# 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.

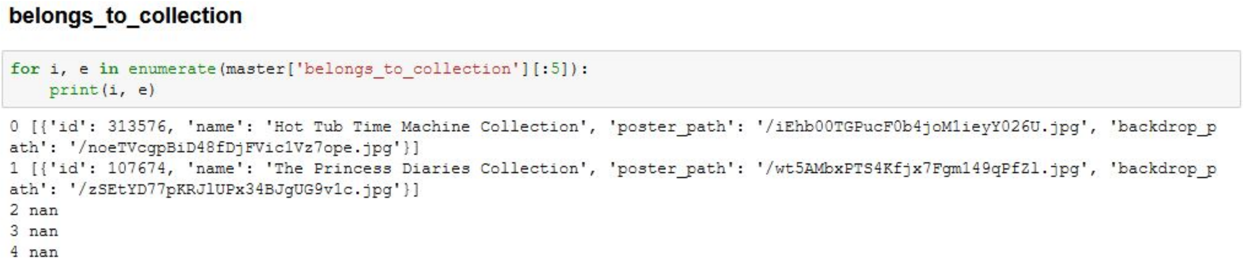
## Data Source

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”:

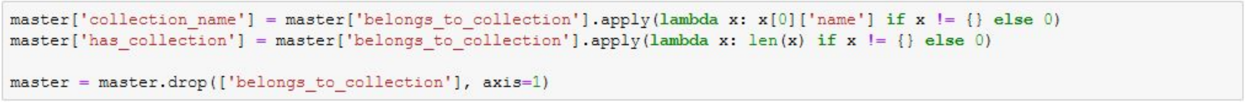


**belongs_to_collection**

```
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': '/zSEtYD77pKRJ1UPx34BJgUG9v1c.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)
```