

분석 방향 참고

<https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset/code?datasetId=3405&sortBy=voteCount>

movie dataset 옆에 code부분 눌러보시면 해당 데이터로 다른 사람들이 짜 놓은 코드들 볼 수 있습니다!

저희가 코드를 전부 이해하긴 쉽지 않으니 다른 사람들은 어떤 column들로 어떤 활용을 해봤는지 정도는 참고하시기 좋은 것 같습니다!

혹시 분석 방향 생각하시는데 감이 안잡히시면 참고해주세요!

예시)

<https://www.kaggle.com/code/rounakbanik/movie-recommender-systems/notebook>

```
In [3]: md['genres'] = md['genres'].fillna('').apply(literal_eval).apply(lambda x: [i['name'] for i in x] if
isinstance(x, list) else [])
```

In [3] : 뭔진 잘 모르겠지만 genres 컬럼을 써서 해봤구나!

```
In [4]: vote_counts = md[md['vote_count'].notnull()]['vote_count'].astype('int')
vote_averages = md[md['vote_average'].notnull()]['vote_average'].astype('int')
C = vote_averages.mean()
C
```

In[4] : vote_count (평균 수) 와 vote_average (평균 평점) 가지고 뭘 했구나!

```
In [6]: md['year'] = pd.to_datetime(md['release_date'], errors='coerce').apply(lambda x: str(x).split('-')[0] if
x != np.nan else np.nan)
```

In [6] : release_date 컬럼을 사용했네! datetime이 보이니 날짜로 바꾸고 뭘 했구나!

```
In [13]:
def build_chart(genre, percentile=0.85):
    df = gen_md[gen_md['genre'] == genre]
    vote_counts = df[df['vote_count'].notnull()]['vote_count'].astype('int')
    vote_averages = df[df['vote_average'].notnull()]['vote_average'].astype('int')
    C = vote_averages.mean()
    m = vote_counts.quantile(percentile)

    qualified = df[(df['vote_count'] >= m) & (df['vote_count'].notnull()) & (df['vote_average'].notnull())][['title', 'year', 'vote_count', 'vote_average', 'popularity']]
    qualified['vote_count'] = qualified['vote_count'].astype('int')
    qualified['vote_average'] = qualified['vote_average'].astype('int')

    qualified['wr'] = qualified.apply(lambda x: (x['vote_count']/(x['vote_count']+m) * x['vote_average']) + (m/(m+x['vote_count']) * C), axis=1)
    qualified = qualified.sort_values('wr', ascending=False).head(250)

    return qualified
```

In [13] : 복잡해 보이지만 일단 genre, vote_count , vote_average, popularity 컬럼을 써서 뭘 했구나!

```
In [27]:
get_recommendations('The Dark Knight').head(10)
```

```
Out[27]:
7931          The Dark Knight Rises
132          Batman Forever
1113          Batman Returns
8227  Batman: The Dark Knight Returns, Part 2
7565          Batman: Under the Red Hood
524          Batman
7901          Batman: Year One
2579  Batman: Mask of the Phantasm
2696          JFK
8165  Batman: The Dark Knight Returns, Part 1
Name: title, dtype: object
```

In[27] : get_recommendations가 뭔진 모르겠지만 뭘 만들어서 'The Dark Knight'를 입력했더니 배트맨 관련 영화가 나오는 걸 만들었구나!

※ 이 코드에서 생각해 볼 수 있는 내용 ※

- get_recommendations라는 걸 만들어서 영화 추천해주는 시스템을 만들었구나!
- 이 친구는 이걸 구현하기 위해서 genre, vote_count, vote_average, popularity, etc... 이런 걸 사용했구나!
- 만약 추천 시스템을 만든다고 하면 어떤 컬럼을 쓰는게 좋을까?