Movies
Investment

Agenda

- Problem description
- Project pipeline
- Data set
- Data visualization and insights
- Preprocessing and feature engineering
- Model building and Training
- Evaluation and accuracy
- Unsuccessful trials

1 Problem description

Why we choose movies ??

movies made

42,500,000,000

IN 2019

Problem description

- Movies industry is growing
- More people are investing
- many risks in this industry and serious problem in case of movie failure

Problem description

- Predict how much revenue a movie will make
- Study factor that affect revenue
- How to maximize revenue

2 Project pipeline

visual board shows stages of project

Project pipeline

Visualization and insights

- Visualizing the data sets and its structure.
- Gettin data insights that are useful for the problem solution

Pre-processing and Feature engineering

- Fix the data formatting and structure for feature extraction
- Manipulate the data to add and transform features

Model building

- Using the extracted features, iteratively build models that attemt to solve our problem
- go back to the feature engineering step with feedback

3 Data set

Collection of data

Data set

- ‘TMDB movies dataset’ by TMDB website.
- Contains movies from 1921 to 2017 for the train dataset.
- contains movies from 1922 to 2018 for the test dataset.

```
RangeIndex: 3000 entries, 0 to 2999
Data columns (total 23 columns):
id                       3000 non-null
belongs_to_collection    604 non-null
budget                   3000 non-null
genres                   2993 non-null
homepage                 946 non-null
imdb_id                  3000 non-null
original_language        3000 non-null
original_title           3000 non-null
overview                 2992 non-null
popularity               3000 non-null
poster_path              2999 non-null
production_companies     2844 non-null
production_countries     2945 non-null
release_date             3000 non-null
runtime                  2998 non-null
spoken_languages         2980 non-null
status                   3000 non-null
tagline                  2403 non-null
title                    3000 non-null
Keywords                 2724 non-null
cast                     2987 non-null
crew                     2984 non-null
revenue                  3000 non-null
```

First 5 records

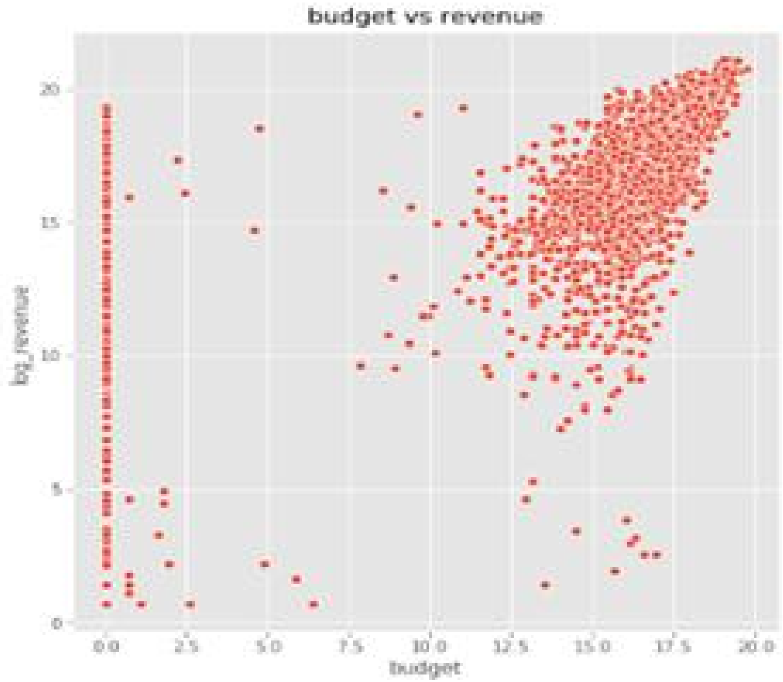
	id	belongs_to_collection	budget	genres	homepage	imdb_id	original_language	original_title	overview	popularity
0	1	[{'id': 313576, 'name': 'Hot Tub Time Machine ...	14000000	[{'id': 35, 'name': 'Comedy'}]	NaN	tt2637294	en	Hot Tub Time Machine 2	When Lou, who has become the "father of the In...	6.575393
1	2	[{'id': 107674, 'name': 'The Princess Diaries ...	40000000	[{'id': 35, 'name': 'Comedy'}, {'id': 18, 'nam...	NaN	tt0368933	en	The Princess Diaries 2: Royal Engagement	Mia Thermopolis is now a college graduate and ...	8.248895
2	3	NaN	3300000	[{'id': 18, 'name': 'Drama'}]	http://sonyclassics.com/whiplash/	tt2582802	en	Whiplash	Under the direction of a ruthless instructor, ...	64.299990
3	4	NaN	1200000	[{'id': 53, 'name': 'Thriller'}, {'id': 18, 'n...	http://kahaanithefilm.com/	tt1821480	hi	Kahaani	Vidya Bagchi (Vidya Balan) arrives in Kolkata ...	3.174936
4	5	NaN	0	[{'id': 28, 'name': 'Action'}, {'id': 53, 'nam...	NaN	tt1380152	ko	마린보이	Marine Boy is the story of a former national s...	1.148070

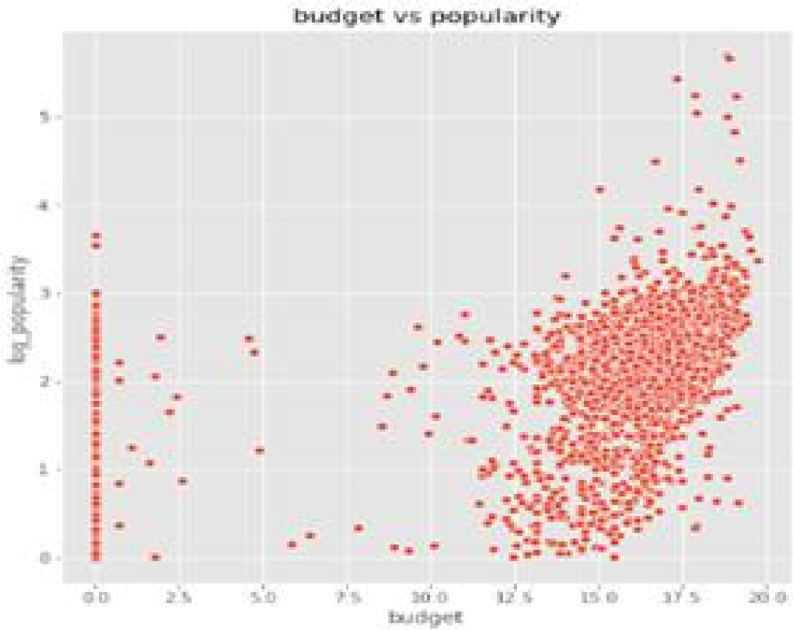
poster_path	production_companies	production_countries	release_date	runtime	spoken_languages	status	tagline	title	Keywords	cast	crew	revenue
/tQtWuwvMf0hCc2QR2tkolwl7c3c.jpg	[{'name': 'Paramount Pictures', 'id': 4}, {'na...	[{'iso_3166_1': 'US', 'name': 'United States o...	2/20/15	93.0	[{'iso_639_1': 'en', 'name': 'English'}]	Released	The Laws of Space and Time are About to be Vio...	Hot Tub Time Machine 2	[{'id': 4379, 'name': 'time travel'}, {'id': 9...	[{'cast_id': 4, 'character': 'Lou', 'credit_id...	[{'credit_id': '59ac067c92514107af02c8c8', 'de...	12314651
w9Z7A0GHEhIp7etpj0vyKOeU1Wx.jpg	[{'name': 'Walt Disney Pictures', 'id': 2}]	[{'iso_3166_1': 'US', 'name': 'United States o...	8/6/04	113.0	[{'iso_639_1': 'en', 'name': 'English'}]	Released	It can take a lifetime to find true love; she'...	The Princess Diaries 2: Royal Engagement	[{'id': 2505, 'name': 'coronation'}, {'id': 42...	[{'cast_id': 1, 'character': 'Mia Thermopolis'...	[{'credit_id': '52fe43fe9251416c7502563d', 'de...	95149435
/lIv1QinFqz4dlp5U4lQ6HaiskOZ.jpg	[{'name': 'Bold Films', 'id': 2266}, {'name': ...	[{'iso_3166_1': 'US', 'name': 'United States o...	10/10/14	105.0	[{'iso_639_1': 'en', 'name': 'English'}]	Released	The road to greatness can take you to the edge.	Whiplash	[{'id': 1416, 'name': 'jazz'}, {'id': 1523, 'n...	[{'cast_id': 5, 'character': 'Andrew Neimann',...	[{'credit_id': '54d5356ec3a3683ba0000039', 'de...	13092000
/aTXRaPrWSinhcmCrcfJK17urp3F.jpg	NaN	[{'iso_3166_1': 'IN', 'name': 'India'}]	3/9/12	122.0	[{'iso_639_1': 'en', 'name': 'English'}, {'iso...	Released	NaN	Kahaani	[{'id': 10092, 'name': 'mystery'}, {'id': 1054...	[{'cast_id': 1, 'character': 'Vidya Bagchi', '...	[{'credit_id': '52fe48779251416c9108d6eb', 'de...	16000000
/m22s7zvkVFDU9ir56PiiqIEWFdT.jpg	NaN	[{'iso_3166_1': 'KR', 'name': 'South Korea'}]	2/5/09	118.0	[{'iso_639_1': 'ko', 'name': '한국어/조선말'}]	Released	NaN	Marine Boy	NaN	[{'cast_id': 3, 'character': 'Chun-soo', 'cred...	[{'credit_id': '52fe464b9251416c75073b43', 'de...	3923970

4 Data visualization and Insights

What the data tell us?

