TMDB Movies Project Proposal

This repository to meet requirements for data science bootcamp with SADIA.

Do the highest revenue films depend on the genres of film and cast?

Film production is such a big industry that has the attention of just about anyone who can see them. Major studios and indie filmmakers alike spend much of their days looking for new sources of revenue this study help filmmakers decide whether or not the cast and genre of the movie affect the revenue.

Dataset

We have data set **tmdb-movies** that represents movies information based on TMDB rating with 10,000 records and 21 features. I use this data set for extract useful information to production companies. This dataset can be found at Kaggle.

This dataset contains Movies for the following genres:

- Action
- Animation
- Comedy
- Crime
- Drama
- Experimental
- Fantasy
- Historical
- Horror
- Romance
- Science Fiction
- Thriller
- Western
- Other Genres

The dataset is available as the `.csv` file. a sample of data is shown in the following table:

id	imdb_id	popularity	budget	revenue	original_title	cast	homepage	director	tagline	keywords	overview	runtime
135397	tt0369610	32.985763	150000000	1513528810	Jurassic World	Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...	http://www.jurassicworld.com/	Colin Trevorrow	The park is open.	monster|dna|tyrannosaurus rex|velociraptor|island	Twenty-two years after the events of Jurassic ...	124

genres	production_companies	release_date	vote_count	vote_average	release_year	budget_adj	revenue_adj
Action|Adventure|Science Fiction|Thriller	Universal Studios|Amblin Entertainment|Legenda...	6/9/15	5562	6.5	2015	1.379999e+08	1.392446e+09

The most important features for this study:
genres, which contained the type of movies
cast, which included the actors name
revenue are used to identify (the target)
The liner Regression Analysis is the process of predicting a Label based on the features at hand so, here we use regression to know relationship between revenue and other features like cast and genres.

Tools

I used Jupyter notebook to write codes and import libraries to achieve the goal of this dataset, such as:
`numby, matplotlib, pandas` used to EDA and prepare data to train a model a.
and for regression use `sklearn, LinearRegression (function)` will be used to analyzes the relationship between two or more features it is type of supervised learning to train the model.

TO DO:

- I will do wrangling data and exploratory data analysis before used the model.