CP1 - What movies make the most money at the box office
Milestone Report 1

Problem Statement

In a world where movies made an estimated $41.7 billion in 2018, the film industry is more popular than ever. But what movies make the most money at the box office? How much does a director matter? Or the budget?
Can we build models, which will be able to accurately predict film revenue?

This project will help film production companies understand key features of having high revenue.

Data Source

The major data source comes from the public dataset uploaded to Kaggle.com (https://www.kaggle.com/c/tmdb-box-office-prediction/overview).

This dataset with metadata on over 7,000 past films from The Movie Database. Data points provided include cast, crew, plot keywords, budget, posters, release dates, languages, production companies, and countries.

Data Cleaning

There are 8 JSON-style columns. We will parse them and create categorical and dummy variables. For example, column “belongs_to_collection”:

belongs_to_collection

```
for i, e in enumerate(master['belongs_to_collection'][:5]):
    print(i, e)
```

```
0 [{'id': 313576, 'name': 'Hot Tub Time Machine Collection', 'poster_path': '/iEhb00TGPucF0b4joM1ieyY026U.jpg', 'backdrop_p
ath': '/noeTVcgpBiD48fDjFVic1Vz7ope.jpg'}]
1 [{'id': 107674, 'name': 'The Princess Diaries Collection', 'poster_path': '/wt5AMbxPTS4Kfjx7Fgm149qPfZl.jpg', 'backdrop_p
ath': '/zSEtYD77pKRJlUPx34BJgUG9v1c.jpg'}]
2 nan
3 nan
4 nan
```

We create two new columns from column "belongs_to collection", first one is collection name and second one has collection or not. We assume that other information from this column we can’t use for future prediction.

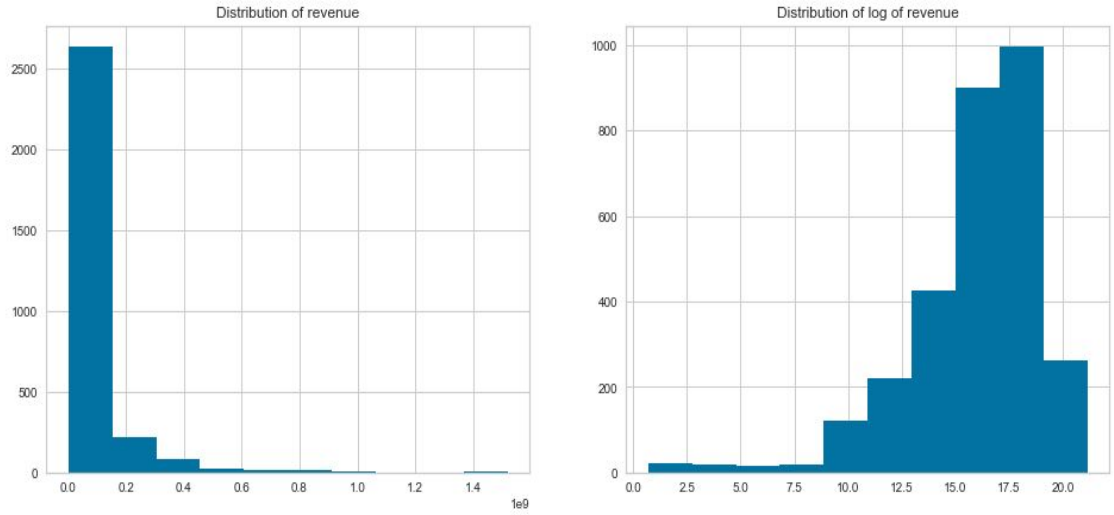
```
master['collection_name'] = master['belongs_to_collection'].apply(lambda x: x[0]['name'] if x != {} else 0)
master['has_collection'] = master['belongs_to_collection'].apply(lambda x: len(x) if x != {} else 0)

master = master.drop(['belongs_to_collection'], axis=1)
```