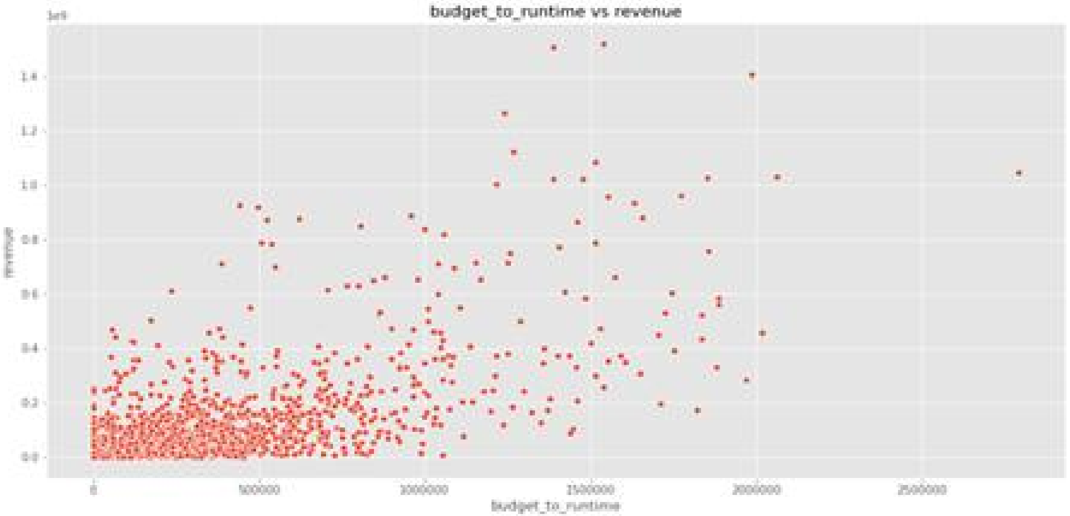
Relationship between Budget and Revenue



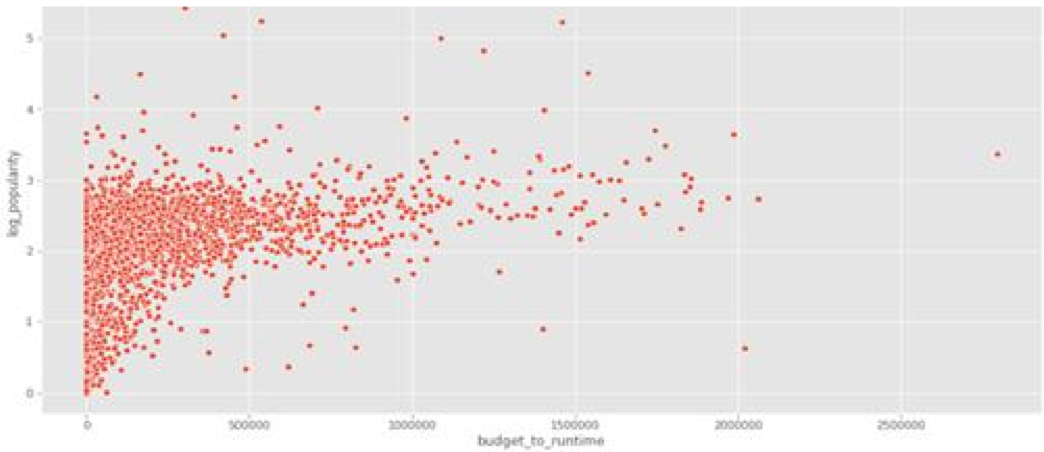


Relationship between Budget and Popularity

Relationship between Budget-to-runtime and Revenue

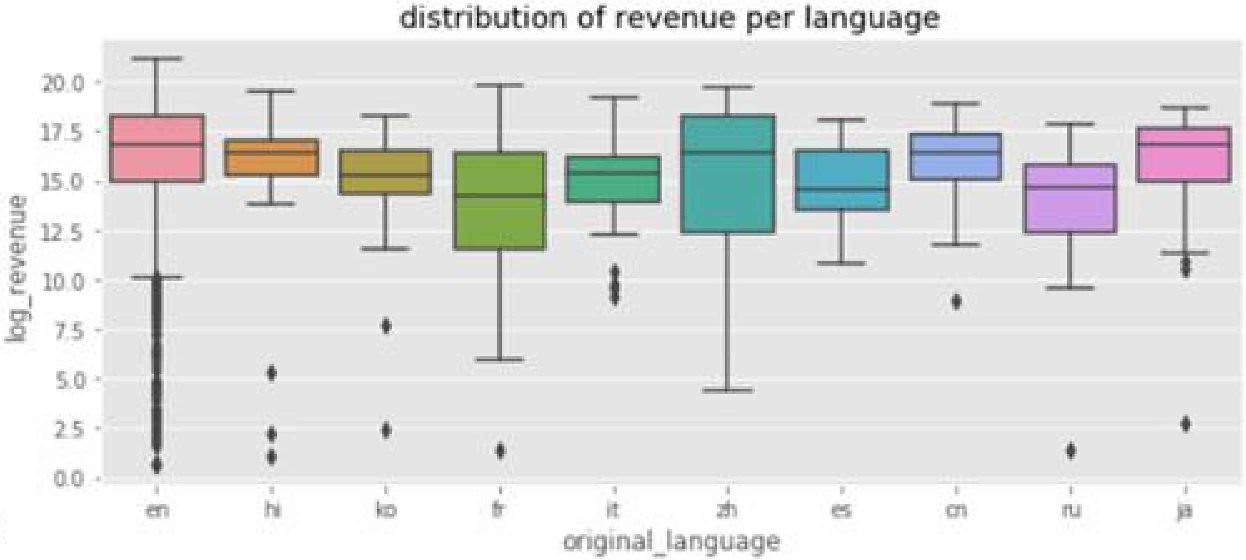


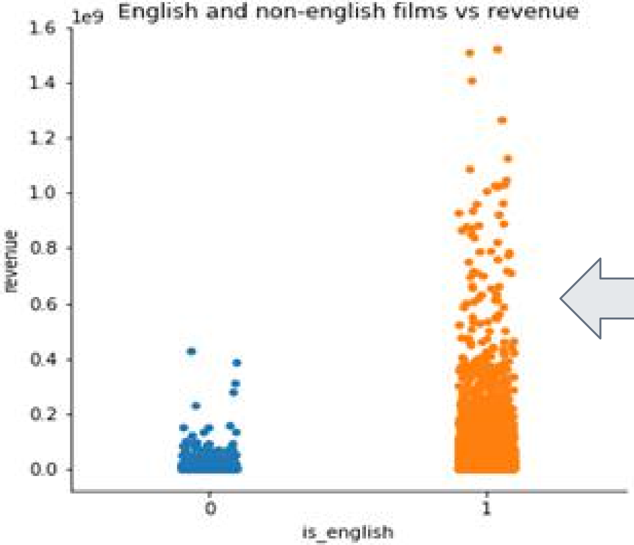
Relationship between Budget-to-runtime and Popularity



Visualizing movies languages

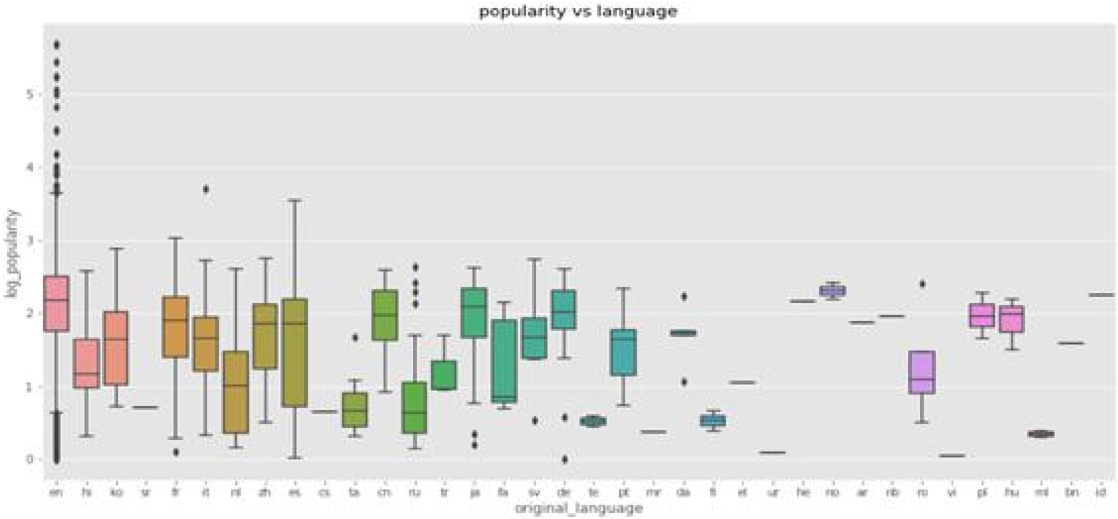
Revenue per language

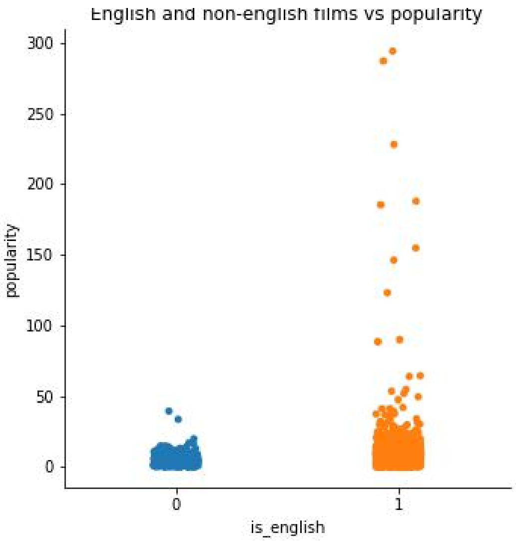




English and non-english movies revenue

Popularity per language

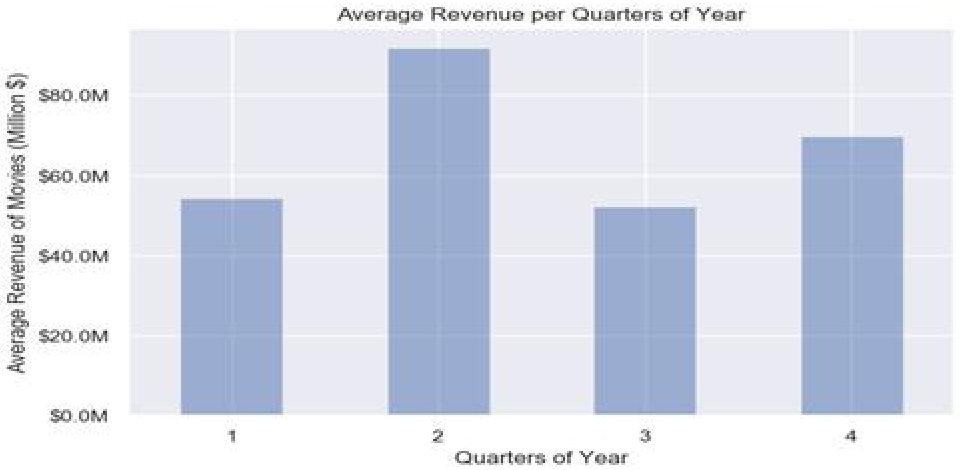




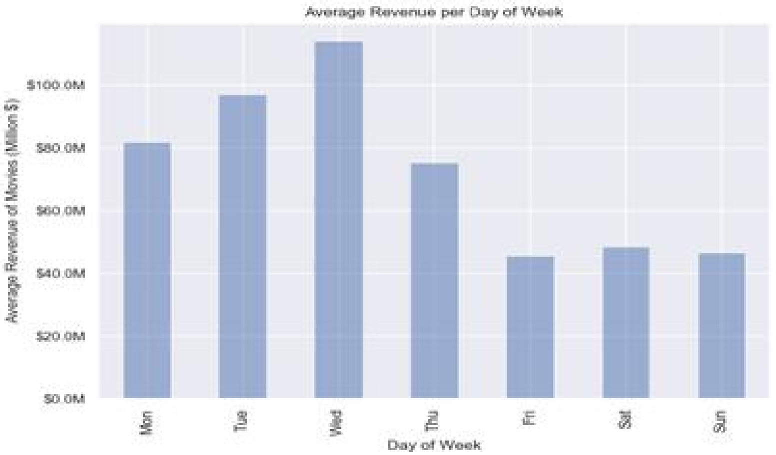
English and non-english movies popularity

