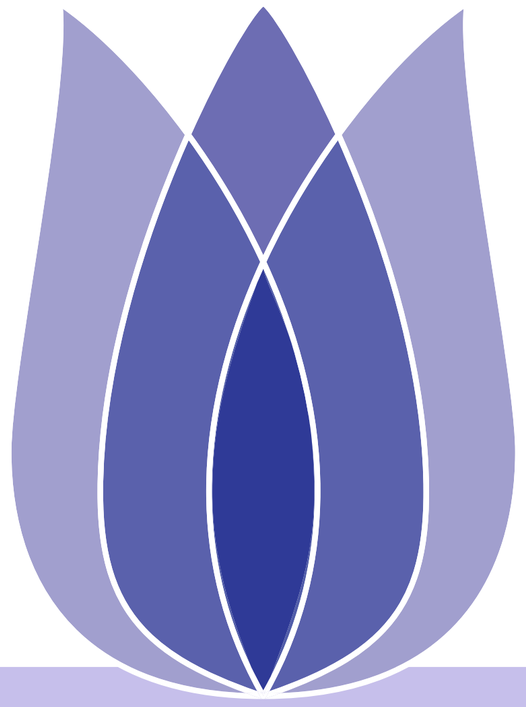
Kaggle Presentation

Huanhuan Ge

Qingdao University of Technology

2022-05-13



Overview

Problem

Data Processing

Data Analysis

Feature Selection And Model

Problem

- Description of TMDB Box Office Prediction

Data Processing

- Basic Information of Data
- Numerical features
- Missing Value
- Genres
- Release date

Data Analysis

- Budget Vs Revenue
- Popularity Vs Revenue
- Runtime Vs Revenue
- genres
- Year And Revenue
- Month And Week

Feature Selection And Model

- Feature Selection
- Model And Result

Problem
Description of TMDB Box Office Prediction
Data Processing
Data Analysis
Feature Selection And Model

Problem

Description of TMDB Box Office Prediction

Problem
Description of TMDB Box Office Prediction
Data Processing
Data Analysis
Feature Selection And Model

Description

In the dataset, it includes 7,398 movies and various metadata from the Movie Database (TMDB), Movies are labeled with id.Data points include cast,crew,plot keywords, budget, posters, release dates, languages, production companies, and countries.

Predict the worldwide revenue for 4398 movies.

Problem

Data Processing

Basic Information of Data

Numerical features

Missing Value

Genres

Release date

Data Analysis

Feature Selection And Model

Data Processing

Basic Information of Data

Problem

Data Processing

Basic Information of Data

Numerical features

Missing Value

Genres

Release date

Data Analysis

Feature Selection And Model

Table 1: Data

Name	Description	Attribute
train.csv	Training set(Movies from 1970-2018)	id,belongs_to_collection,budget,genres,homepage, imdb_id,original_language,original_title,overview,popularity, poster_path,production_companies,production_countries, release_date,runtime,spoken_languages,status,tagline, title,Keywords,cast,crew,revenue
test.csv	Test set(Predict revenue)	id,belongs_to_collection,budget,genres,homepage, imdb_id,original_language,original_title,overview,popularity, poster_path,production_companies,production_countries, release_date,runtime,spoken_languages,status,tagline, title,Keywords,cast,crew
sample_submission.csv	Format of submission	id,revenue

- There are 3000 samples in train set.

  There are 4398 samples in test set.

Problem

Data Processing

Basic Information of Data

Numerical features

Missing Value

Genres

Release date

Data Analysis

Feature Selection And Model

- There are 4 numerical features in total.
- The minimum of budget is 0.
- There are some missing values in the runtime, and the minimum of runtime is 0.

	id	budget	popularity	runtime	revenue
count	3000.000000	3.000000e+03	3000.000000	2998.000000	3.000000e+03
mean	1500.500000	2.253133e+07	8.463274	107.856571	6.672585e+07
std	866.169729	3.702609e+07	12.104000	22.086434	1.375323e+08
min	1.000000	0.000000e+00	0.000001	0.000000	1.000000e+00
25%	750.750000	0.000000e+00	4.018053	94.000000	2.379808e+06
50%	1500.500000	8.000000e+06	7.374861	104.000000	1.680707e+07
75%	2250.250000	2.900000e+07	10.890983	118.000000	6.891920e+07
max	3000.000000	3.800000e+08	294.337037	338.000000	1.519558e+09

