University of Science and Technology
Communications & Information Engineering Program
CIE 427:Big Data Analytics

مدينة زويل للعلوم والتكنولوجيا
Zewail City of Science and Technology

Final Course Project:

IMDB Analysis

CIE 427

Under Supervision of
Dr. Elsayed Hemayed
Eng. Ahmed Weal

Prepared By

Omar Gaballah	201801697
Ahmed AbdelSalam	201801597
Mazen Hassan	201801897
Ibrahim Hamada	201800739

2022/2023

Table of Contents

Introduction

Our Idea is to analyze the IMDB reviews and predict whether a movie review is helpful using machine learning models and do sentiment analysis on reviews. We want to do extensive data analysis on the movies dataset, which should give us some crucial insights that would help us learn plenty of information about the data and apply machine learning models to predict how much help a review is and sentiment analysis on the reviews.Movie recommender systems help people in narrowing their ranges of which movie to watch. However, people still look for other people's reviews to determine whether they would like to watch this movie or not. We believe that ranking people's reviews can be of great help for websites and viewers.

The Lowest & Highest Rated Movies

1) Lowest Rated Movies:

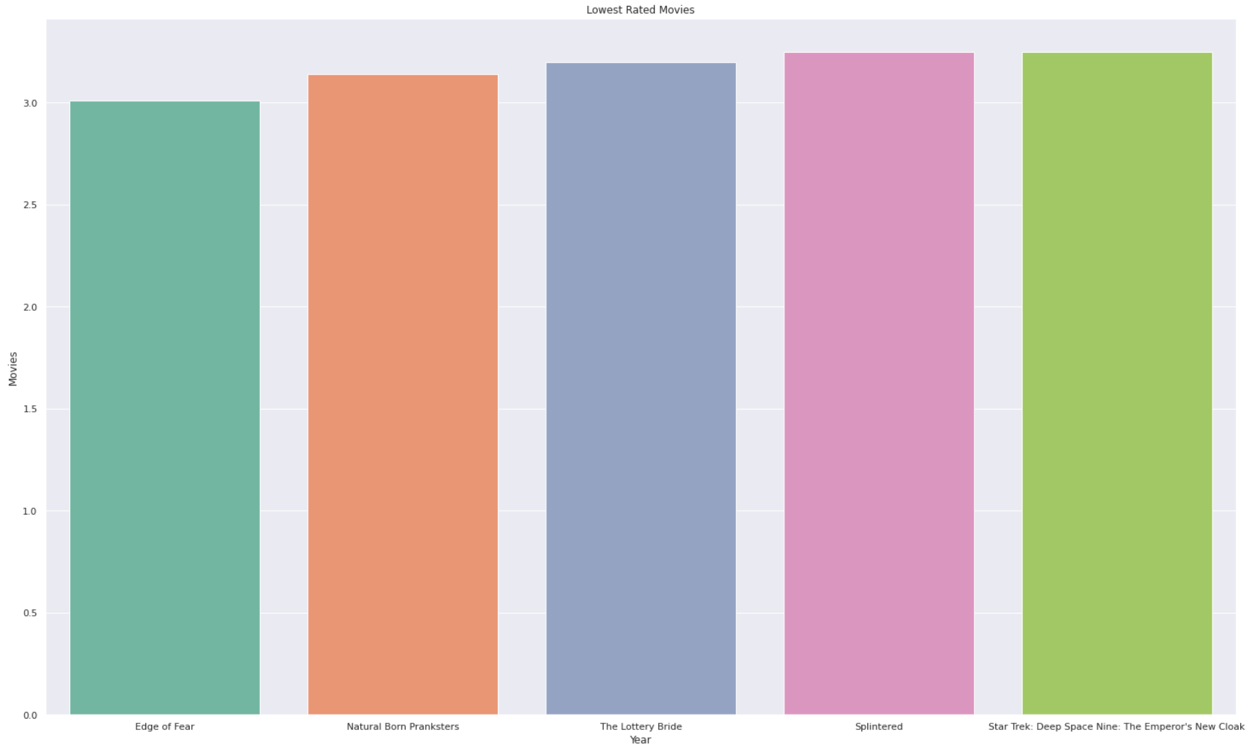

Figure 1. 5 Least Movies Rated.

We found the least 5 rated movies:Edge of Fear,The lottery bride, Star Trek Deep Space, Splintered, Catherine the Great. , we searched for a reference to compare our results with and we searched on IMDB website to compare and found similar results.