Visualizing movies release dates

Average revenue of all movies per year

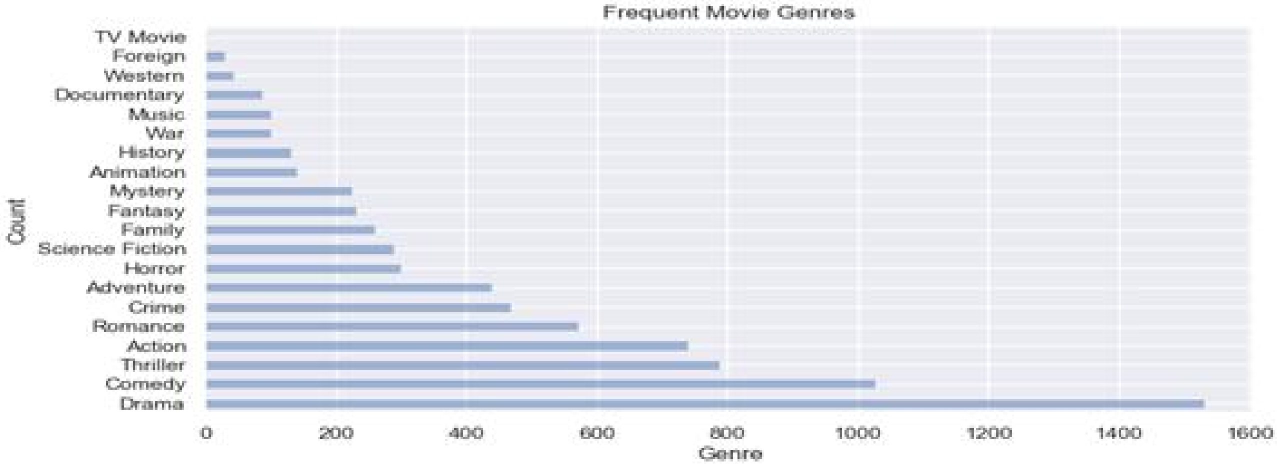

Average Revenue per Quarters of Year
Average Revenue of Movies (Million $)
$80.0M
$60.0M
$40.0M
$20.0M
$0.0M
1
2
3
4
Quarters of Year

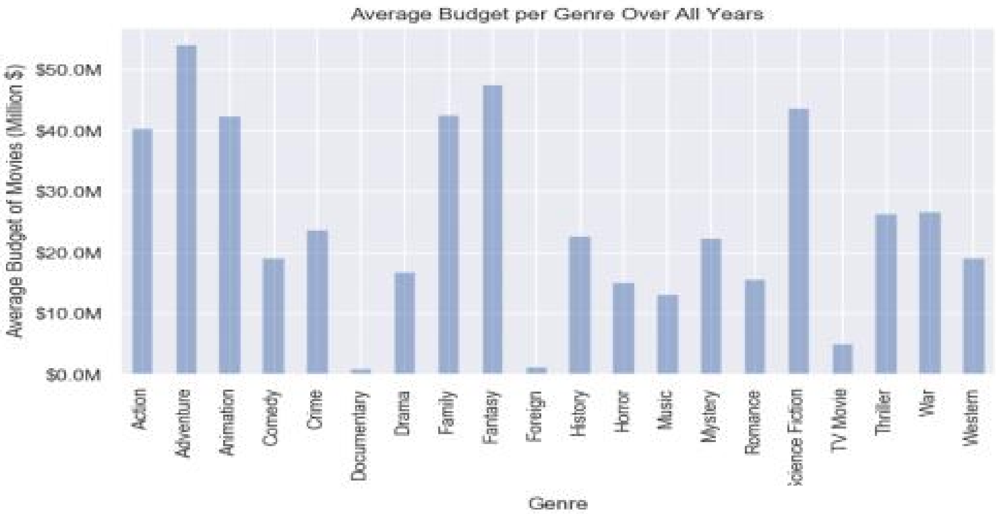

Average Revenue per Months of Year
Average Revenue of Movies (Million $)
$120.0M
$100.0M
$80.0M
$60.0M
$40.0M
$20.0M
$0.0M
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Months of Year


Average Revenue per Day of Week
Average Revenue of Movies (Million $)
$100.0M
$80.0M
$60.0M
$40.0M
$20.0M
$0.0M
Mon
Tue
Wed
Thu
Fri
Sat
Sun
Day of Week

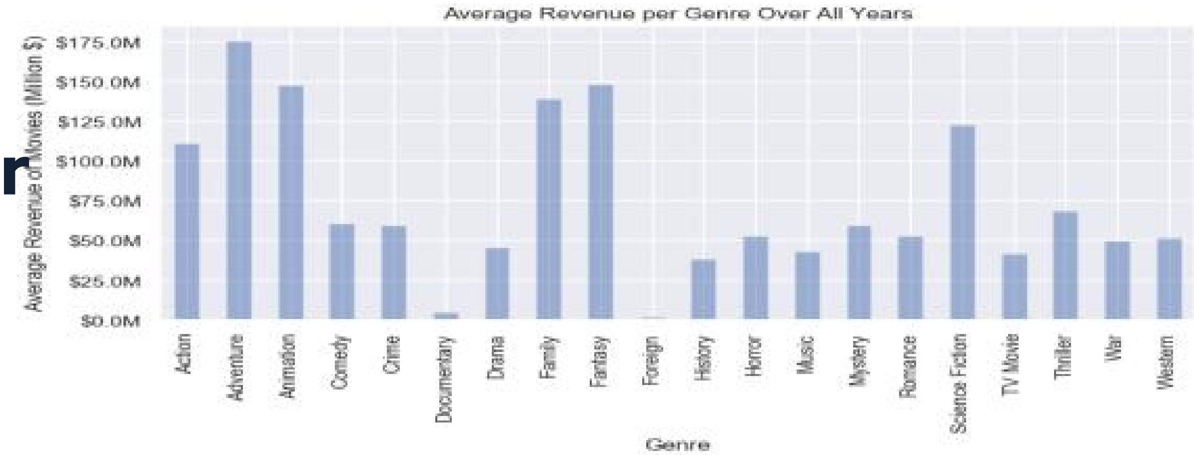

Visualizing movies genres

Movies genres frequency

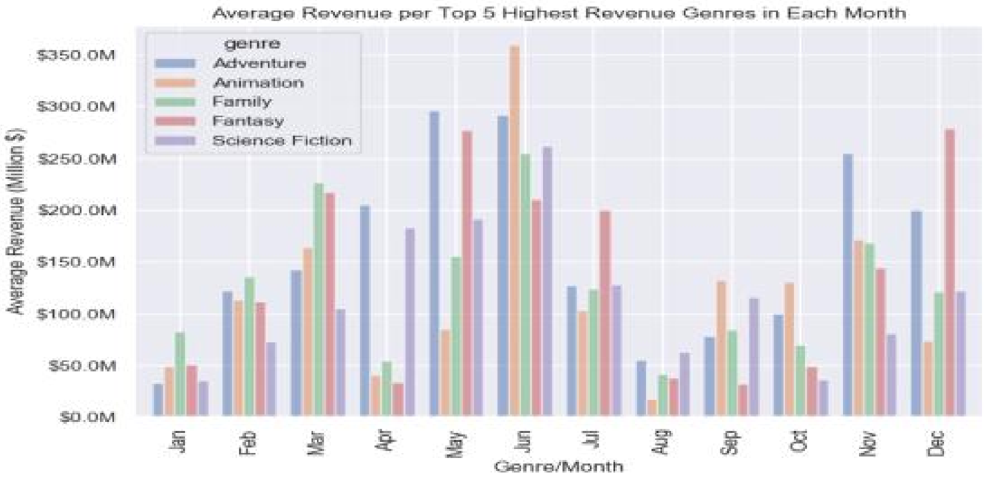




Average **revenue** per movie **genre**

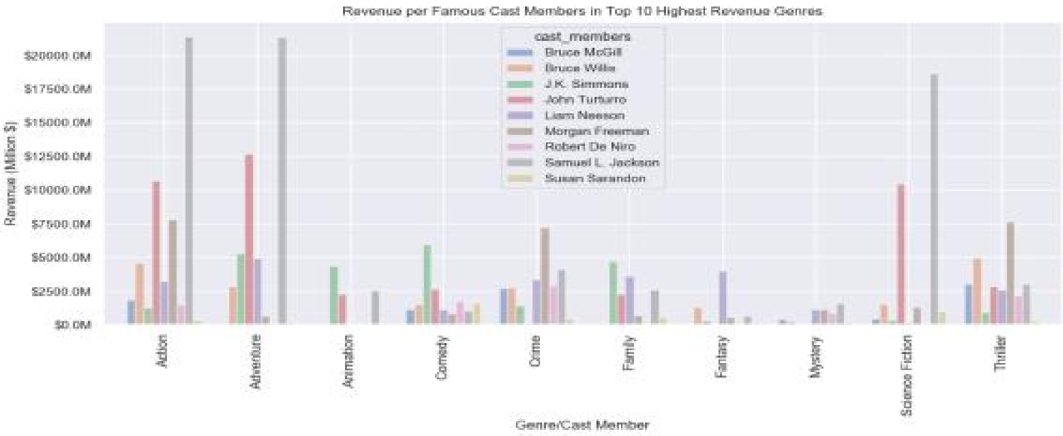


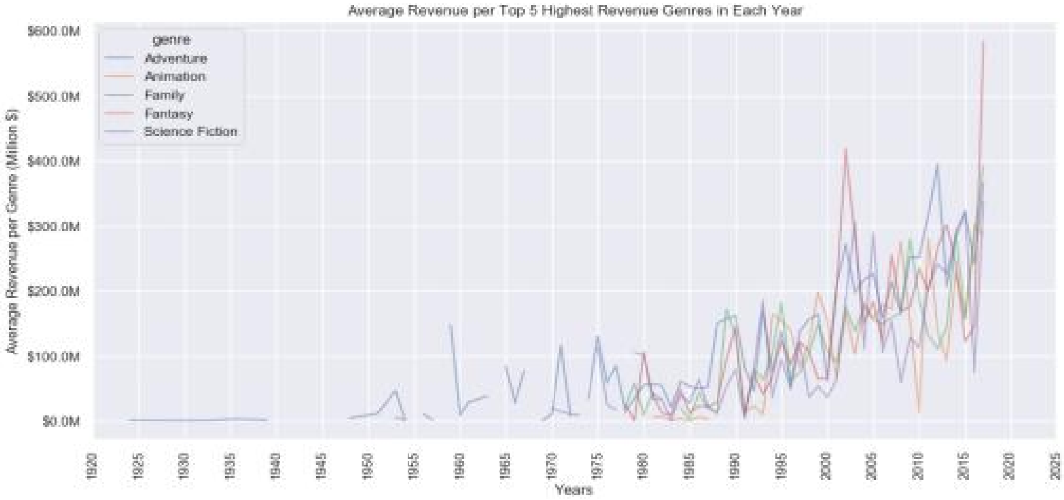
Average **revenue** per movie **genre**



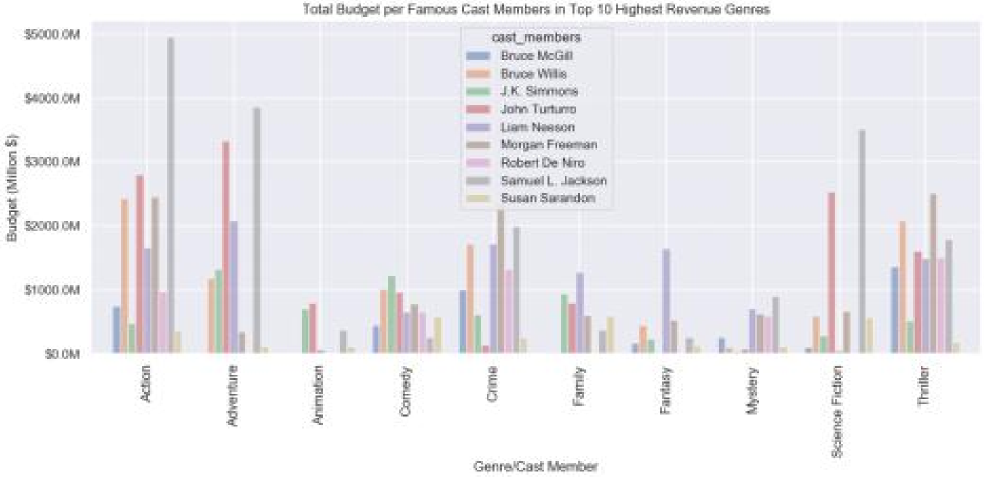
Total revenue of famous cast in top 10 highest revenue genres