Figure 1: Numerical features

Missing Value

Problem

Data Processing

Basic Information of Data

Numerical features

Missing Value

Genres

Release date

Data Analysis

Feature Selection And Model

- Remove columns that contain many null-valued features.
- Remove Some columns from which 'Prediction of Revenue' doesn't affect.

id	0
belongs_to_collection	2396
budget	0
genres	7
homepage	2054
imdb_id	0
original_language	0
original_title	0
overview	8
popularity	0
poster_path	1
production_companies	156
production_countries	55
release_date	0
runtime	2
spoken_languages	20
status	0
tagline	597
title	0
Keywords	276
cast	13
crew	16
revenue	0

dtype: int64

Figure 2: Missing Value

id	0.000000
belongs_to_collection	79.866667
budget	0.000000
genres	0.233333
homepage	68.466667
imdb_id	0.000000
original_language	0.000000
original_title	0.000000
overview	0.266667
popularity	0.000000
poster_path	0.033333
production_companies	5.200000
production_countries	1.833333
release_date	0.000000
runtime	0.066667
spoken_languages	0.666667
status	0.000000
tagline	19.900000
title	0.000000
Keywords	9.200000
cast	0.433333
crew	0.533333
revenue	0.000000

dtype: float64

Figure 3: Percentage of Missing Value

Genres

Problem

Data Processing

Basic Information of Data

Numerical features

Missing Value

Genres

Release date

Data Analysis

Feature Selection And Model

- Parse genres in type of JSON format.

	genres
0	[Comedy]
1	[Comedy, Drama, Family, Romance]
2	[Drama]
3	[Thriller, Drama]
4	[Action, Thriller]
...	...
2995	[Comedy, Romance]
2996	[Drama, Music]
2997	[Crime, Action, Mystery, Thriller]
2998	[Comedy, Romance]
2999	[Thriller, Action, Mystery]

3000 rows × 1 columns

Figure 4: genres

Release date

Problem

Data Processing

Basic Information of Data

Numerical features

Missing Value

Genres

Release date

Data Analysis

Feature Selection And Model

- Parse release date.

	release_month	release_day	release_year	day_of_Week
0	2	20	2015	4
1	8	6	2004	4
2	10	10	2014	4
3	3	9	2012	4
4	2	5	2009	3
...	...	...	...	...
2995	4	22	1994	4
2996	3	28	2013	3
2997	10	11	1996	4
2998	1	16	2004	4
2999	9	22	2011	3

Problem

Data Processing

Data Analysis

Budget Vs Revenue

Popularity Vs Revenue

Runtime Vs Revenue

genres

Year And Revenue

Month And Week

Feature Selection And Model

Data Analysis

Budget Vs Revenue

Problem

Data Processing

Data Analysis

Budget Vs Revenue

Popularity Vs Revenue

Runtime Vs Revenue

genres

Year And Revenue

Month And Week

Feature Selection And Model

	id	title	budget	revenue
1126	1127	The Avengers	220000000	1519557910
1761	1762	Furious 7	190000000	1506249360
2770	2771	Avengers: Age of Ultron	280000000	1405403694
684	685	Beauty and the Beast	160000000	1262886337
2322	2323	Transformers: Dark of the Moon	195000000	1123746996
906	907	The Dark Knight Rises	250000000	1084939099
2135	2136	Pirates of the Caribbean: On Stranger Tides	380000000	1045713802
2562	2563	Finding Dory	200000000	1028570889
881	882	Alice in Wonderland	200000000	1025491110
734	735	Zootopia	150000000	1023784195

