Structure of folder

```
/MovieAnalyse                       ----- root directory
|   datasets_2745_4700_movies.csv   ----- movie dataset
|   gdps_1970_2024.csv              ----- gdp dataset
|   movie_dataset_analyse.ipynb     ----- notebook
|   movie_directors_information.csv ----- director dataset
|   my_dataset.csv          ----- training/testing dataset
|   parser_director.py      ----- movie director parser
|   parser_gdp.py           ----- gdp parser
|   README.md               ----- readme file
```

Our Work

This project aimes to analyse the movie dataset feaures.

We download a movie dataset (**datasets_2745_4700_movies.csv**) from kaggle, which contains the information of movies including the movie country, the movie comapny, the movie director, the movie name, the movie genre, the movie budget, vote, score, and the gross.

In order to practice python scratching and to build our own traing/testing dataset, we write two script: **parser_director.py** and **parser_gdp.py**. These two scratching script can parser wikipedia websites and output two csv files: **movie_directors_information.csv** and **gdps_1970_2024.csv**

To execute these two scratching scripts, you might need **bs4**, **urllib3**, **csv** and **re**. Attention: the **parser_director.py** script might take 20-40 minutes, beacause it will parser several thousands wikipedia websites.

The **movie_directors_information.csv** contains the movie director information such as the director name, his education, his birth day and birth place, his roles, ect. The **gdps_1970_2024.csv** contains the gdp of different countries and areas between 1970 and 2024. For more details, please check the source wikipedia websites.

To build our own training/testing dataset, we will merge **datasets_2745_4700_movies.csv** and **movie_directors_information.csv** using director's name as the union key to get an intermediate file. After that we merge this intermediate file and **gdps_1970_2024.csv** using the county name as the union key to get the final dataset. The final dataset is saved as **my_dataset.csv**. For this part, you can check the notebook to find more information.

After executing the two scratching scripts, you can read the code in the notebook **movie_dataset_analyse.ipynb**. This notebook aimes to proccess the dataset, to split the dataset into traing dataset and testing dataset, to use the model of SVM of **sklearn**, to visualize the dataset features, .etc. You can read the comments of the notebook to understand each step.