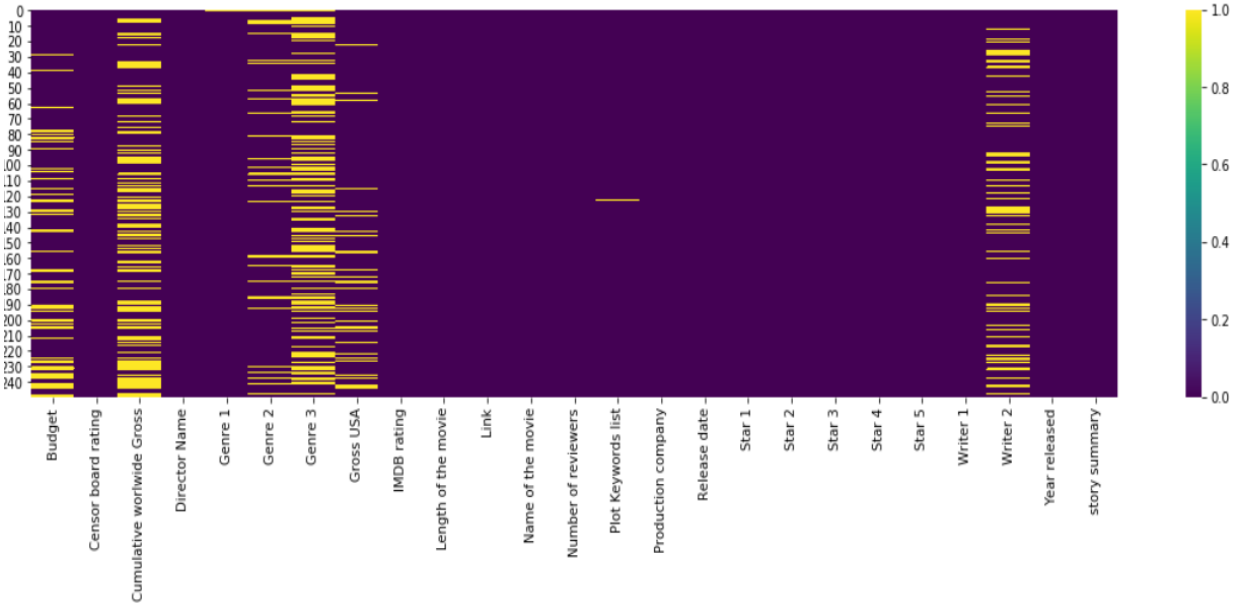
APPROACH FOR THE MODEL BUILDING:

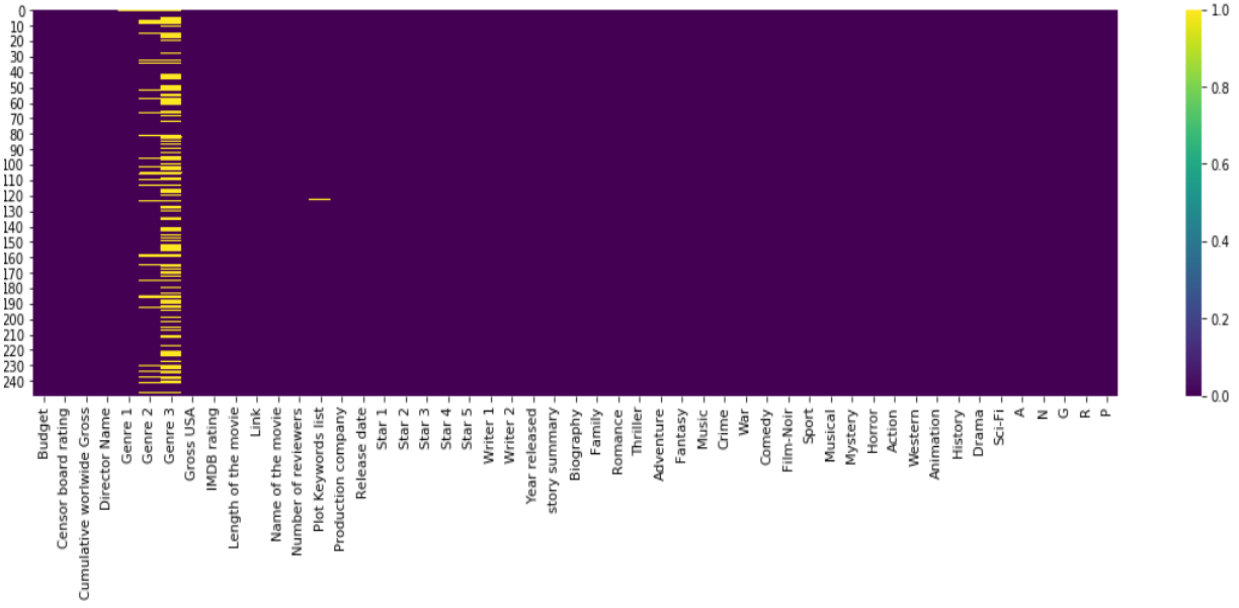
1. DATA SCRAPING:

- Used python based selenium webdriver script to harvest data for the given parameters.
- Collected the harvested data into a mongodb database (json based db).
- Exported the csv format data and imported into code processing.

