Project: Investigate a Dataset (TMDb Movie Database)

Table of Contents

- Introduction
- Data Wrangling
- Exploratory Data Analysis
- Conclusions

Introduction

This dataset contains information about 10,000 movies collected from The Movie Database (TMDb), including user ratings and revenue. Dataset consistes of 21 columns and 10866 rows. Data types of columns: float64(4 columns), int64(6 columns), object(11 columns) There are 9512 unique movie titles, 4505 unique director names, and 7406 production companies. the database contains movies that were produced from 1960 to 2015.

Questions asked:

1. what is line of production by years?
2. Is Budget related to popularity?
3. What is most runtime desired?
4. Research Question 4 *what is the relationship between popularity and vote_average?
5. who are the most important directors that achieve the most popular movies?
6. what is leading production companies in the industry?
7. what genre is most produced?

Data Wrangling¶

General Properties

Database columns and Datatypes of columns:
RangeIndex: 10866 entries, 0 to 10865
Data columns (total 21 columns):

```
id                      10866 non-null int64
imdb_id                 10856 non-null object
popularity              10866 non-null float64
budget                  10866 non-null int64
revenue                 10866 non-null int64
original_title          10866 non-null object
cast                    10790 non-null object
homepage                 2936 non-null object
director                10822 non-null object
tagline                  8042 non-null object
keywords                 9373 non-null object
overview                10862 non-null object
runtime                 10866 non-null int64
genres                  10843 non-null object
production_companies     9836 non-null object
release_date            10866 non-null object
vote_count              10866 non-null int64
vote_average            10866 non-null float64
release_year            10866 non-null int64
budget_adj              10866 non-null float64
revenue_adj             10866 non-null float64
dtypes: float64(4), int64(6), object(11)
```

Unique Values:

```
id                      10865
imdb_id                 10855
popularity              10814
budget                    557
revenue                  4702
original_title          10571
cast                    10719
homepage                 2896
director                 5067
tagline                  7997
keywords                 8804
overview                10847
runtime                   247
genres                   2039
production_companies     7445
release_date             5909
```

```
vote_count          1289
vote_average          72
release_year          56
budget_adj          2614
revenue_adj         4840
```

Inofmation about numeric datatype columns:

	id	popularity	budget	revenue	runtime	vote_count	vote_average	release_year	budget_adj	revenue_adj
count	10866.000000	10866.000000	1.086600e+04	1.086600e+04	10866.000000	10866.000000	10866.000000	10866.000000	1.086600e+04	1.086600e+04
mean	66064.177434	0.646441	1.462570e+07	3.982332e+07	102.070863	217.389748	5.974922	2001.322658	1.755104e+07	5.136436e+07
std	92130.136561	1.000185	3.091321e+07	1.170035e+08	31.381405	575.619058	0.935142	12.812941	3.430616e+07	1.446325e+08
min	5.000000	0.000065	0.000000e+00	0.000000e+00	0.000000	10.000000	1.500000	1960.000000	0.000000e+00	0.000000e+00
25%	10596.250000	0.207583	0.000000e+00	0.000000e+00	90.000000	17.000000	5.400000	1995.000000	0.000000e+00	0.000000e+00
50%	20669.000000	0.383856	0.000000e+00	0.000000e+00	99.000000	38.000000	6.000000	2006.000000	0.000000e+00	0.000000e+00
75%	75610.000000	0.713817	1.500000e+07	2.400000e+07	111.000000	145.750000	6.600000	2011.000000	2.085325e+07	3.369710e+07
max	417859.000000	32.985763	4.250000e+08	2.781506e+09	900.000000	9767.000000	9.200000	2015.000000	4.250000e+08	2.827124e+09

- There are meaningless zero values in budjet, revenu and runtime
- There are some columns will not be used in analysis ('imdb_id','homepage','tagline','keywords','overview')
- There are one duplicated row.
- There are missing values in some columns

Data Cleaning

1.drop unnecessary columns

2.drop duplicated row

3.drop null values

4.drop rows with zero values in budjet, revenu and runtime.