Exploratory Analysis

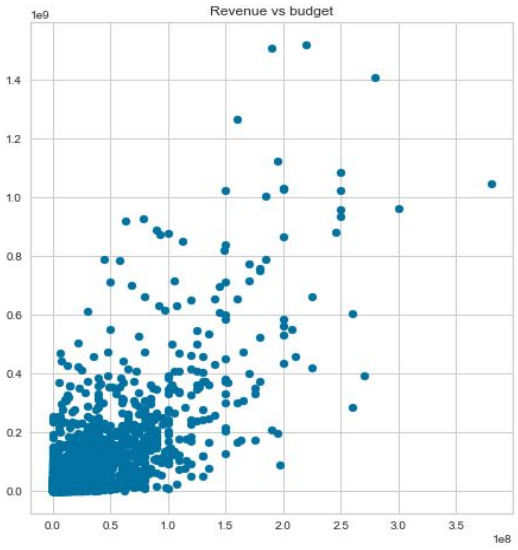
SOME INSIGHTS

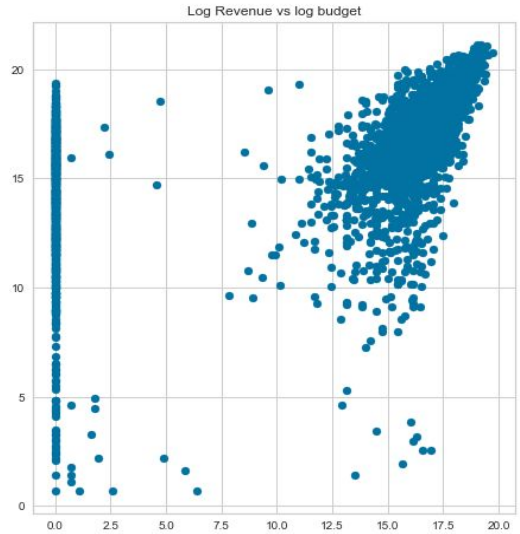
- Average length of a movie is about **107 minutes**
- The **longest** movie: *"Carlos"* is **5hrs 38min** long
- The **most popular** movies: *"Wonder Woman"* and *"Beauty and the Beast"*
- Movies with the **biggest budget**: *"Pirates of the Caribbean: On Stranger Tides"* and *"Pirates of the Caribbean: At World's End"*
- The **biggest** movie **producers**: *Warner Bros.* and *Universal Pictures*
- The most popular **genres**: *Drama* and *Comedy*

Let first look at the histograms of the target feature - **revenue**. As we can see revenue distribution has a high skewness. It is better to use np.log1p of revenue.

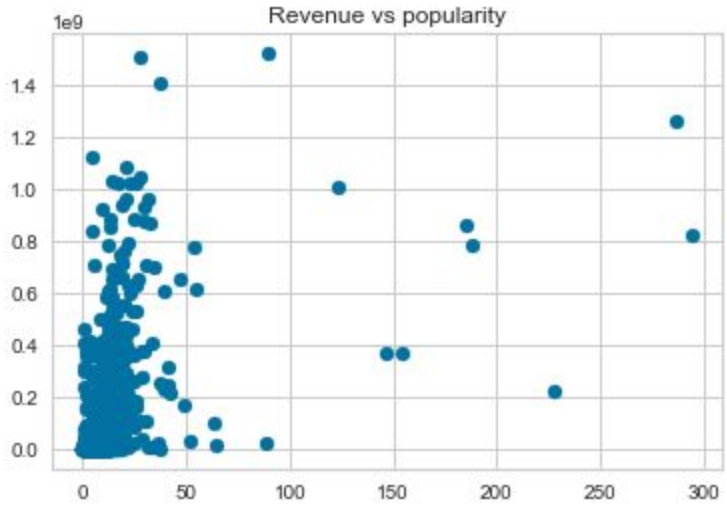


First feature we will look at is **budget**, intuitively the budget should be somehow correlated with revenue.