Top 5 monthly revenue per genre



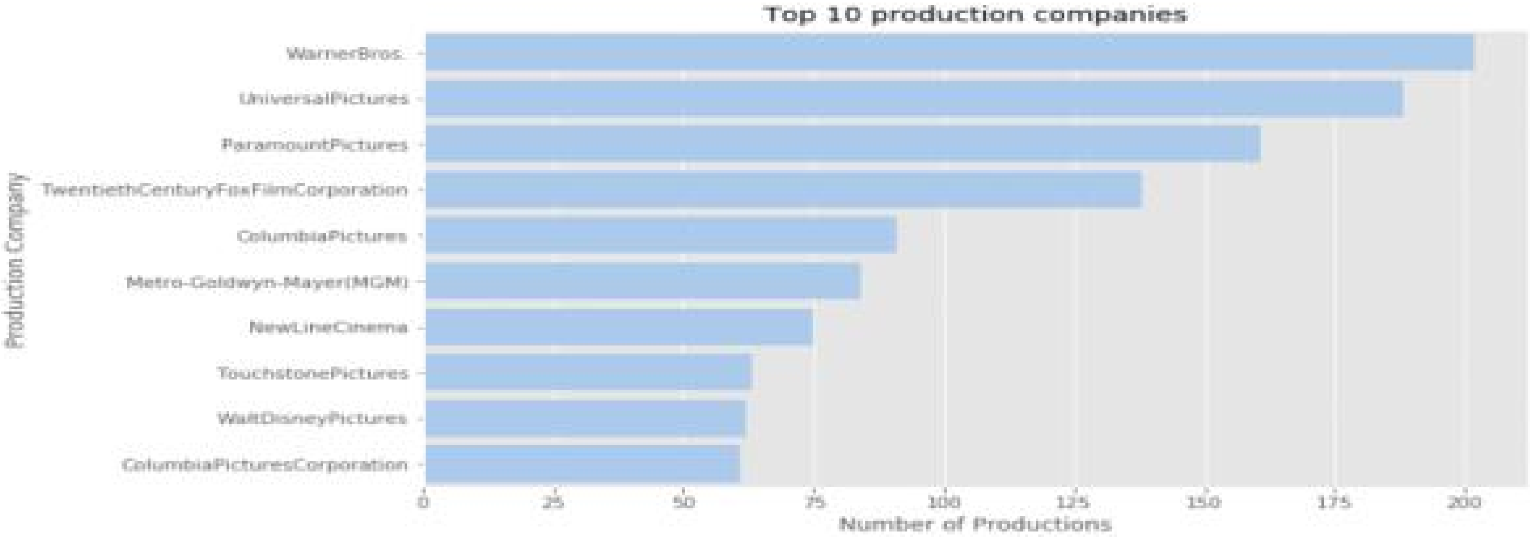


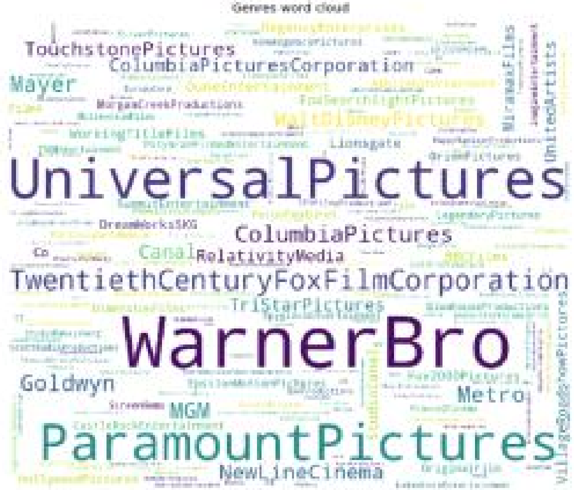
Total **budget** of **famous** cast in top 10 highest revenue **genres**



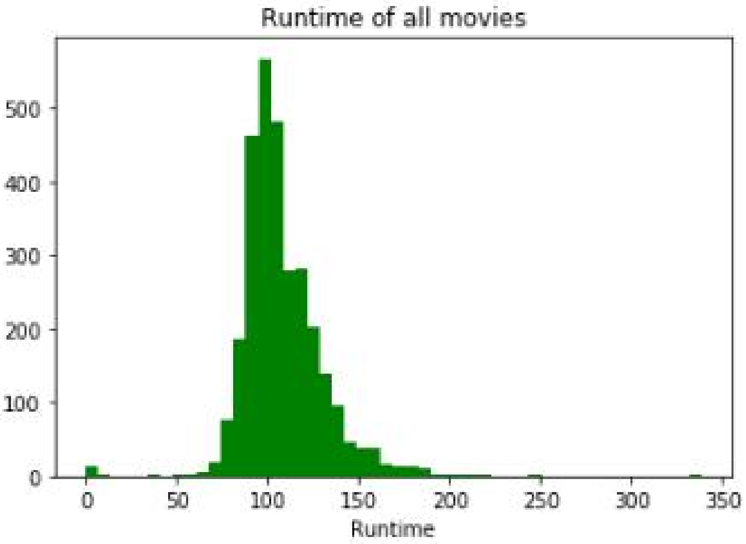
Average monthly **revenue** per **genre**

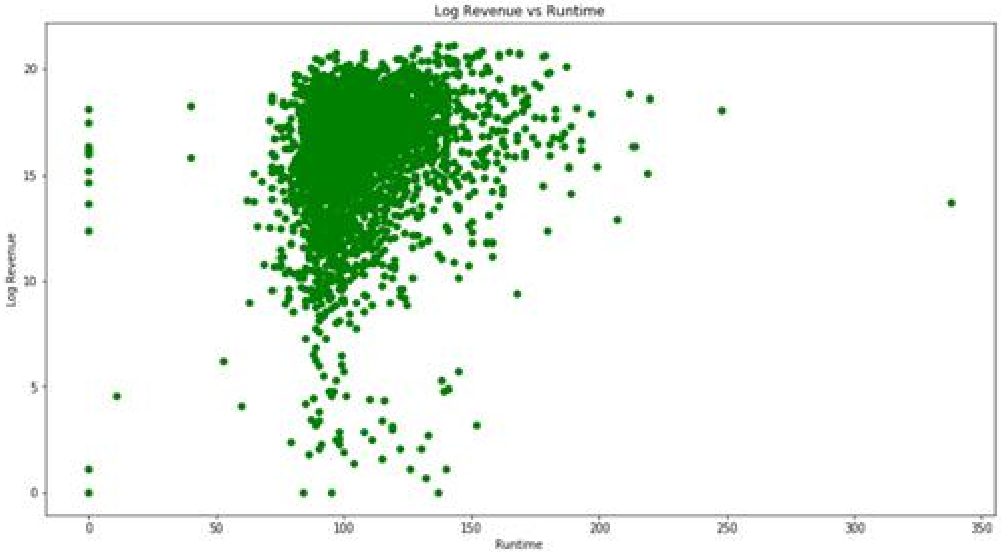
Top common words

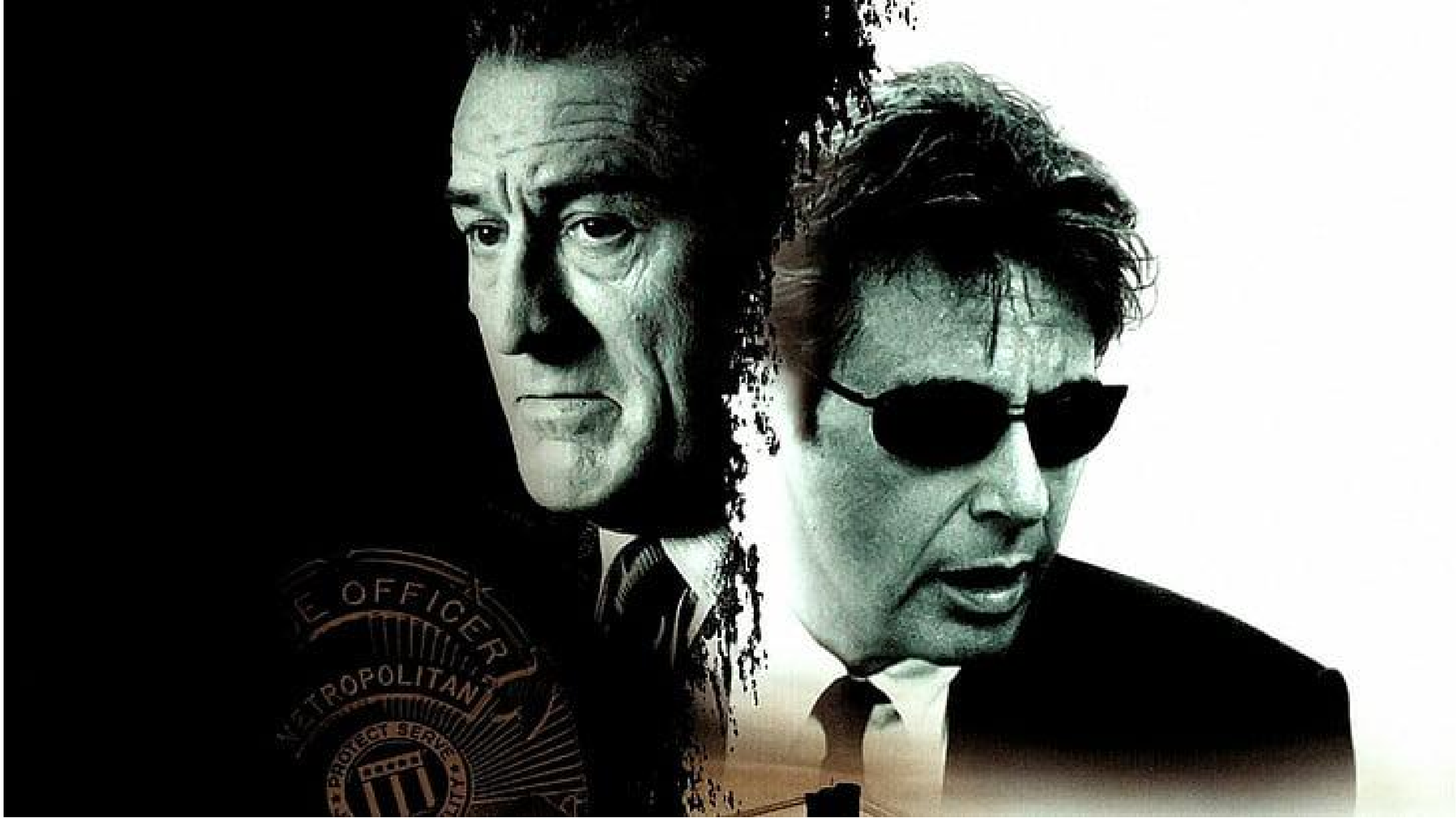

Top 10 production companies
WarnerBros.
UniversalPictures
ParamountPictures
TwentiethCenturyFoxFilmCorporation
ColumbiaPictures
Metro-Goldwyn-Mayer(MGM)
NewLineCinema
TouchstonePictures
WaltDisneyPictures
ColumbiaPicturesCorporation
Production Company
0
25
50
75
100
125
150
175
200
Number of Productions


Genres word cloud
TouchstonePictures
ColumbiaPicturesCorporation
Mayer
WaltDisneyPictures
Lionsgate
UniversalPictures
ColumbiaPictures
RelativityMedia
TwentiethCenturyFoxFilmCorporation
TriStarPictures
WarnerBro
Goldwyn
MGM
Metro
ParamountPictures
NewLineCinema
UnitedArtists