Exploratory Data Analysis

The histograms of dataset columns:

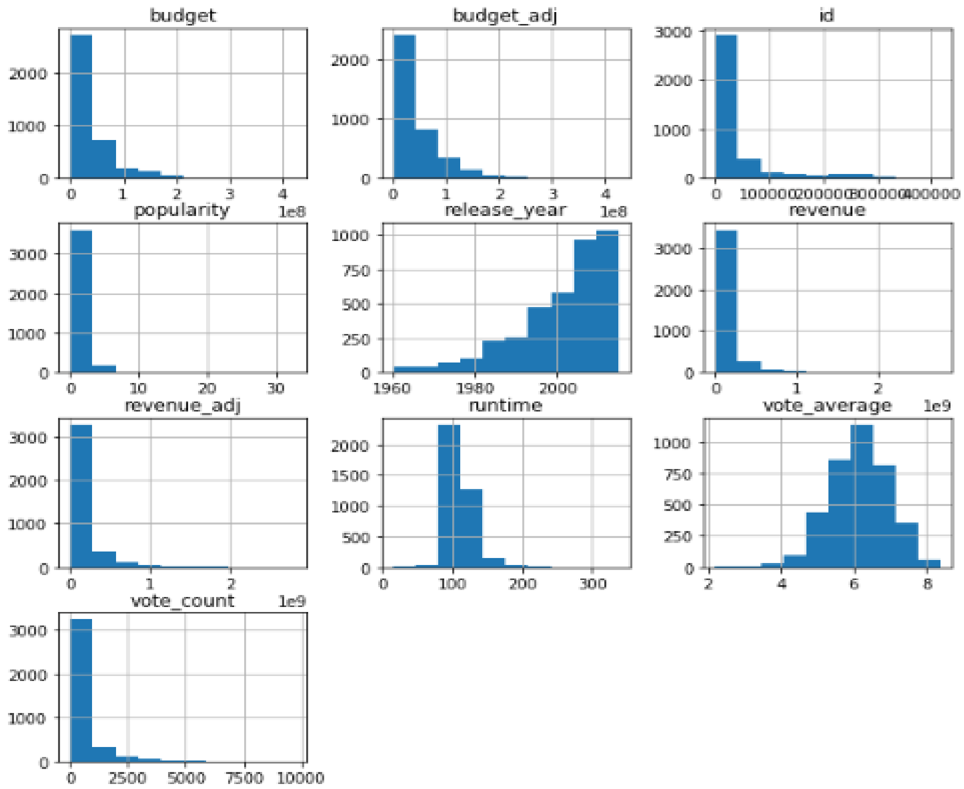


Figure 1 hoistogram of numeric dataset columns

Histogram shows the flowing:

1-budget histograms are left-skewed, which means most movies were produced with a low budget.

2-popularity histogram is left-skewed, which means most movies had low popularity.

3- relaese_year histogram is right-skewed which means most movies are produced in recent years.

4- revenue histograms are left-skewed which means most revenue of movies is low.

5- runtime histogram is left-skewed which means that most movies produced with runtime duration approximately in a range of 100 minutes.

Research Question 1 * what is line of production by years?

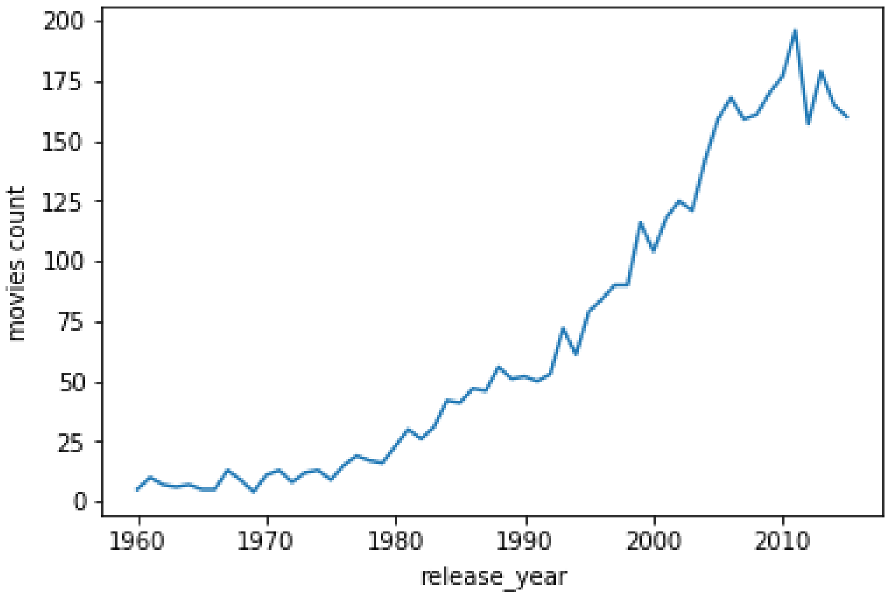


Figure 2 movies produced every year

The graph shows that movies production has been on a steady increase since 1960.