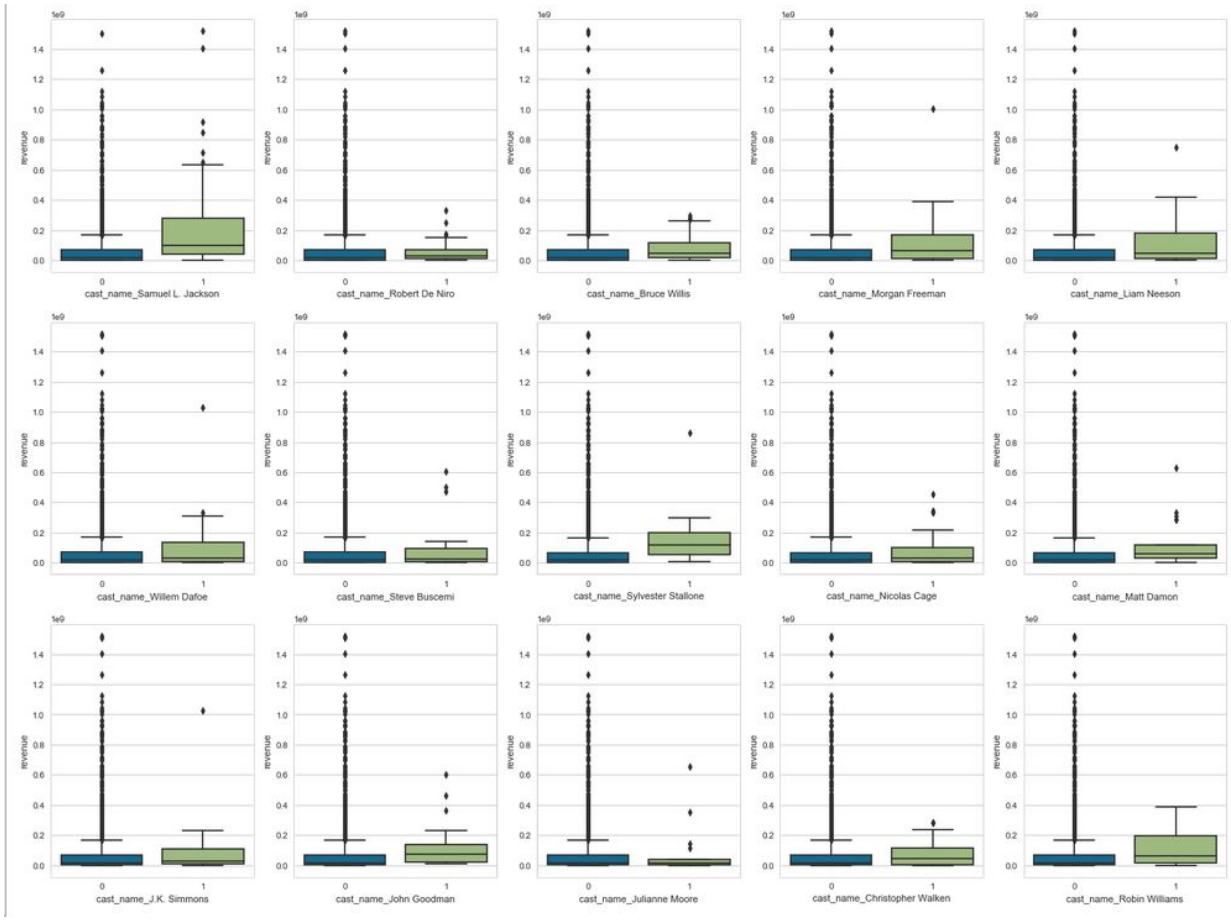




Let's look at popularity. We can see some clear trends that an increase in popularity tends to lead to higher revenue.

