The Movie Database - tmdb

Team12:
Gabriela Trindade - gabrielatrindade
Kayen Fung - kayen88

What is tmdb?

- The Movie Database is a free and open source database on Movies
  - 'movie_id', 'title',
  - 'budget', 'revenue',
  - 'genres',
  - 'original_language', 'spoken_languages',
  - 'popularity', 'vote_average', 'vote_count'
  - 'production_companies', 'production_countries'
- Dataset provided by https://www.kaggle.com/tmdb/tmdb-movie-metadata?select=tmdb_5000_movies.csv
- Almost 5000 observations

Steps

1. Load the dataset
2. Data exploration
3. Ask questions
4. Cleaning
5. Merging
6. Answer questions

Data Exploration

Data Exploration

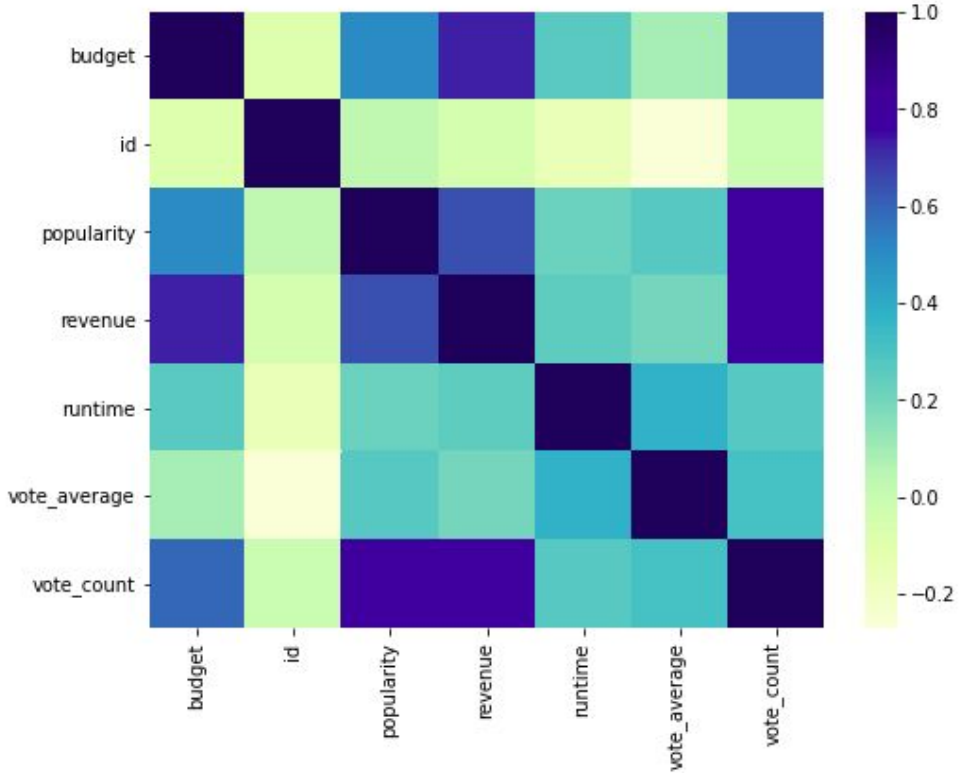
	budget	popularity	revenue	vote_average	vote_count
count	4.803000e+03	4803.000000	4.803000e+03	4803.000000	4803.000000
mean	2.904504e+07	21.492301	8.226064e+07	6.092172	690.217989
std	4.072239e+07	31.816650	1.628571e+08	1.194612	1234.585891
min	0.000000e+00	0.000000	0.000000e+00	0.000000	0.000000
25%	7.900000e+05	4.668070	0.000000e+00	5.600000	54.000000
50%	1.500000e+07	12.921594	1.917000e+07	6.200000	235.000000
75%	4.000000e+07	28.313505	9.291719e+07	6.800000	737.000000
max	3.800000e+08	875.581305	2.787965e+09	10.000000	13752.000000

```
tmdb_movies.info()
```