Research Question 2 *Is Budget related to popularity?

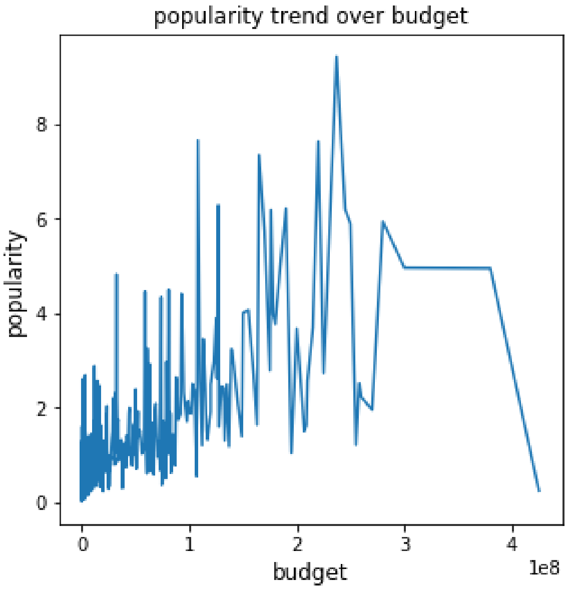


Figure 3 popularity trend over budget, line_ chart

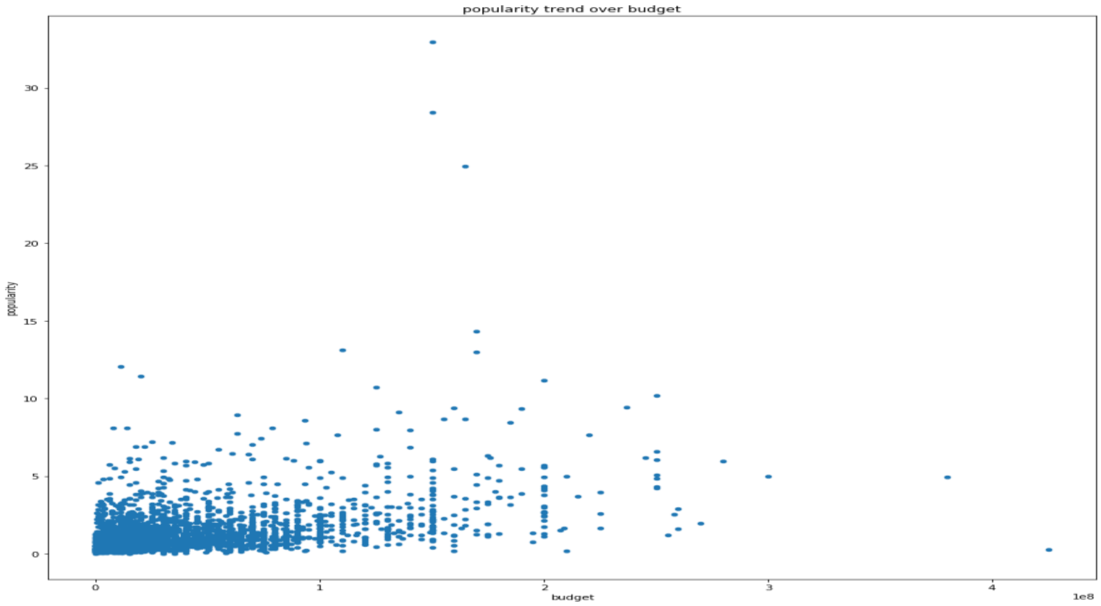


Figure 4 popularity trend over budget , scatter_chart

The graphs shows that there is no relationship between the popularity of movies and popularity, which means that a high budget is not necessary to make popular movies.

Research Question 3 What is most runtime desired?

