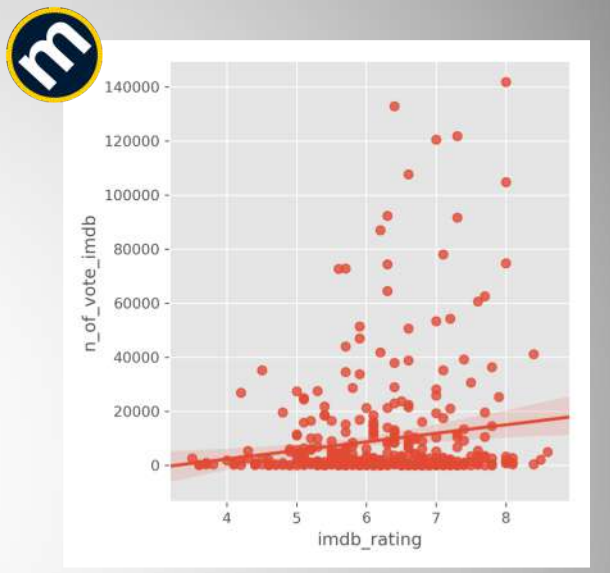



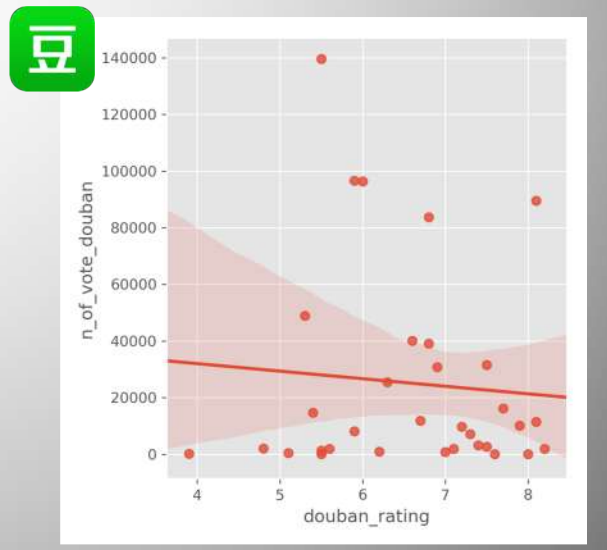



# DATA MANIPULATION & VISUALIZATION

POPULARITY
INFLUENCE





# DATA MANIPULATION & VISUALIZATION

GROUP BY GENRE

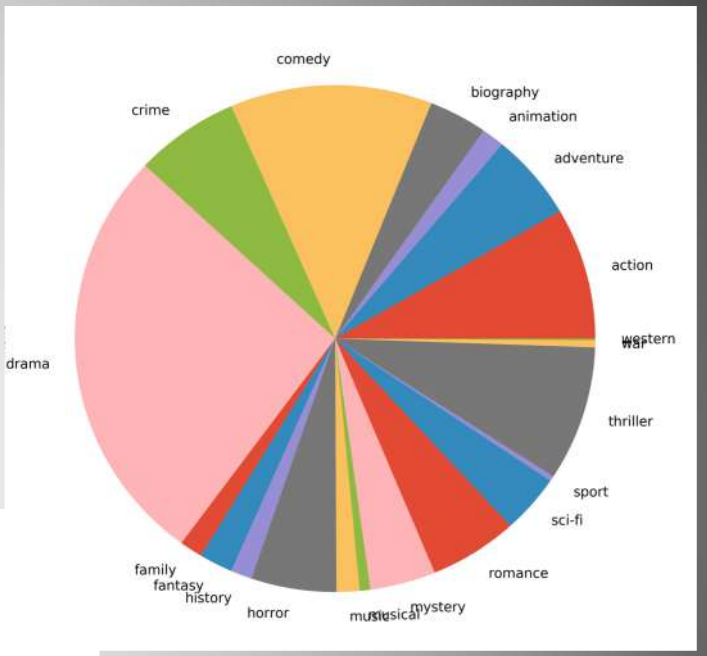
| genre |
| --- |
| drama |
| thriller |
| comedy |
| biography, comedy, music |
| drama |
| horror, thriller |