Figure 5: Budget And Revenue

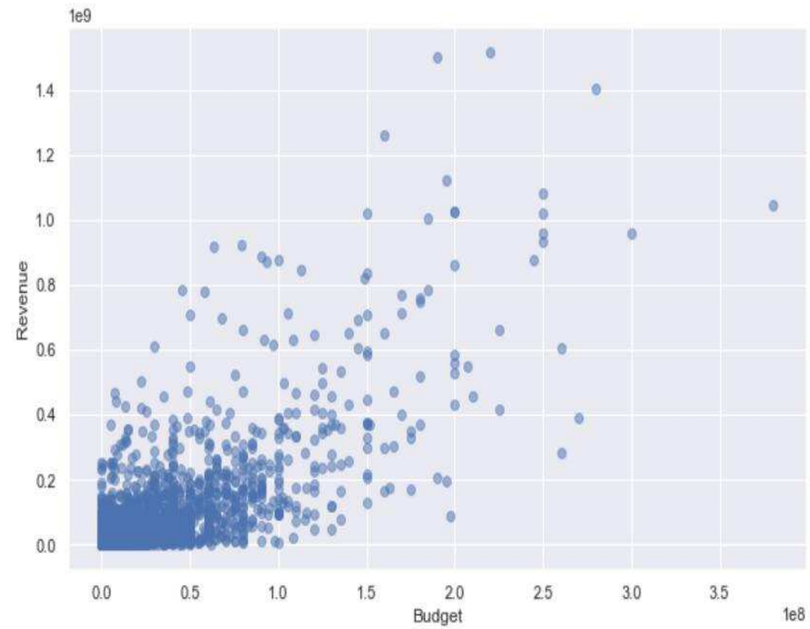


Figure 6: Budget And Revenue Scatter Plot

Popularity Vs Revenue

Problem

Data Processing

Data Analysis

Budget Vs Revenue

Popularity Vs Revenue

Runtime Vs Revenue

genres

Year And Revenue

Month And Week

Feature Selection And Model

	id	title	popularity	revenue
1126	1127	The Avengers	89.887648	1519557910
1761	1762	Furious 7	27.275687	1506249360
2770	2771	Avengers: Age of Ultron	37.379420	1405403694
684	685	Beauty and the Beast	287.253654	1262886337
2322	2323	Transformers: Dark of the Moon	4.503505	1123746996
906	907	The Dark Knight Rises	20.582580	1084939099
2135	2136	Pirates of the Caribbean: On Stranger Tides	27.887720	1045713802
2562	2563	Finding Dory	14.477677	1028570889
881	882	Alice in Wonderland	17.285093	1025491110
734	735	Zootopia	26.024868	1023784195

Figure 7: Popularity And Revenue

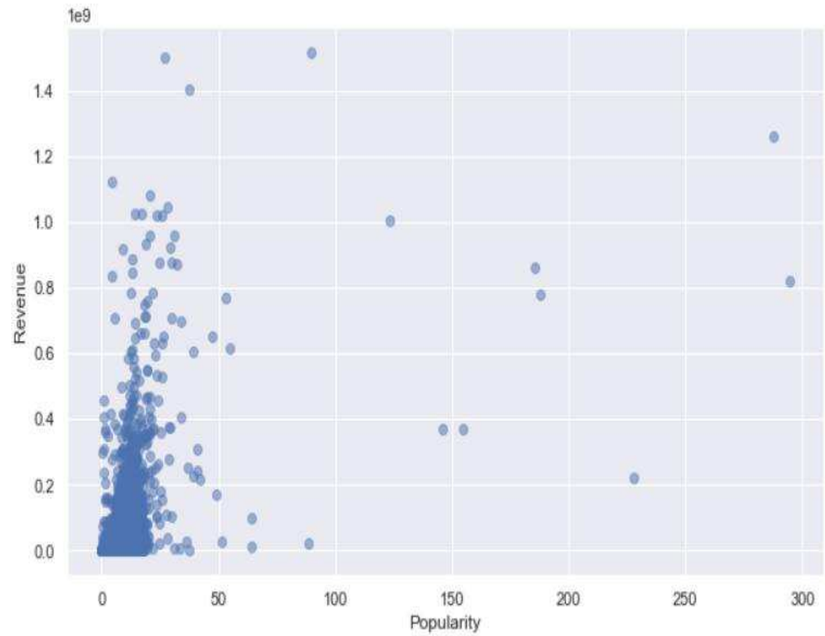


Figure 8: Popularity And Revenue Scatter Plot

Runtime Vs Revenue

Problem

Data Processing

Data Analysis

Budget Vs Revenue

Popularity Vs Revenue

Runtime Vs Revenue

genres

Year And Revenue

Month And Week

Feature Selection And Model

- Most movies are around two hours long.
- View in reverse order of runtime.