Visualizing movie runtime

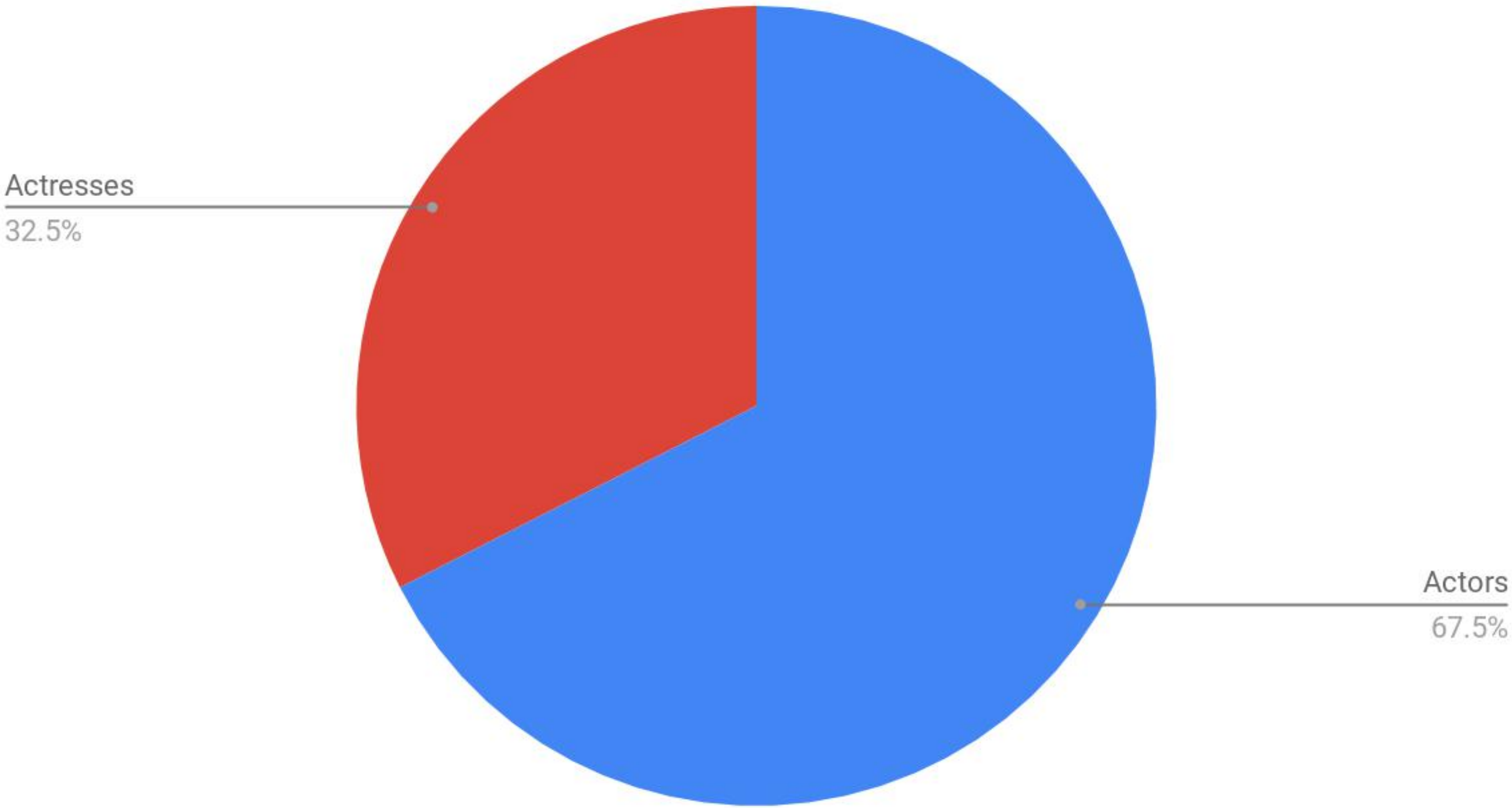




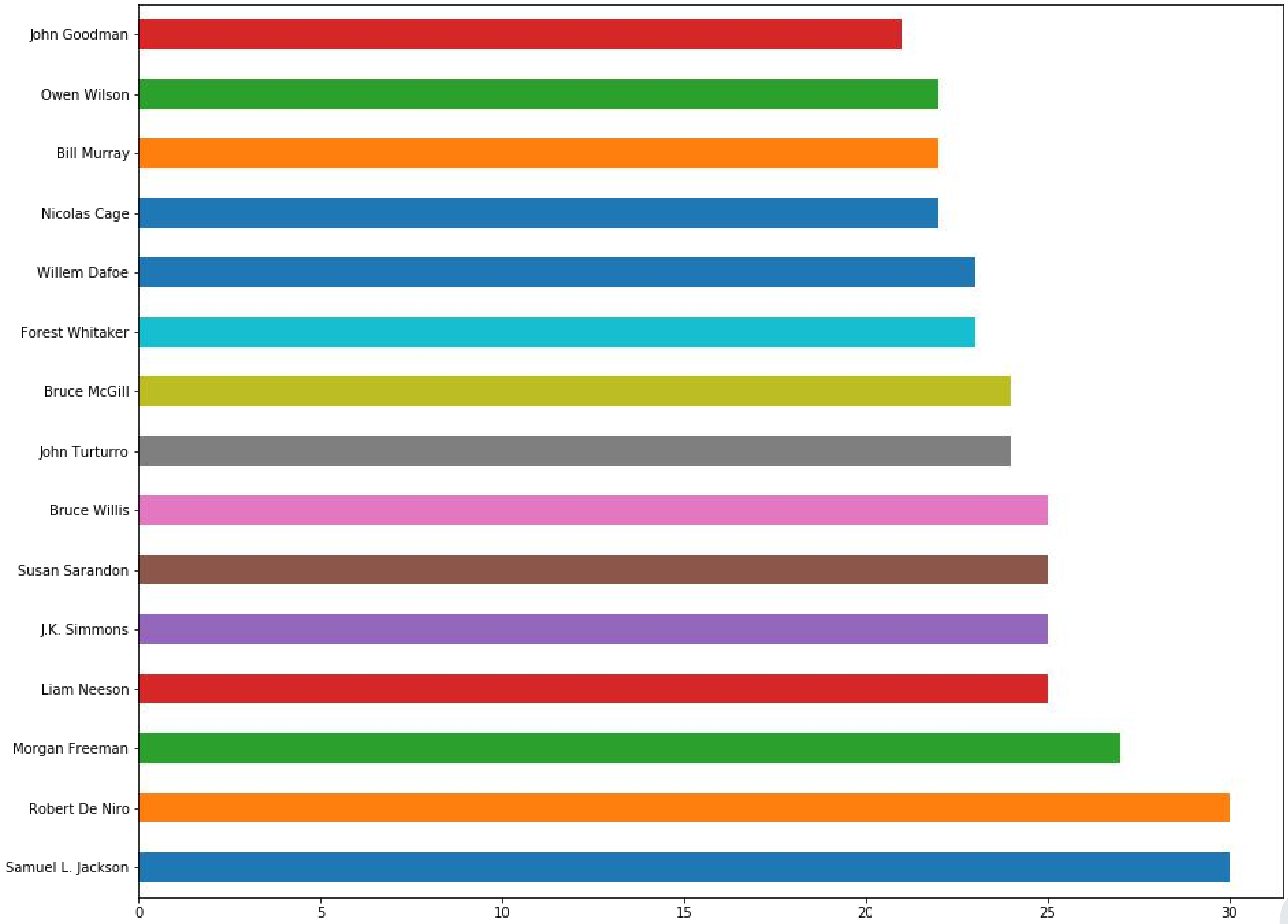


Visualizing cast

Gender of cast



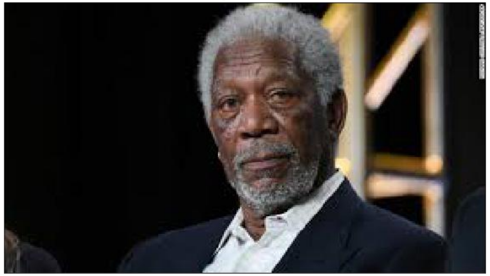
Most appearing Actors



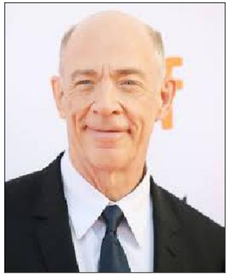
Samuel L. Jackson

Robert De Niro

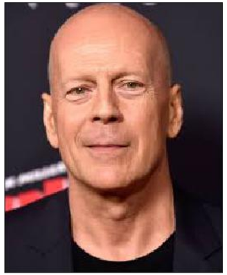
Morgan Freeman



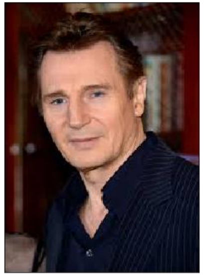
J.K. Simmons



Bruce Willis

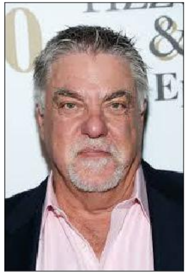


Liam Neeson

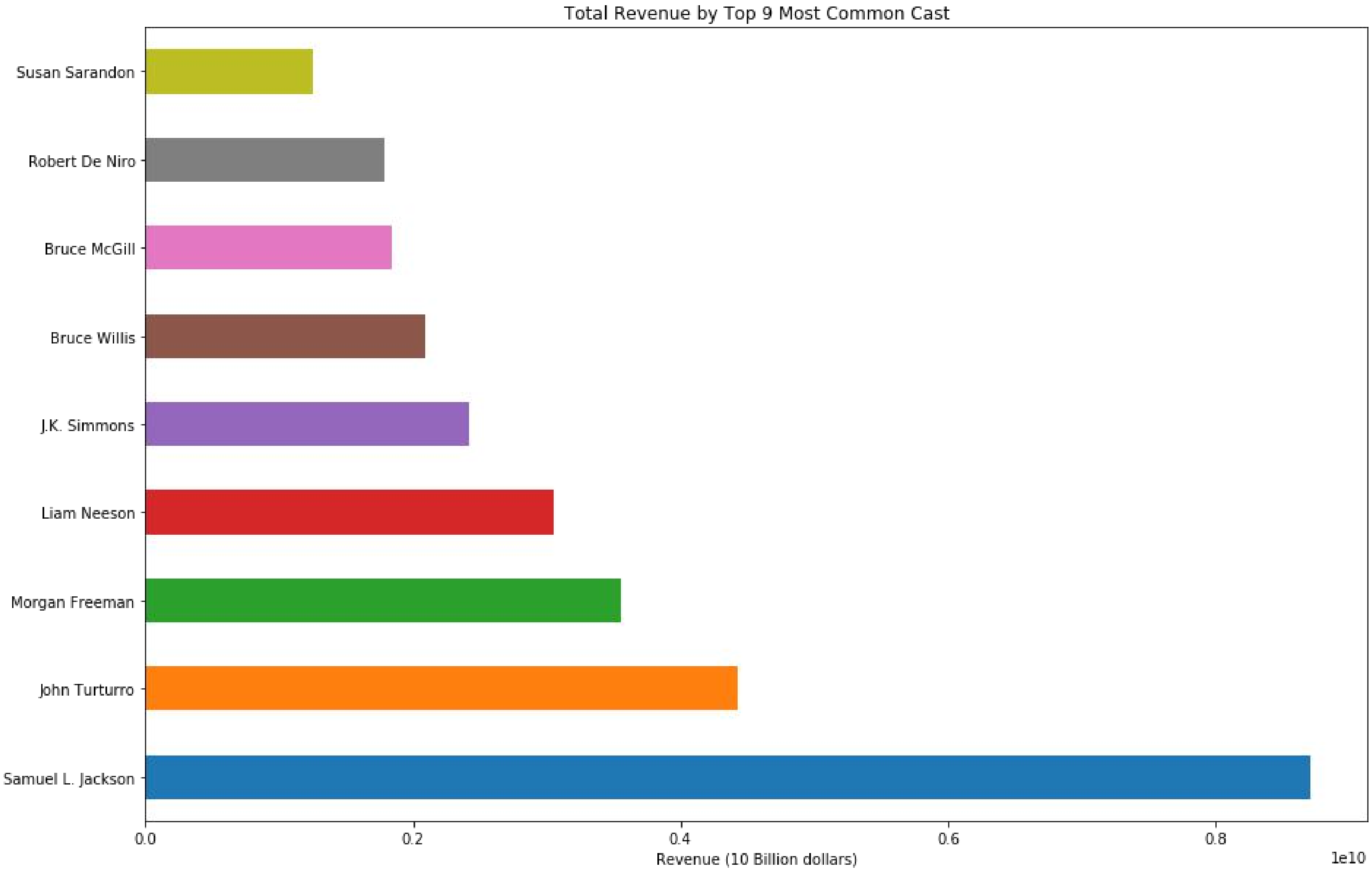


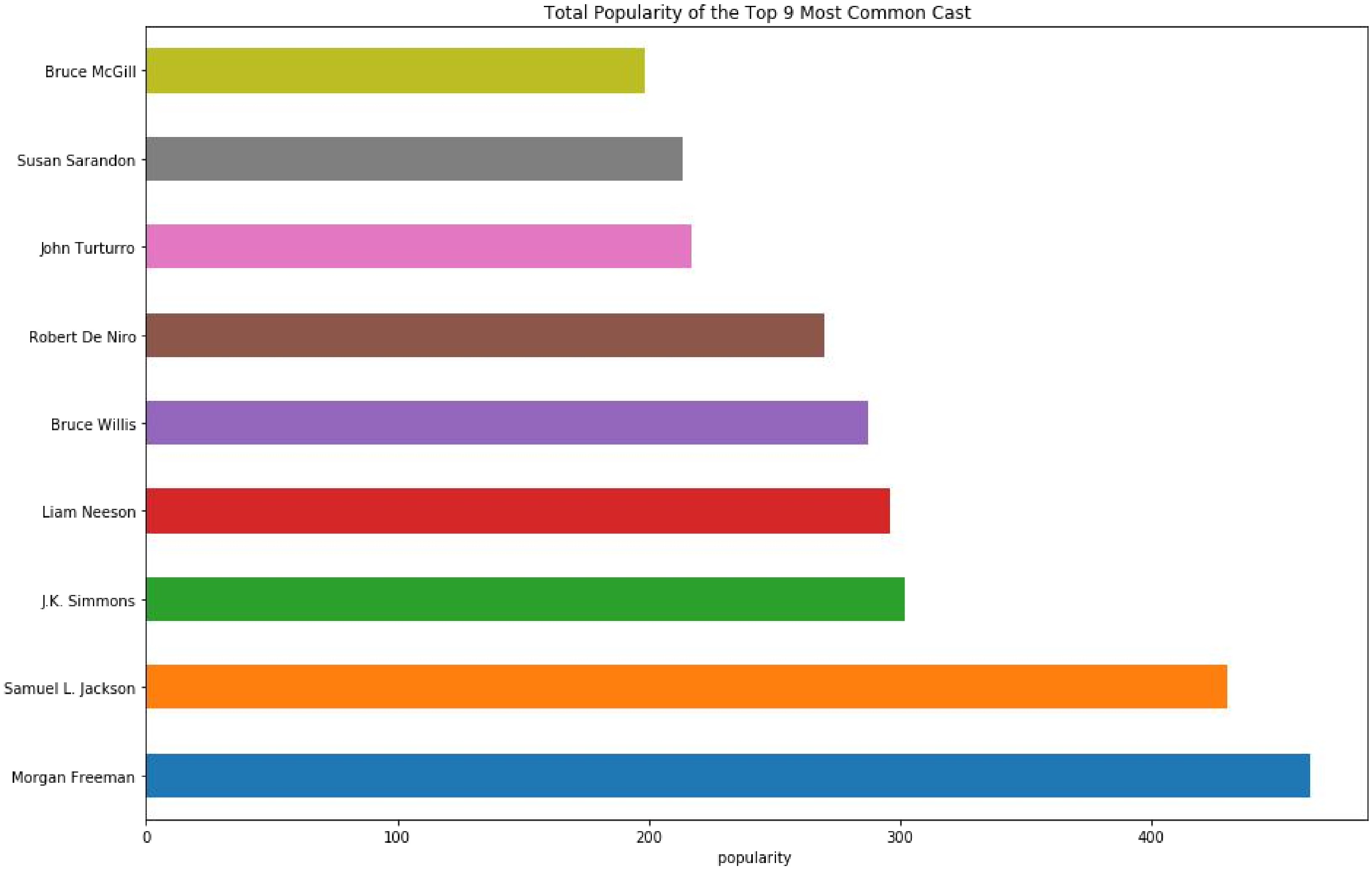