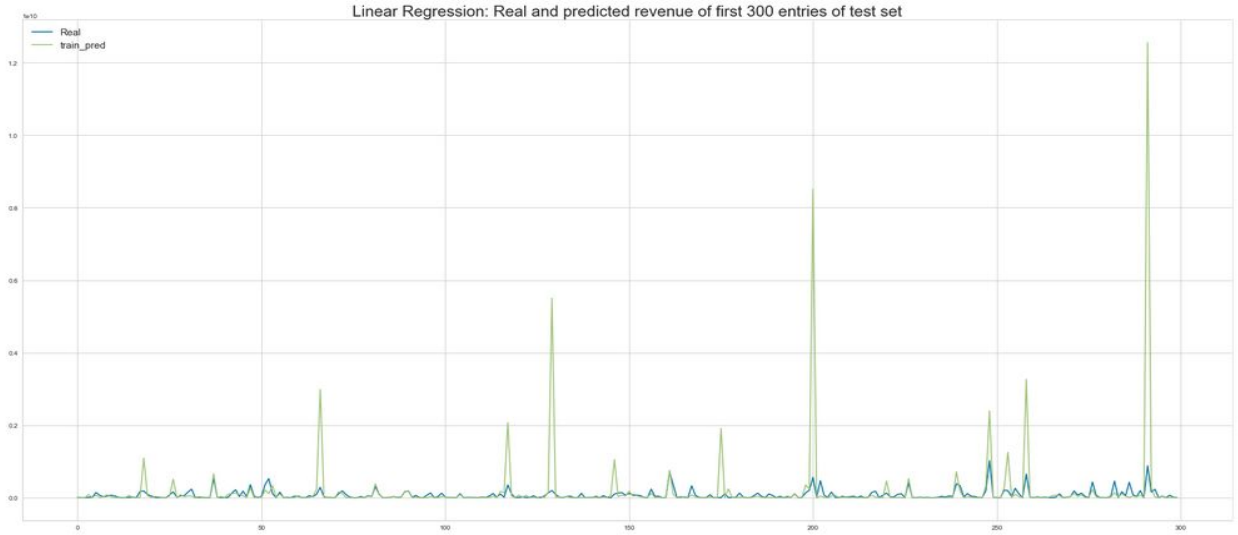


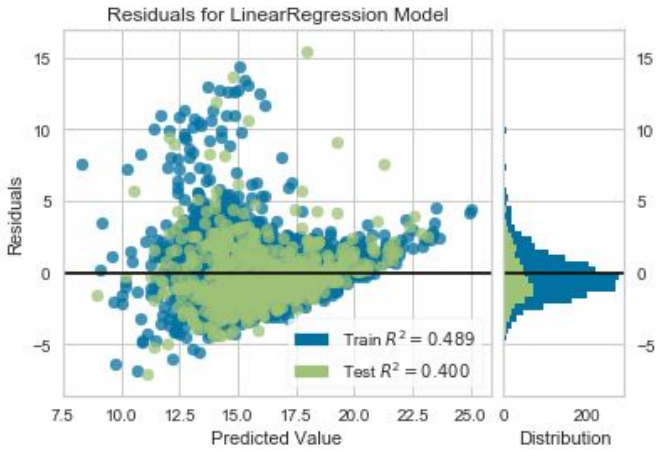
We have columns with most common cast members, so let's plot boxplots.



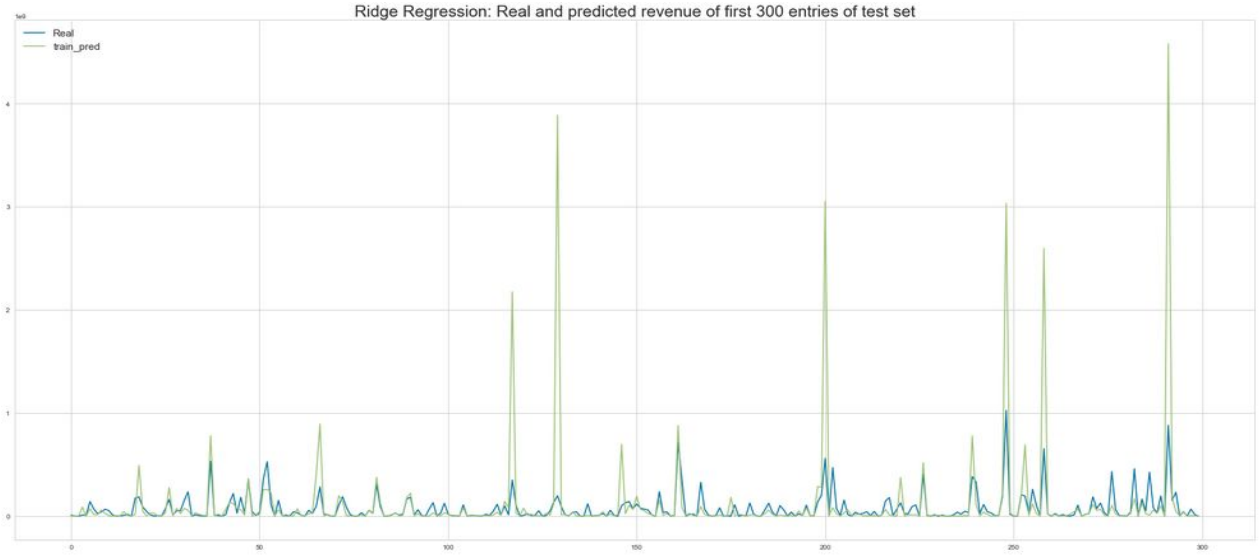
As you can see mostly films with these actors tend to have higher revenue.

