Predicting US Movies Gross Using Linear Regression

Leena AlQasem
Leenabdulh@gmail.com

Randa Mohammed
randa1414@gmail.com

Introduction:

The second project for Data science Bootcamp T5 is predicting US movies gross using multiple linear Regression. Hence, the idea behind this project is to find the best ML models to predict US movie gross based on many factors. In addition, this project sheds light on the IMDb dataset that we will use, dataset description, and finally, the tools used in this project.

- **Company information:** The Internet Movie Database (IMDb), a website that provides the most extensive information about millions of movies, TV shows, and their cast and crew. It's a highly detailed source of movie data that features top movies, news, free movies, reviews, celebrity profiles, and more.
- **Problem statement:** Prediction of the correct box office gross is essential for the movie industry and decreasing market risk. The success and failure of the movie depend on movie-related variables, for instance: the duration of a movie and release date.
- **Value for the company:** Due to the importance of gross in the production of the movie, we will determine the factors that impact the revenue. Therefore, accurate prediction of box office revenues will improve the financial aspects as well as customer satisfaction.

Dataset: The datasets that will be used in this project will be scraped from the IMDB website. Then, the data will be cleaned and analyzed in order to be used for prediction, using multiple factors to predict the gross. Those factors include the movie's rate, ticket price, duration, release date, IMDb's rate, etc.

Tools:

- **Technologies:** Jupyter Notebook, Python.
- **Libraries:** Pandas, NumPy, Matplotlib, Seaborn, Request, BeautifulSoup, Selenium, Statsmodel and Scikit-learn.