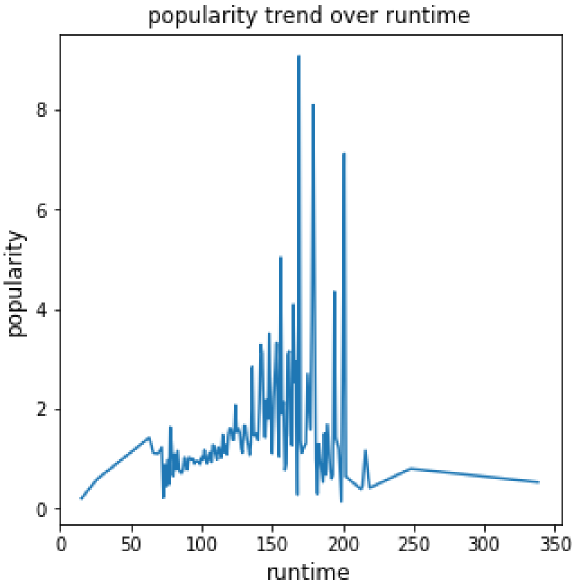


Figure 5 popularity trend over runtime

The chart shows that movies with a runtime between(150-200) have top popularity rates.

Research Question 4 *what is the relationship between popularity and vote_average?

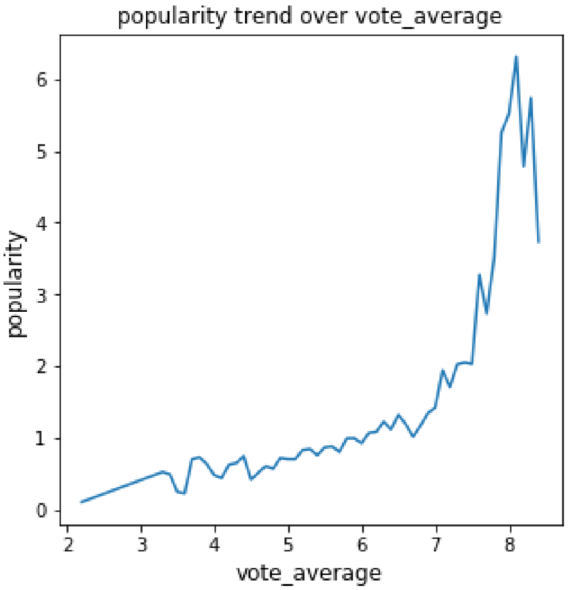
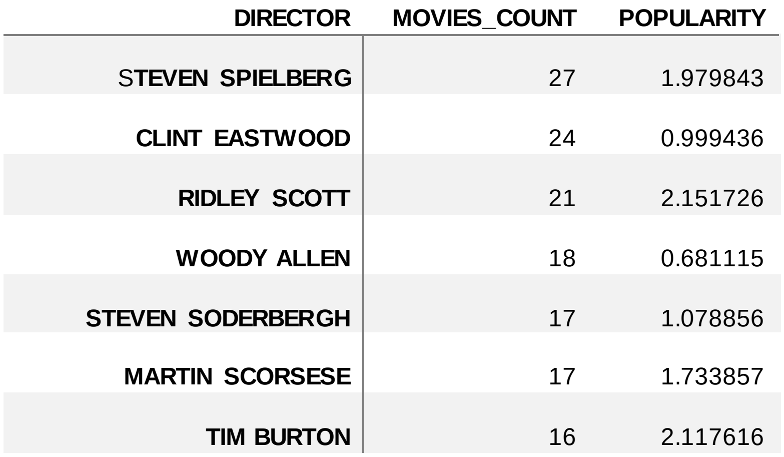


Figure 6 popularity trend over vote_average

The chart shows that the vote average rate increases with popularity increasing.

Research Question 5 *who are the most important directors that achieve the most popular movies?