Susan Sarandon

Bruce McGill

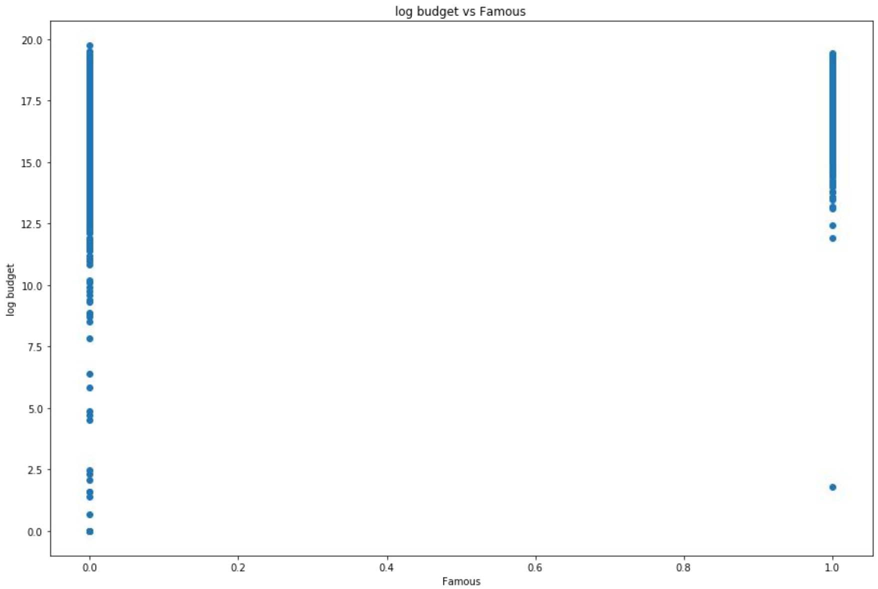


John Turturro


Total Revenue by Top 9 Most Common Cast
Susan Sarandon
Robert De Niro
Bruce McGill
Bruce Willis
J.K. Simmons
Liam Neeson
Morgan Freeman
John Turturro
Samuel L. Jackson
0.0
0.2
0.4
0.6
0.8
1e10
Revenue (10 Billion dollars)

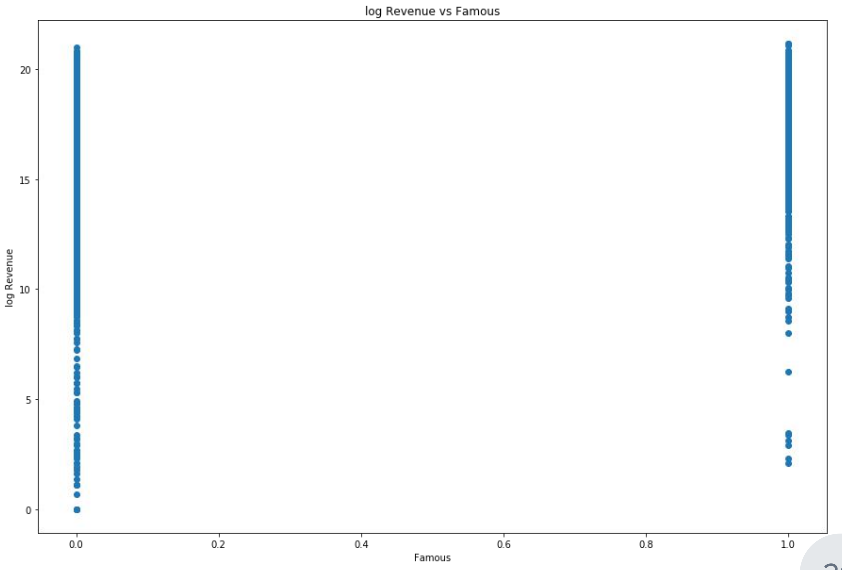

Total Popularity of the Top 9 Most Common Cast
Bruce McGill
Susan Sarandon
John Turturro
Robert De Niro
Bruce Willis
Liam Neeson
J.K. Simmons
Samuel L. Jackson
Morgan Freeman
0
100
200
300
400
popularity