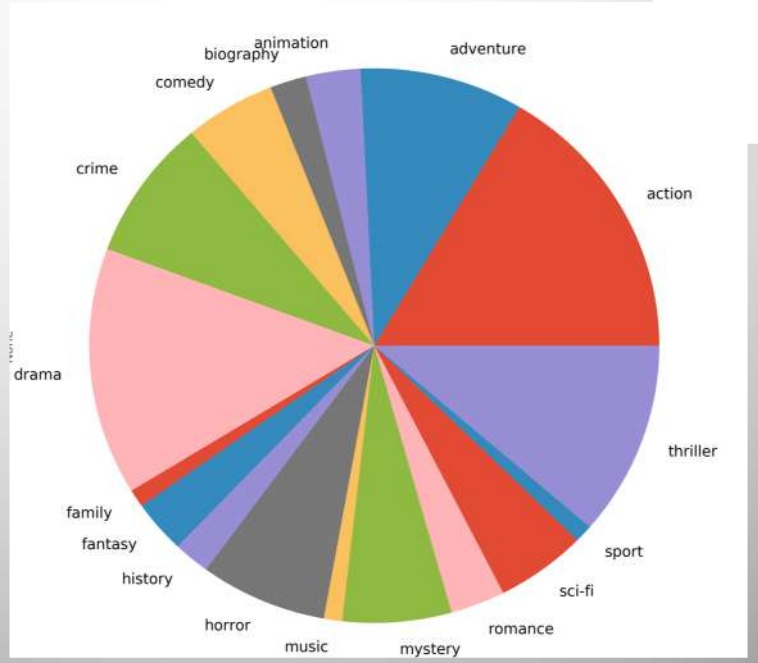
MANIPULATIONS

1. REMOVE COMMA
2. STACK THE DATA

   FRAMES
3. GROUP BY GENRE

**(907 X 7)**



**(97 X 9)**

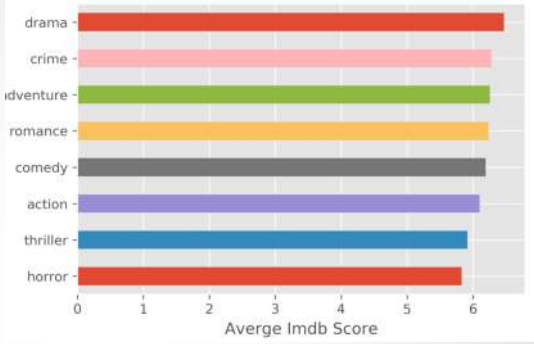
# DATA MANIPULATION & VISUALIZATION

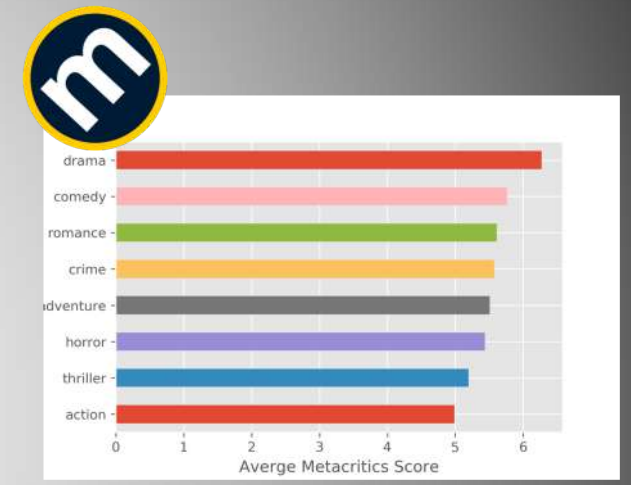
IMDb

| movie_name |
| --- |
| the wild pear tree |
| avengers: infinity war |
| thunder road |
| a star is born |
| dragged across concrete |

| movie_name |
| --- |
| roma |
| shoplifters |
| sunday's illness |
| ray & liz |
| the favourite |

IMDb



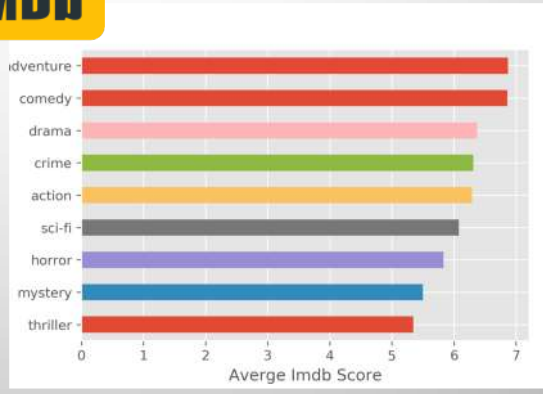
IMDb

| movie_name |
| --- |
| deadpool 2 |
| annihilation |
| solo: a star wars story |
| tomb raider |
| red sparrow |

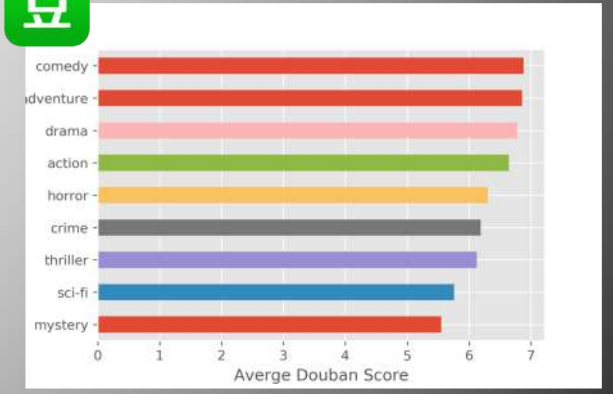
豆

| movie_name |
| --- |
| annihilation |
| pacific rim: uprising |
| deadpool 2 |
| the meg |
| tomb raider |

IMDb




豆



# Question?