Project Proposal

Predict Movie Revenue using Linear Regression

**Purpose:**

Gather movie dataset scraped from imdb.com / boxofficemojo.com and build linear regression model to predict the movie revenue.

**Goal:**

Predict the movie revenue

**Data Description:**

Using web scraping to get the movie dataset from public websites:

* <https://www.boxofficemojo.com/>
* <https://www.imdb.com/>

Preliminary Data Explorations –

Clean the scraped movie dataset.

Transform the categorical variables using dummy variables (1 hot encoding).

Target (y) *is Movie revenue* that we want to predict

Features (x) for the regression

Features for the regression:

1. Movie Name
2. Budget
3. No of Theatres
4. Weekly Gross Revenue
5. Dates (week wise)
6. Weekend Revenue
7. Percent change last week
8. Percent change overall last weel
9. Number of releases (weekly)
10. Rank
11. IMDB Rating
12. Rotten Tomatoes Rating
13. Movie Duration
14. Release date
15. MPAA Rating (G, G/PG, R, etc.)
16. Genre (Comedy, Thriller, Romance, etc.)
17. Special Days (eg – holidays / long weekends)

Out of the total data set, I will randomly divide it into 2 buckets –

1. Training Data (80%)
2. Test data (20%)

I will use this training data to train the model and build linear regression on.

**Assumptions –**

1. The residuals (errors) form a Normal Distribution
2. Linearity: The relationship between the features X and the target of Y is linear
3. No Correlation
4. No Multicollinearity

**Tools:**

Web Scraping tools –

1. Beautiful Soup
2. Selenium

Google Chrome driver – to enable selenium code access chrome browser pages

Python packages for Linear Regression –

1. Scikit-learn – build Linear Regression model
2. Pandas – manipulating and working on data frames
3. NumPy – working with arrays
4. matplotlib and seaborn – data visualization

**MVP Goal:**

Have a web scraped, cleaned with some preliminary analysis on the movie data set ready.