Visualizing famous actors’ movies

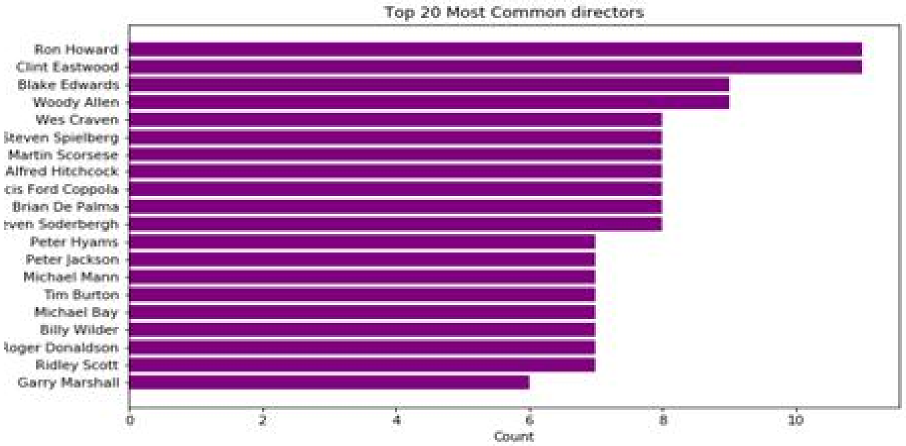


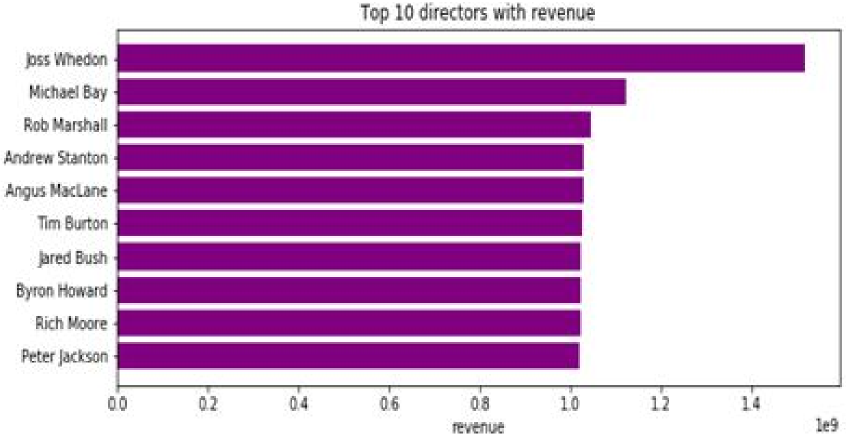
Famous and infamous VS budget

Famous and infamous VS revenue

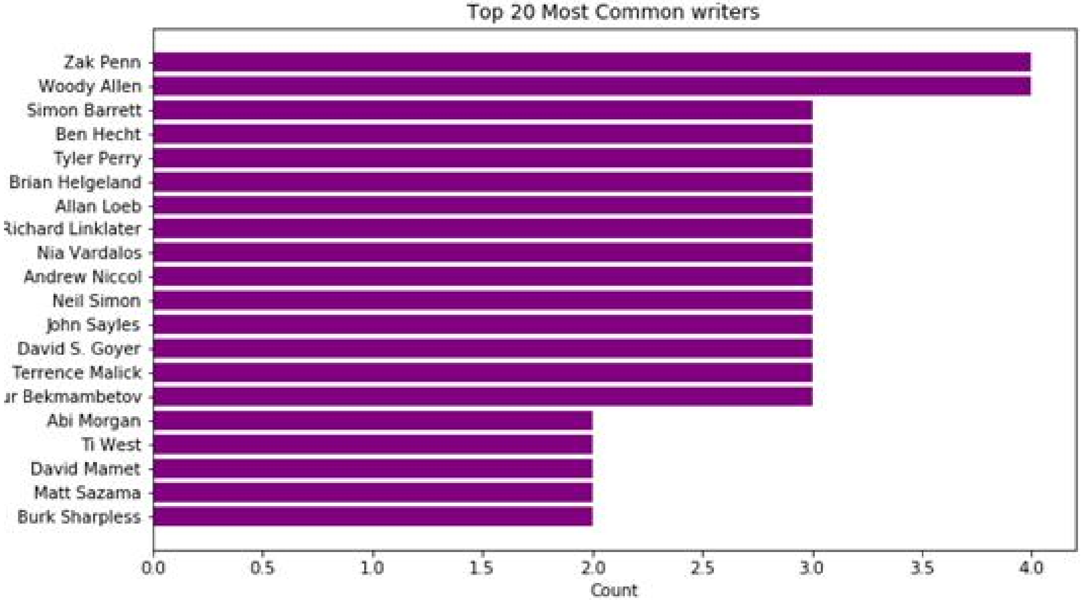


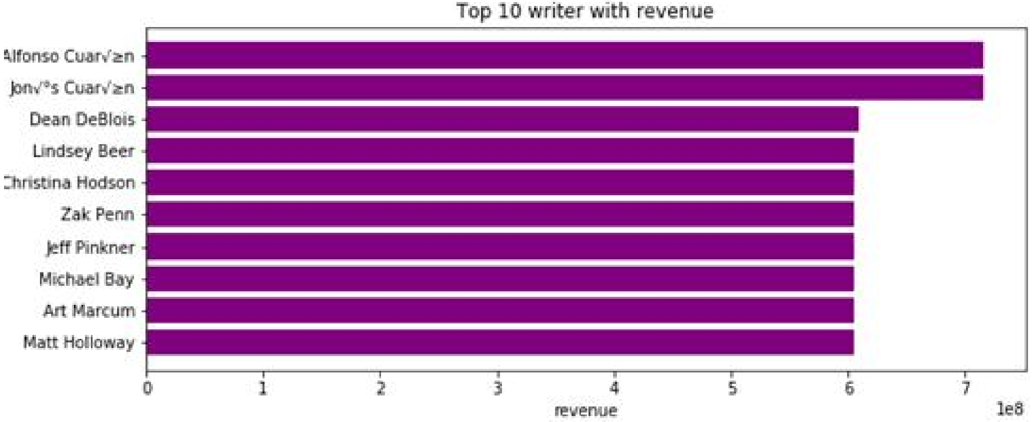
Movie Directors

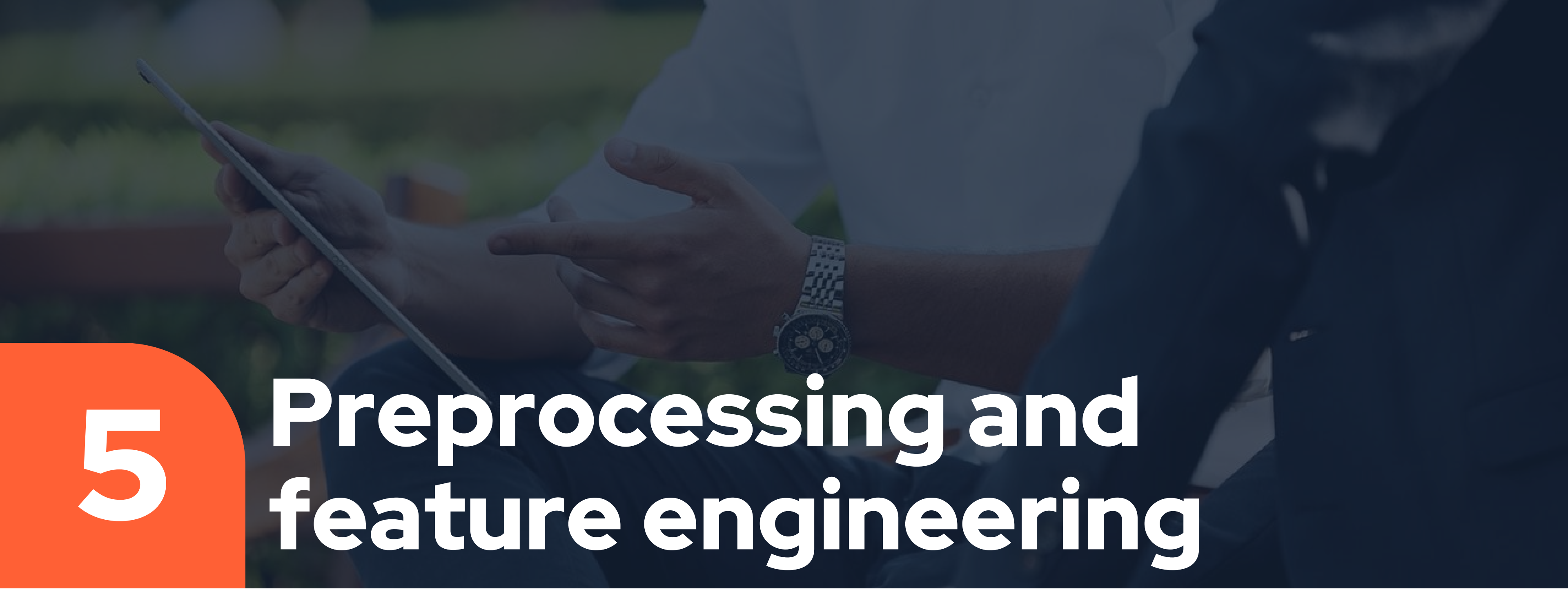




Movie Writers







5 Preprocessing and feature engineering

Transforming raw data into an understandable format. Then extract features from raw data via data mining techniques.

Preprocessing

- Columns **‘belongs_to_collection’, 'genres', 'production_companies', 'production_countries', 'spoken_languages', 'Keywords', and 'crew'** are provided as JSON strings
- convert them into list of dictionaries

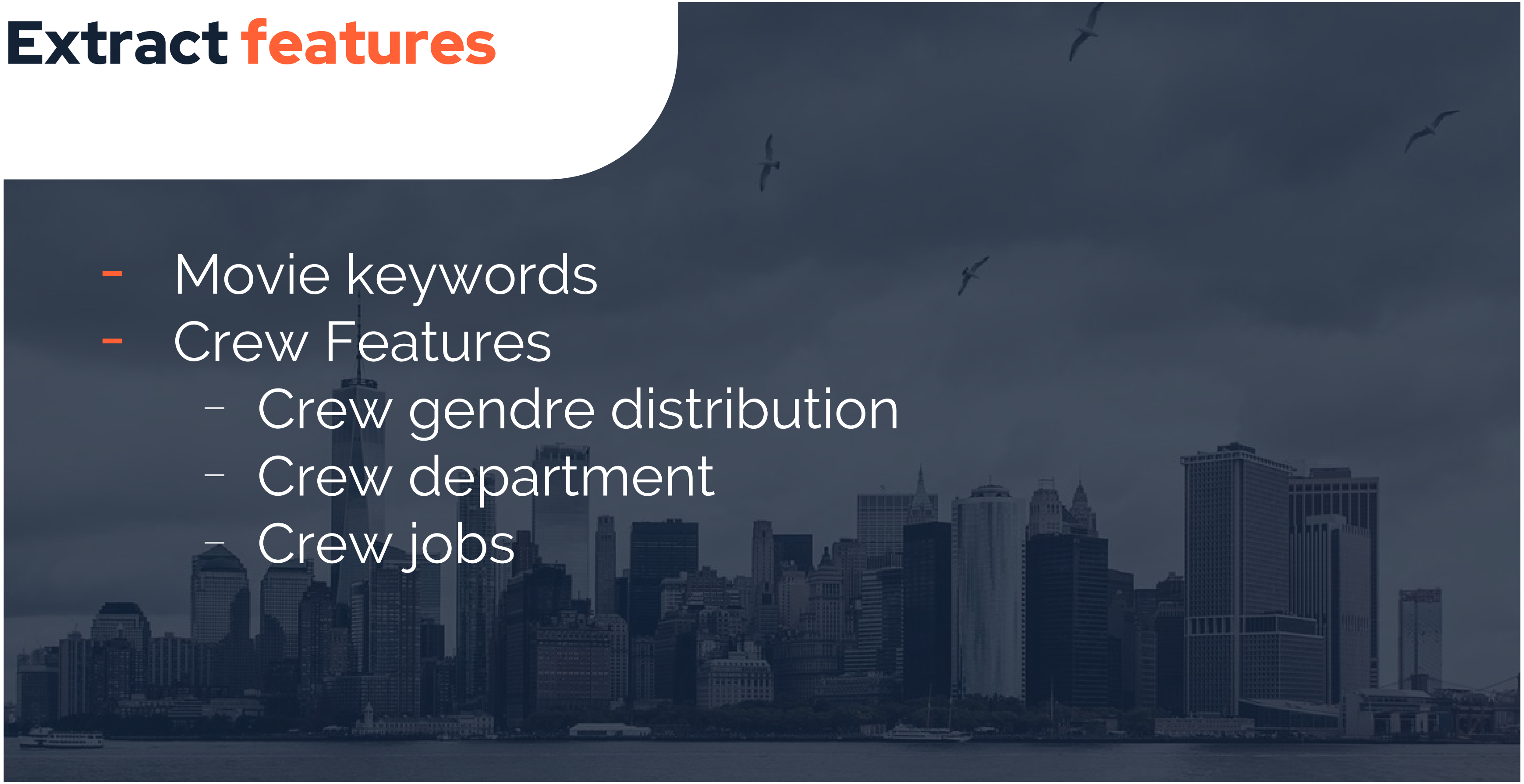
```
train['genres']

0                          [{'id': 35, 'name': 'Comedy'}]
1       [{'id': 35, 'name': 'Comedy'}, {'id': 18, 'nam...
2                           [{'id': 18, 'name': 'Drama'}]
3       [{'id': 53, 'name': 'Thriller'}, {'id': 18, 'n...
4       [{'id': 28, 'name': 'Action'}, {'id': 53, 'nam...
                              ...
2995    [{'id': 35, 'name': 'Comedy'}, {'id': 10749, '...
2996    [{'id': 18, 'name': 'Drama'}, {'id': 10402, 'n...
2997    [{'id': 80, 'name': 'Crime'}, {'id': 28, 'name...
2998    [{'id': 35, 'name': 'Comedy'}, {'id': 10749, '...
2999    [{'id': 53, 'name': 'Thriller'}, {'id': 28, 'n...
Name: genres, Length: 3000, dtype: object
```