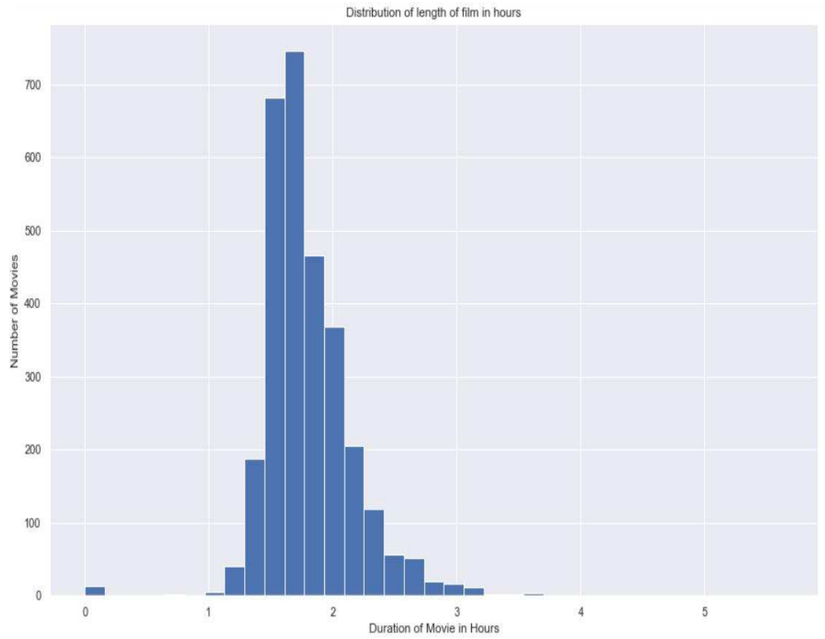


Figure 9: Runtime Histogram

	id	title	runtime	budget	revenue
1211	1212	Carlos	338.000000	18000000	871279
1922	1923	Cleopatra	248.000000	31115000	71000000
523	524	The Ten Commandments	220.000000	13000000	122700000
1302	1303	Heaven's Gate	219.000000	44000000	3484331
1914	1915	Gods and Generals	214.000000	56000000	12923936
2353	2354	Jodhaa Akbar	213.000000	8376800	13000000
625	626	Ben-Hur	212.000000	15000000	146900000
1975	1976	Chapiteau-Show	207.000000	2000000	393816
1731	1732	Hey Ram	199.000000	3900000	4900000
2120	2121	Spartacus	197.000000	12000000	60000000

Figure 10: Runtime And Revenue

genres

Problem

Data Processing

Data Analysis

Budget Vs Revenue

Popularity Vs Revenue

Runtime Vs Revenue

genres

Year And Revenue

Month And Week

Feature Selection And Model

- Drama and comedy dominate.