DIRECTOR	MOVIES_COUNT	POPULARITY
STEVEN SPIELBERG	27	1.979843
CLINT EASTWOOD	24	0.999436
RIDLEY SCOTT	21	2.151726
WOODY ALLEN	18	0.681115
STEVEN SODERBERGH	17	1.078856
MARTIN SCORSESE	17	1.733857
TIM BURTON	16	2.117616

RENNY HARLIN	15	0.765630
BRIAN DE PALMA	15	1.012305
ROBERT ZEMECKIS	15	2.143790
OLIVER STONE	15	0.789349
RON HOWARD	14	1.699220
WES CRAVEN	14	0.651942
JOEL SCHUMACHER	14	1.009877
TONY SCOTT	14	1.097664
FRANCIS FORD COPPOLA	13	1.479083
RICHARD DONNER	13	1.147632
BARRY LEVINSON	12	0.732921
JOHN CARPENTER	12	0.846316
ROBERT RODRIGUEZ	12	0.740549
ROB REINER	12	1.042801
MICHAEL BAY	11	2.023170
PETER JACKSON	11	4.382200
RICHARD LINKLATER	11	0.980503
WALTER HILL	11	0.649902
KEVIN SMITH	11	0.905482
SHAWN LEVY	10	1.748588
SPIKE LEE	10	0.656607
ROMAN POLANSKI	10	1.041919
PAUL W.S. ANDERSON	10	1.318314
DAVID CRONENBERG	10	0.574948
BOBBY FARRELLY|PETER FARRELLY	10	1.050880
DAVID FINCHER	10	3.447978
SAM RAIMI	10	1.428279
JOHN LANDIS	10	0.862271

IVAN REITMAN	10	1.256998
MICHAEL MANN	10	1.124872
ROB COHEN	10	1.096025

Research Question 6 *what is leading production companies in the industry?

PRODUCTION COMPANY	MOVIES_COUNT	POPULARITY
PARAMOUNT PICTURES	77	0.873188
UNIVERSAL PICTURES	57	0.669668
COLUMBIA PICTURES	39	0.849457
NEW LINE CINEMA	38	0.816173
WARNER BROS.	33	0.829843
METRO-GOLDWYN-MAYER (MGM)	26	0.638193
TOUCHSTONE PICTURES	24	0.677981
TWENTIETH CENTURY FOX FILM CORPORATION	23	0.722115
WALT DISNEY PICTURES	22	1.322252
20TH CENTURY FOX	22	0.692946
MIRAMAX FILMS	17	0.666078
ORION PICTURES	17	0.625379
DIMENSION FILMS	16	0.642095
COLUMBIA PICTURES CORPORATION	16	0.703489
TRISTAR PICTURES	15	0.599546
UNITED ARTISTS	15	0.922827
DREAMWORKS ANIMATION	15	1.744172
WALT DISNEY PICTURES|PIXAR ANIMATION STUDIOS	13	3.336180
WALT DISNEY PICTURES|WALT DISNEY FEATURE ANIMATION	12	1.970178
IMAGINE ENTERTAINMENT|UNIVERSAL PICTURES	11	1.152110
EON PRODUCTIONS	10	2.186643

Research Question 7 *what genre is most produced?

GENRE	MOVIE_COUNT
DRAMA	243
COMEDY	230
DRAMA|ROMANCE	106
COMEDY|ROMANCE	103
COMEDY|DRAMA|ROMANCE	87
COMEDY|DRAMA	85
HORROR|THRILLER	80
HORROR	57
DRAMA|THRILLER	47
ACTION|THRILLER	39
CRIME|DRAMA|THRILLER	37
DRAMA|COMEDY	36
COMEDY|FAMILY	32
ACTION|CRIME|THRILLER	31
DRAMA|CRIME	27
DRAMA|HISTORY	27
CRIME|DRAMA	26
DRAMA|COMEDY|ROMANCE	25
THRILLER	25
ACTION|CRIME|DRAMA|THRILLER	25
ACTION|THRILLER|CRIME	22
ADVENTURE|ACTION|THRILLER	22
COMEDY|CRIME	22
HORROR|MYSTERY|THRILLER	21
ACTION	21
ACTION|COMEDY	20

Conclusions

The analysis concludes the follwing:

#1The movie production increases stedily since 1960, which means that movies industry is improving and a popular demand..

#2-there is no relationship between budget and popularity, which means high budjet is not the factor of a success movie.

#3- the vote average increases by popularity .

#4-the most popularity of movies goes to those with runtime between 150 and 200 minutes.

#5-The director who produced most movies is 'Steven Spielberg' and the mean popularity of his movies is not the top, that means there is no relationship between more movies production of director and popularity.

#6- Analysis shows the The production company which produced most movies is 'Paramount Pictures'

#7-MOst produced genres are Drama,Comedy,Drama|Romance,Comedy|Romance,Comedy|Drama|Romance,Comedy|Drama,Horror|Thriller,Horror