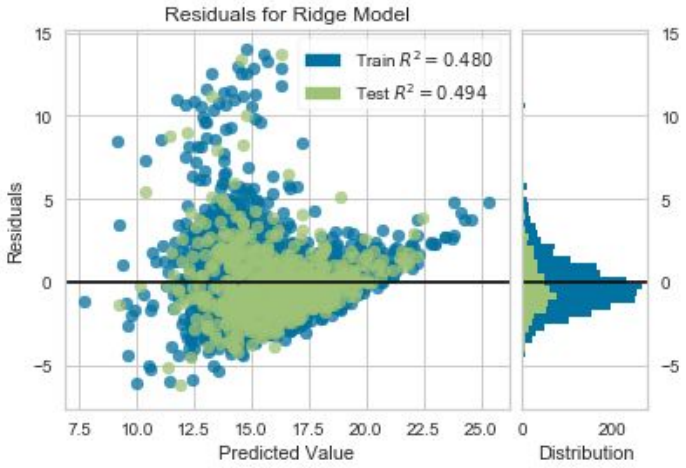
Linear Regression



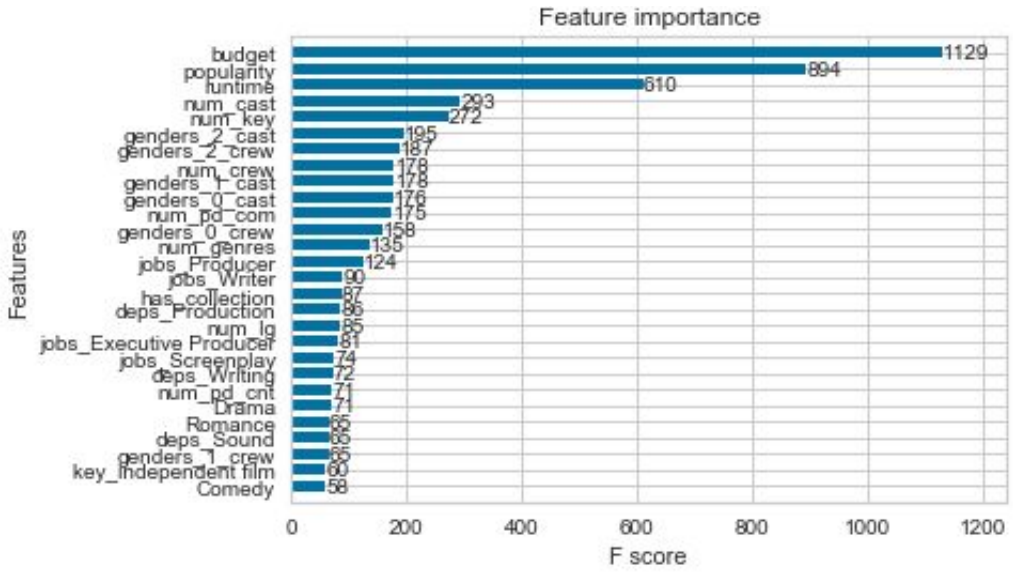


Ridge Regression





XGBOOST



	budget	popularity	runtime	num_cast	num_key	genders_2_cast	genders_2_crew	genders_1_cast	num_crew	genders_0_cast	num_pd_com
0	3500000	0.556	90.000	11	2	3	2	5	2	3	1
1	0	2.087	100.000	7	7	2	5	1	13	4	2
2	2000000	1.189	89.000	3	1	2	0	1	1	0	1
3	98000000	7.284	119.000	31	6	20	12	2	16	9	4
4	0	1.219	101.000	7	0	3	0	0	2	4	1

genders_0_crew	num_genres	jobs_Producer	jobs_Writer	has_collection	deps_Production	num_lg	jobs_Executive Producer	jobs_Screenplay	deps_Writing
0	1	0	1	0	0	1	0	0	1
8	2	2	2	0	2	3	0	0	2
1	3	0	0	0	0	1	0	0	0
2	2	4	0	0	6	2	1	2	4
2	4	0	1	1	0	1	0	0	1

num_pd_cnt	Drama	genders_1_crew	deps_Sound	Romance	key_independent film	Comedy	revenue	predicted_revenue	title
1	0	0	0	0	0	1	16.040	14.407	Ringmaster
1	0	0	3	0	0	0	2.079	13.655	He-Man and She-Ra: The Secret of the Sword
1	1	0	0	0	1	1	10.425	11.509	Cowboys & Angels
4	0	2	0	0	0	0	16.120	18.579	Cutthroat Island
1	1	0	0	0	0	0	16.003	13.237	We Are from the Future 2

1) Rmse xgboost how to choose the right one train or test?
2) How to extract properly rmse from output of code?
3) Is it good to have opposite features in data?
4) Contacts for informational interviews