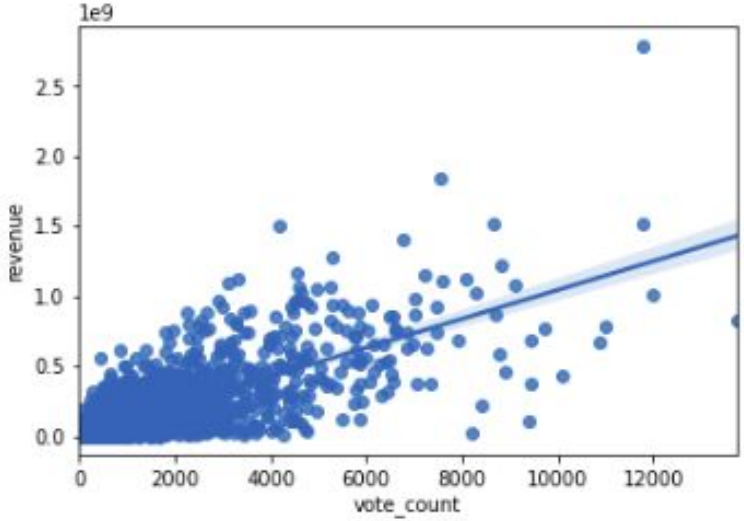
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4803 entries, 0 to 4802
Data columns (total 13 columns):
budget                  4803 non-null int64
genres                  4803 non-null object
id                      4803 non-null int64
original_language       4803 non-null object
popularity              4803 non-null float64
production_companies    4803 non-null object
production_countries    4803 non-null object
release_date            4802 non-null object
revenue                 4803 non-null int64
runtime                 4801 non-null float64
spoken_languages        4803 non-null object
vote_average            4803 non-null float64
vote_count              4803 non-null int64
dtypes: float64(3), int64(4), object(6)
memory usage: 487.9+ KB
```

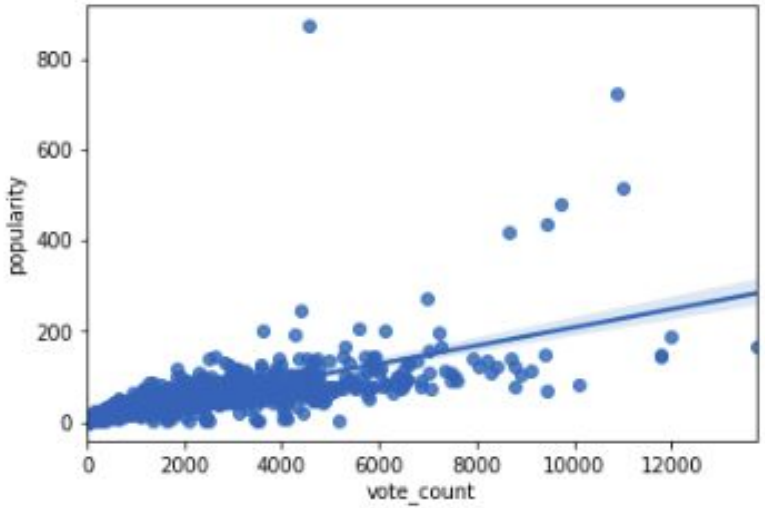
Data Exploration - corr, heatmap

	budget	id	popularity	revenue	r
budget	1.000000	-0.089377	0.505414	0.730823	0.
id	-0.089377	1.000000	0.031202	-0.050425	-0.
popularity	0.505414	0.031202	1.000000	0.644724	0.
revenue	0.730823	-0.050425	0.644724	1.000000	0.
runtime	0.269851	-0.153536	0.225502	0.251093	1.
vote_average	0.093146	-0.270595	0.273952	0.197150	0.
vote_count	0.593180	-0.004128	0.778130	0.781487	0.271944	0.312997	1.000000



Data Exploration - scatterplot





Data Exploration

```
tmdb_credits.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 4803 entries, 0 to 4802
Data columns (total 2 columns):
movie_id    4803 non-null int64
title       4803 non-null object
dtypes: int64(1), object(1)
memory usage: 112.6+ KB
```

Data Cleaning

Data Cleaning

genres	id	original_language	popularity	production_companies	production_countries	release_date	revenue	runtime	spoken_languages	vote_
[{"id": 28, "name": "Action"}, {"id": 12, "nam...	19995	en	150.437577	[{"name": "Ingenious Film Partners", "id": 289...	[{"iso_3166_1": "US", "name": "United States o...	2009-12-10	2787965087	162.0	[{"iso_639_1": "en", "name": "English"}, {"iso...	
[{"id": 12, "name": "Adventure"}, {"id": 14, "...	285	en	139.082615	[{"name": "Walt Disney Pictures", "id": 2}, {"...	[{"iso_3166_1": "US", "name": "United States o...	2007-05-19	961000000	169.0	[{"iso_639_1": "en", "name": "English"}]	

genres	id	original_language	popularity	production_companies	production_countries	release_date	revenue	runtime	spoken_languages	vote_a
[Action, Adventure, Fantasy, Science Fiction]	19995	en	150.437577	[Ingenious Film Partners, Twentieth Century Fo...	[United States of America, United Kingdom]	2009-12-10	2787965087	162.0	[English, Español]	
[Adventure, Fantasy, Action]	285	en	139.082615	[Walt Disney Pictures, Jerry Bruckheimer Films...	[United States of America]	2007-05-19	961000000	169.0	[English]	

Merging

Merging