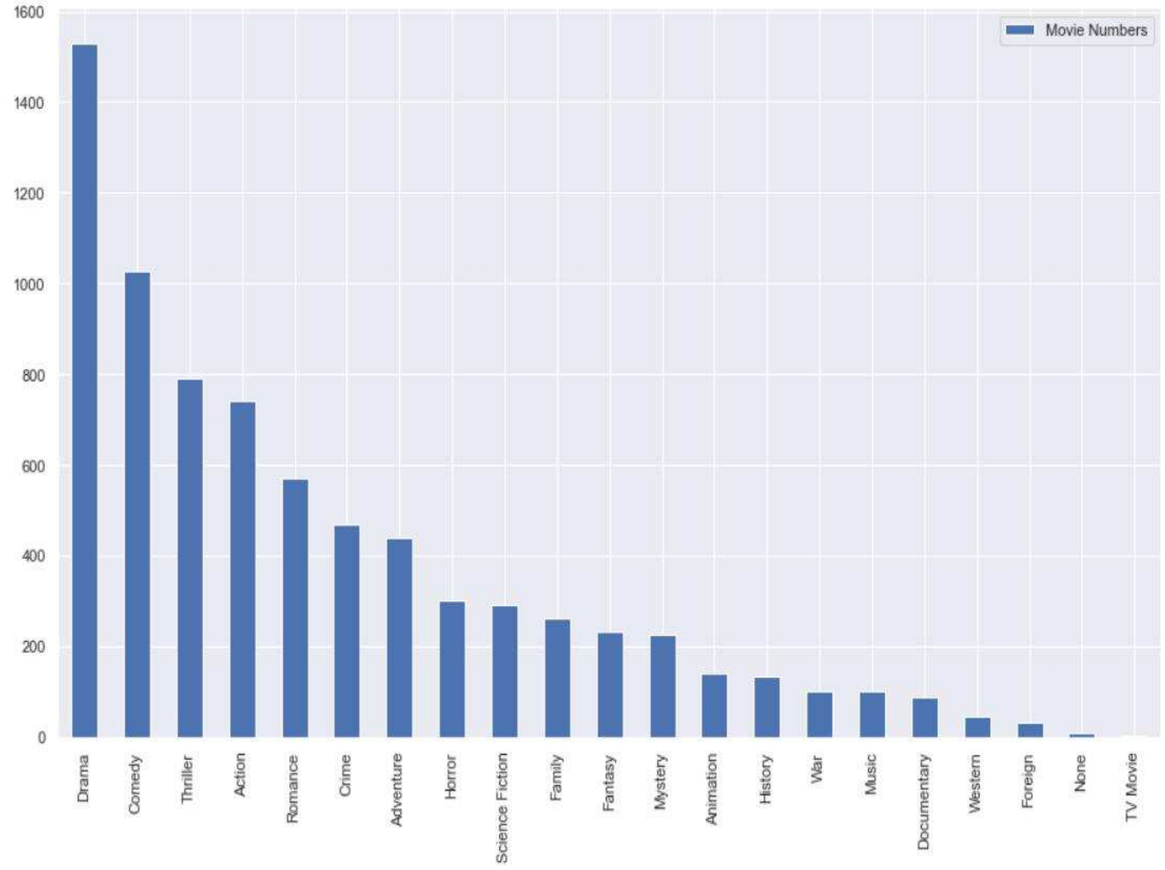


Figure 11: genres bar plot

Year And Revenue

Problem

Data Processing

Data Analysis

Budget Vs Revenue

Popularity Vs Revenue

Runtime Vs Revenue

genres

Year And Revenue

Month And Week

Feature Selection And Model

- More and more movies have been released in recent years.
- There may be a positive correlation between Year and Revenue.