Extract **features**

- Release date
- Genres
- Famous cast
- Production companies and countries
- Original Language

Extract features



- Movie keywords
- Crew Features
  - Crew gendre distribution
  - Crew department
  - Crew jobs

6 Model building and Training

Build a ml model for revenue predictions

Model building and Training

- Regression model called Light Gradient Boosting Machine
- The hyper-parameters:

Number of leaves = 30 per regressor
Minimum data in leaf = 20 per regressor
Learning rate = 0.01
Bagging frequency = 1
Bagging seed = 11
Boosting = gradient boosted decision trees 'gbdt'
Max depth of boosting trees = 5
1 regularization using lambda = 0.2
Feature fraction = 0.9
Bagging fraction = 0.9
Number of regressors = 20,000
Early stopping = 200 rounds

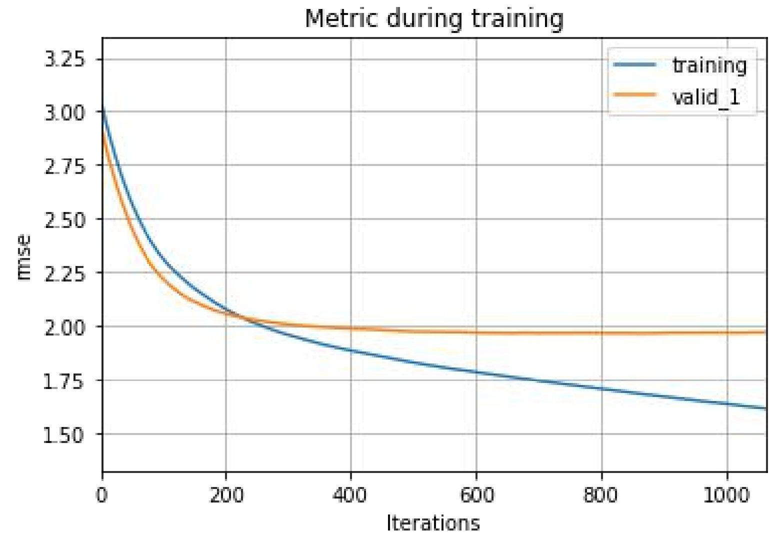
7 Evaluation and accuracy

How good the model is??

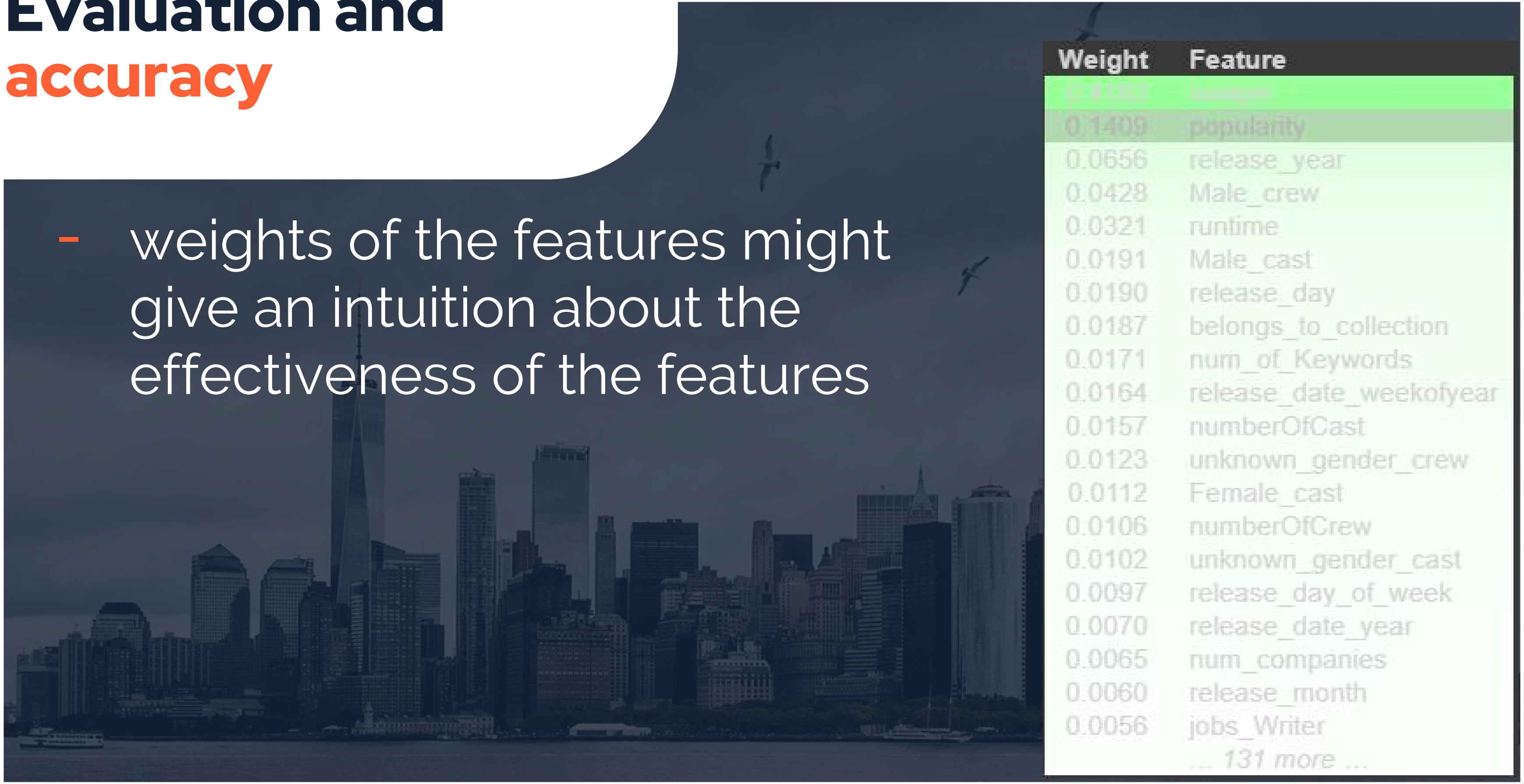
Evaluation and accuracy

Training RMSE = 1.68197,
Validation RMSE = 1.96516

Test RMSE = 2.04048



Evaluation and **accuracy**



- weights of the features might give an intuition about the effectiveness of the features

Weight	Feature
[illegible]	[illegible]
0.1409	popularity
0.0656	release_year
0.0428	Male_crew
0.0321	runtime
0.0191	Male_cast
0.0190	release_day
0.0187	belongs_to_collection
0.0171	num_of_Keywords
0.0164	release_date_weekofyear
0.0157	numberOfCast
0.0123	unknown_gender_crew
0.0112	Female_cast
0.0106	numberOfCrew
0.0102	unknown_gender_cast
0.0097	release_day_of_week
0.0070	release_date_year
0.0065	num_companies
0.0060	release_month
0.0056	jobs_Writer
	... 131 more ...

8 Unsuccessful trials

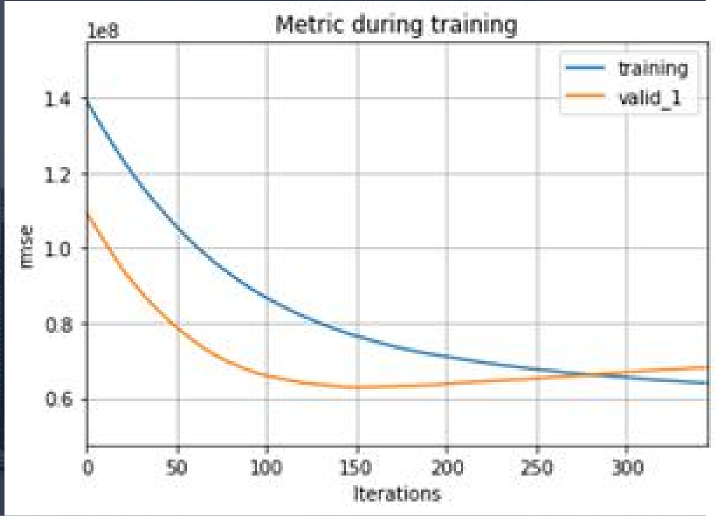
Models with bad accuracy

Unsuccessful **trials**

- Using basic feature set

1. Light GBM Regressor

training RMSE: 7.74808e+07, validation RMSE: 6.20101e+07

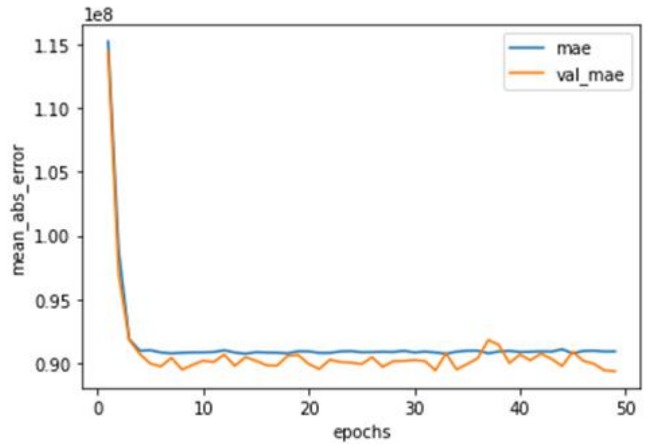


Unsuccessful trials

- Using basic feature set

1. Keras sequential neural network
   training RMSE: 9.0934216e+07, validation RMSE: 8.9378792e+07

1. Random forest regressors
   training RMSE: 2.8186479e+07, validation RMSE: 6.1342776+07



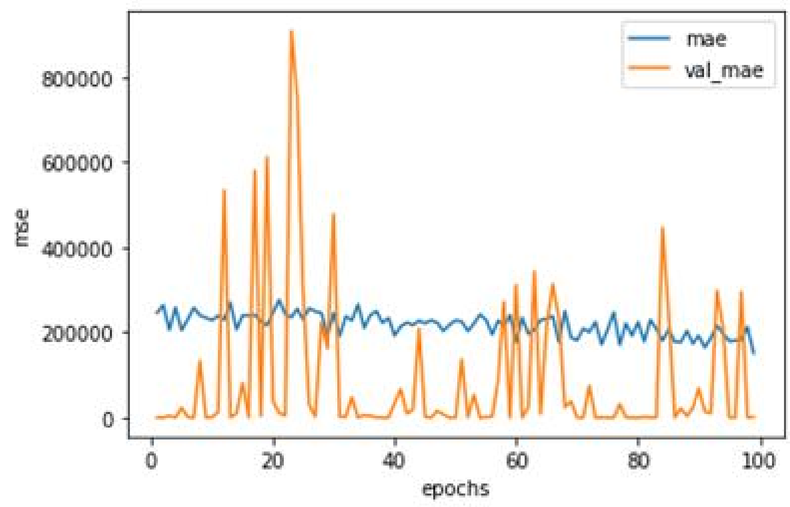
Unsuccessful **trials**

- Using modified feature set

1. Keras sequential neural network

2. Random forest regressors

   training RMSE was 0.8 and validation RMSE is 2.4



9 Future work

Try to enhance in future

Future work

- Try XGBOOST rather than the LGBM
- Hyperparameter search and fine tune
- Combine models and make an ensemble of the models

Thanks!