```
tmdb_credits.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 4803 entries, 0 to 4802
Data columns (total 2 columns):
movie_id    4803 non-null int64
title       4803 non-null object
dtypes: int64(1), object(1)
memory usage: 112.6+ KB
```

```
tmdb_movies.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4803 entries, 0 to 4802
Data columns (total 13 columns):
budget                  4803 non-null int64
genres                  4803 non-null object
id                      4803 non-null int64
original_language       4803 non-null object
popularity              4803 non-null float64
production_companies    4803 non-null object
production_countries    4803 non-null object
release_date            4802 non-null object
revenue                 4803 non-null int64
runtime                 4801 non-null float64
spoken_languages        4803 non-null object
vote_average            4803 non-null float64
vote_count              4803 non-null int64
dtypes: float64(3), int64(4), object(6)
memory usage: 487.9+ KB
```

Questions and Answers

Questions

- Which movies have the highest vote_average?
- Which movies people voted the most?
- Which movie has the most budget?
- Which movie has the most popularity?
- Which production_companies have the most movies?
- Which production_companies have the highest budget?
- Which production_companies have the highest revenue?
- Which genres have the best popularity?
- Which spoken_languages have the highest popularity?

Questions

- Which movies have the highest vote_average?
- Which movies people voted the most?
- Which movie has the most budget?
- Which movie has the most popularity?
- Which production_companies have the most movies?

- ~~Which production_companies have the highest budget?~~
- ~~Which production_companies have the highest revenue?~~
- ~~Which genres have the best popularity?~~
- ~~Which spoken_languages have the highest popularity?~~

1. Which movies have the highest vote_a

	title	vote_average
3519	Stiff Upper Lips	10.0
4247	Me You and Five Bucks	10.0
4045	Dancer, Texas Pop. 81	10.0
4662	Little Big Top	10.0
3992	Sardaarji	9.5
2386	One Man's Hero	9.3
2970	There Goes My Baby	8.5
1881	The Shawshank Redemption	8.5
2796	The Prisoner of Zenda	8.4
3337	The Godfather	8.4

1. Which movies have the highest vote_average?

	title	vote_average	vote_count
3519	Stiff Upper Lips	10.0	1
4247	Me You and Five Bucks	10.0	2
4045	Dancer, Texas Pop. 81	10.0	1
4662	Little Big Top	10.0	1
3992	Sardaarji	9.5	2
2386	One Man's Hero	9.3	2
2970	There Goes My Baby	8.5	2
1881	The Shawshank Redemption	8.5	8205
2796	The Prisoner of Zenda	8.4	11
3337	The Godfather	8.4	5893

2. Which movies people voted the most?

	title	vote_count	vote_average
96	Inception	13752	8.1
65	The Dark Knight	12002	8.2
0	Avatar	11800	7.2
16	The Avengers	11776	7.4
788	Deadpool	10995	7.4
95	Interstellar	10867	8.1
287	Django Unchained	10099	7.8
94	Guardians of the Galaxy	9742	7.9
426	The Hunger Games	9455	6.9
127	Mad Max: Fury Road	9427	7.2

3. Which movie has the most budget?

	title	budget
17	Pirates of the Caribbean: On Stranger Tides	380000000
1	Pirates of the Caribbean: At World's End	300000000
7	Avengers: Age of Ultron	280000000
10	Superman Returns	270000000
4	John Carter	260000000
6	Tangled	260000000
5	Spider-Man 3	258000000
13	The Lone Ranger	255000000
46	X-Men: Days of Future Past	250000000
22	The Hobbit: The Desolation of Smaug	250000000

4. Which movie has the most popularity?

	title	popularity
546	Minions	875.581305
95	Interstellar	724.247784
788	Deadpool	514.569956
94	Guardians of the Galaxy	481.098624
127	Mad Max: Fury Road	434.278564
28	Jurassic World	418.708552
199	Pirates of the Caribbean: The Curse of the Bla...	271.972889
82	Dawn of the Planet of the Apes	243.791743
200	The Hunger Games: Mockingjay - Part 1	206.227151
88	Big Hero 6	203.734590

5. Which production_companies have the most movies?

	companies	count
4827	Warner Bros.	319
4680	Universal Pictures	311
3362	Paramount Pictures	285

Future work

Challenges and Learnings

Thank you!

repository:
gabrielatrindade/movies_tmdb_pyladies_bootcamp