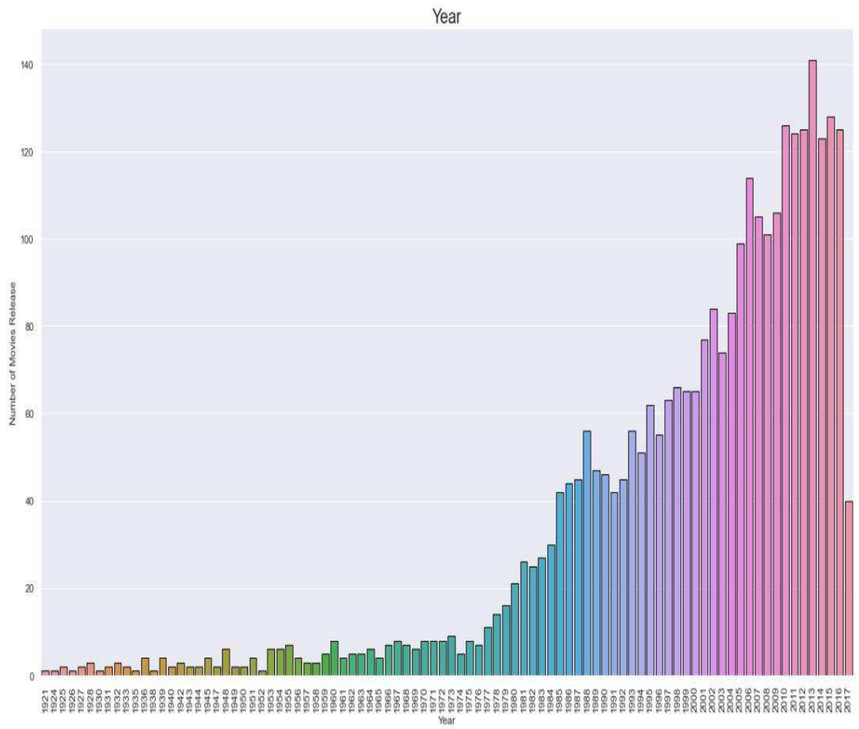


Figure 12: Histogram

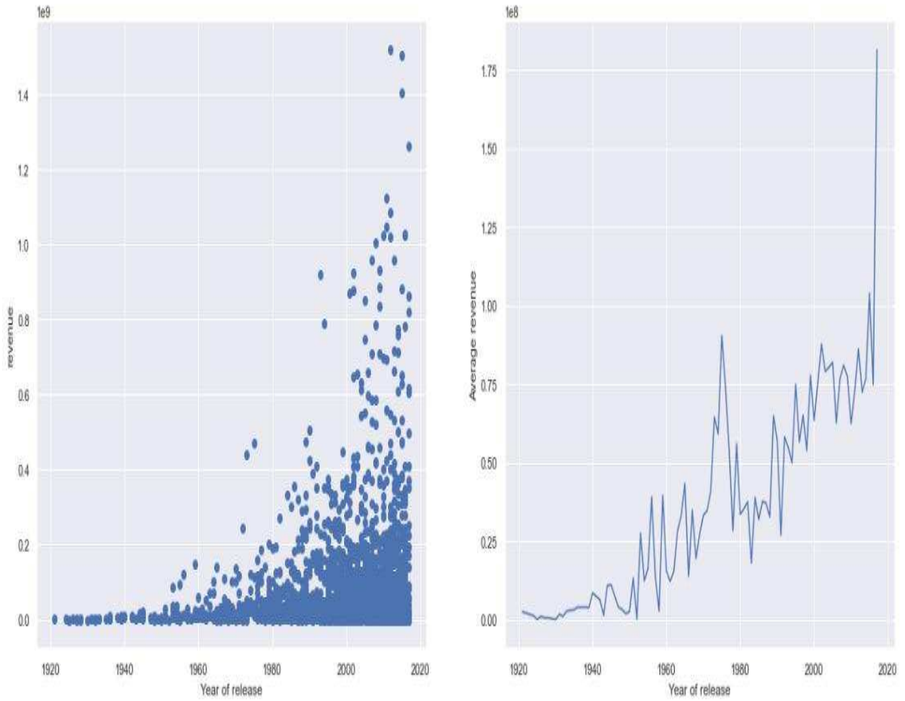


Figure 13: Scatter Plot

Month And Week

Problem

Data Processing

Data Analysis

Budget Vs Revenue

Popularity Vs Revenue

Runtime Vs Revenue

genres

Year And Revenue

Month And Week

Feature Selection And Model

- The number of movies released in September and October is higher.
- The number of films released on Friday accounted for the majority.