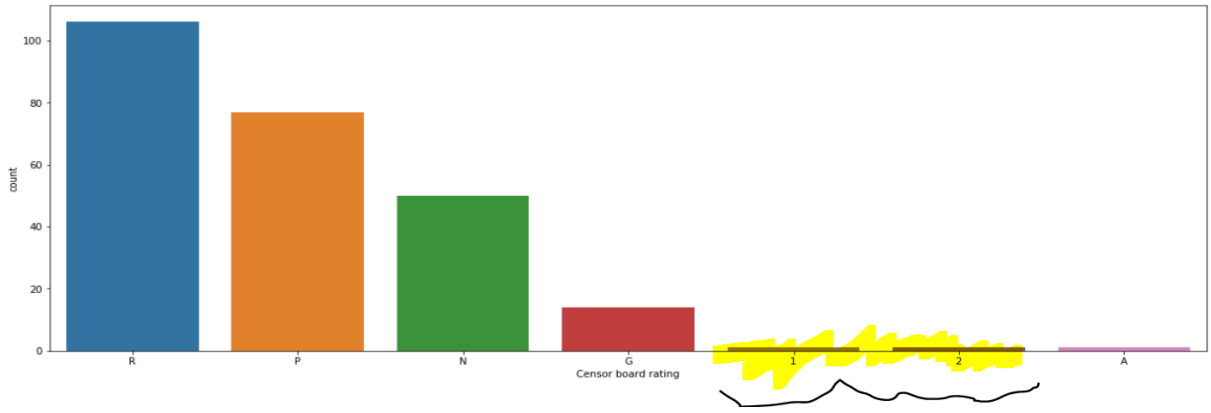
1.1 DATA CLEANING AND PREPROCESSING:

- Removed the Null/absent data values from the following essential features:
  - Budget
  - Cumulative worlwide Gross
  - Gross USA





- Textual cleaning in the following features:
  - Writer 2
  - Censor board rating
  - Number of Reviewers

- Rectification/Replacement in the following feature:
  - Censor board rating
  - 

- Allocating appropriate data types to following essential features:
  - Budget (float)
  - Cumulative worlwide Gross (float)
  - Gross USA (float)
  - IMDB rating (float)
  - Number of reviewers (int)

- Feature extraction/encoding as separate features for the following:
  - Genre 1
  - Genre 2
  - Genre 3
  - Censor board rating
  - ```
    imdb.iloc[:, 25:].head()
    ```

	Biography	Family	Romance	Thriller	Adventure	Fantasy	Music	Crime	War	Comedy	...	Western	Animation	History	Drama	Sci-Fi	A	N	G	R	P
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	1	0	0	0	0	1	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	1	0
2	0	0	0	0	0	0	0	1	0	0	...	0	0	0	1	0	0	0	0	0	1
3	1	0	0	0	0	0	0	0	0	0	...	0	0	1	1	0	0	0	0	1	0
4	0	0	0	0	1	1	0	0	0	0	...	0	0	0	1	0	0	0	0	0	1

  -
    ```

2. MODEL BUILDING:

- Input features:
  - 'Budget', 'Cumulative worlwide Gross',
  - 'Gross USA',
  - 'Number of reviewers',
  - 'Biography', 'Family', 'Romance', 'Thriller', 'Adventure', 'Fantasy', 'Music',
  - 'Crime', 'War', 'Comedy', 'Film-Noir', 'Sport', 'Musical', 'Mystery',
  - 'Horror', 'Action', 'Western', 'Animation', 'History', 'Drama','Sci-Fi',
  - 'A', 'N', 'G', 'R', 'P'

- To predict:
  - 'IMDB rating'

- Splitting the dataset:
  - 80% train
  - 20% test

- Using Linear Regression model since other models such as Decision Trees and SVM tend to over fit on small datasets

- Evaluating the model on test set with MAPE to 1.7%

- Cross Validating (cv=10) the model to find model overfitting or underfitting

- The model has optimal bias variance tradeoff