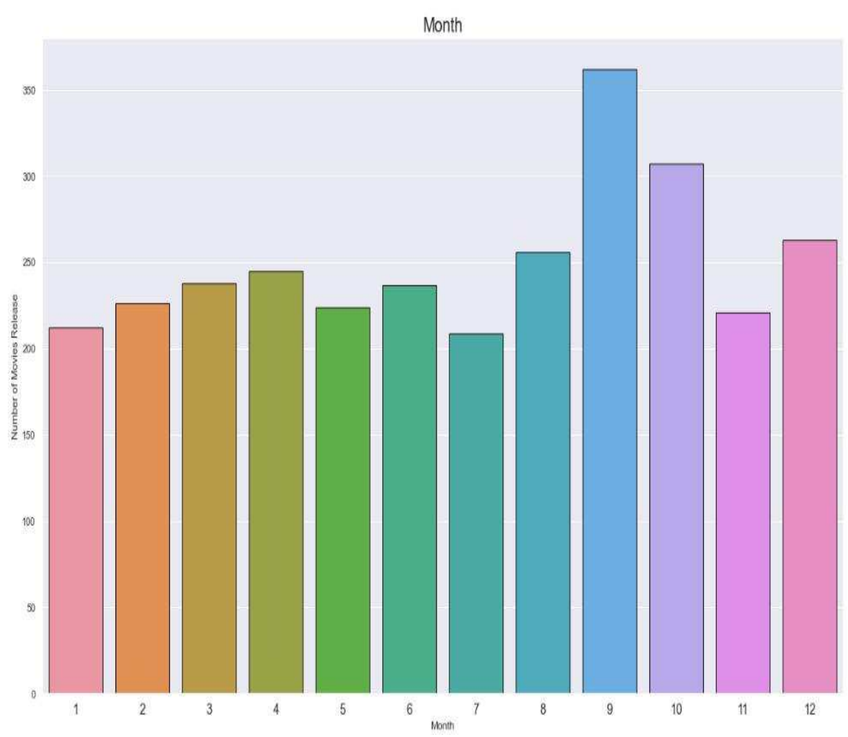


Figure 14: Month

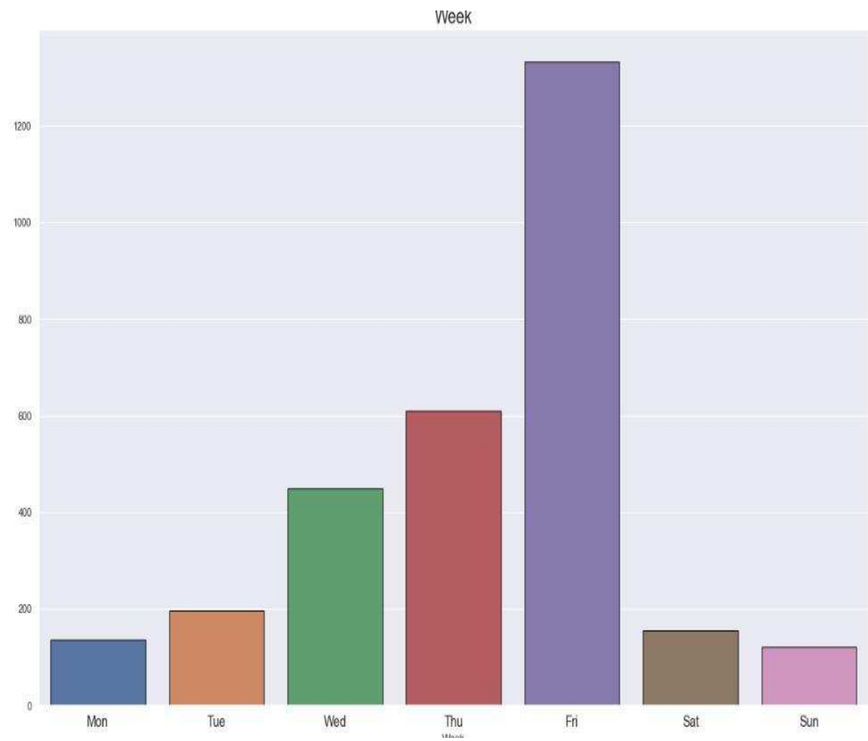


Figure 15: Week

Problem

Data Processing

Data Analysis

Feature Selection And Model

Feature Selection

Model And Result

Feature Selection And Model

Feature Selection

Problem

Data Processing

Data Analysis

Feature Selection And Model

Feature Selection

Model And Result

- release_data: release_year, release_day, release_month
- original_language,number_companies,crew_size
- budget,popularity, runtime

	release_year	release_day	release_month	original_language	budget	popularity	runtime	number_companies	crew_size
0	2015	20	2	7	14000000	6.575393	33	3	72
1	2004	6	8	7	40000000	8.248895	53	1	9
2	2014	10	10	7	3300000	64.299990	45	3	64
3	2012	9	3	13	1200000	3.174936	62	0	3
4	2009	5	2	18	0	1.148070	58	0	2
...	...	...	...	...	...	...	...	...	...
2995	1994	22	4	7	0	9.853270	42	2	17
2996	2013	28	3	29	0	3.727996	42	2	15
2997	1996	11	10	7	65000000	14.482345	60	3	10
2998	2004	16	1	7	42000000	15.725542	30	2	89
2999	2011	22	9	7	35000000	10.512109	46	6	48

Figure 16: Feature

Model And Result

Problem

Data Processing

Data Analysis

Feature Selection And Model

Feature Selection

Model And Result

- Random Forest
  - A random forest is a classifier that uses multiple trees to train and predict samples.
- RMSE
  - The root mean square error is the square root of the ratio of the deviation between the predicted value and the true value and the number of observations n,which is used to measure the deviation between the observed value and the true value.
- Train set: 0.8 , Test set: 0.2
- Score:1.19