2) Highest Rated Movies:

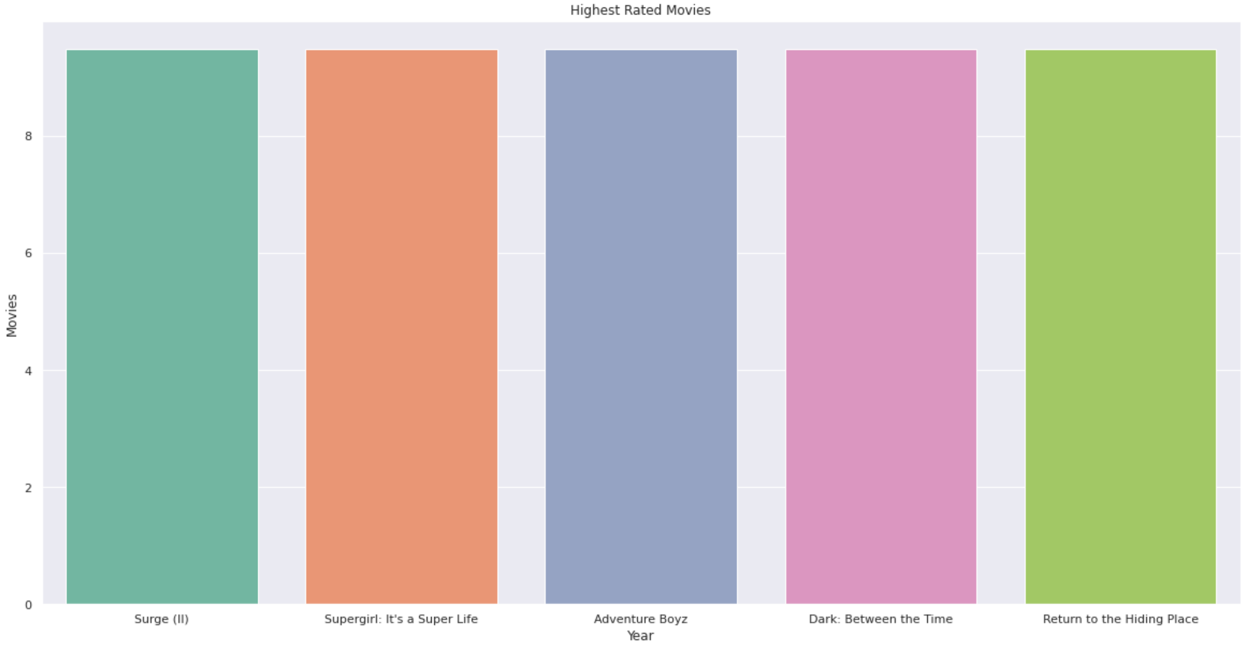

Figure 2. 6 Highest Movies Rated.

We found the most 5 rated movies: Surge (2) ,Supergirl, Adventure Boyz, Dark Between The Time, Return to the Hiding Place. , we searched for a reference to compare our results with and we searched on IMDB website to compare and found similar results.

3) Lowest Rated Tv:

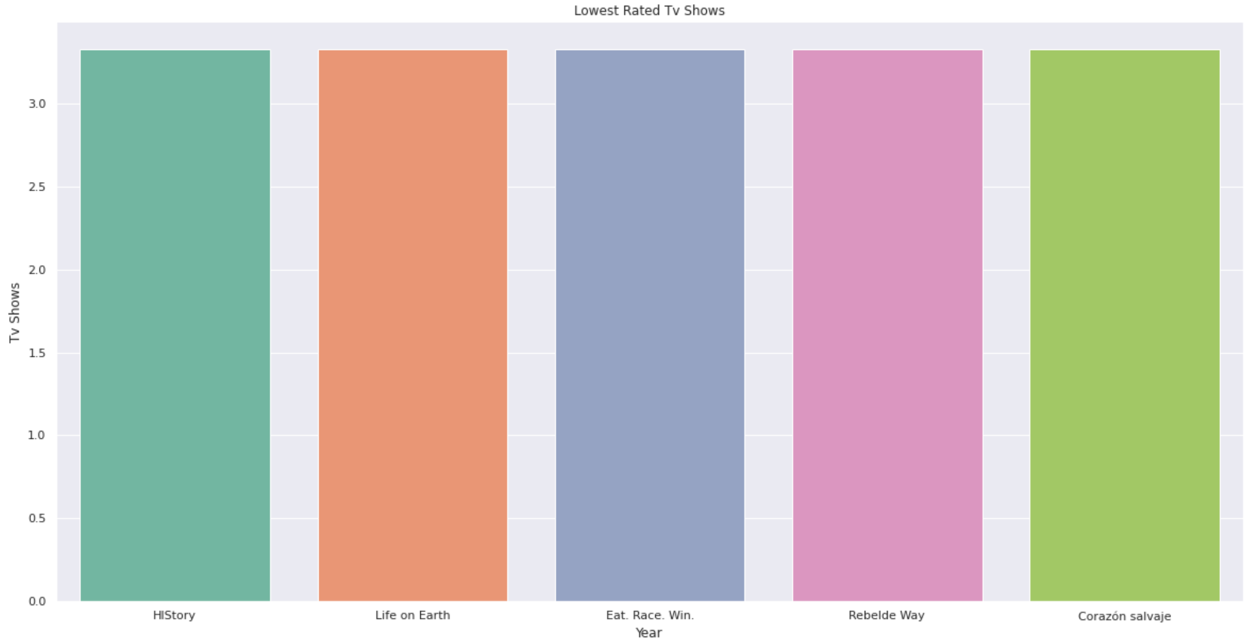

Figure 3. 5 Lowest Tv Shows Rated.

We found the least 5 rated Tv Shows :HIStory, Life on earth, Eat. Race. Win Rebelde way and Corazon Salvaje. We searched for a reference to compare our results with and we searched on IMDB website to compare and found similar results.

4) Highest Rated Tv:

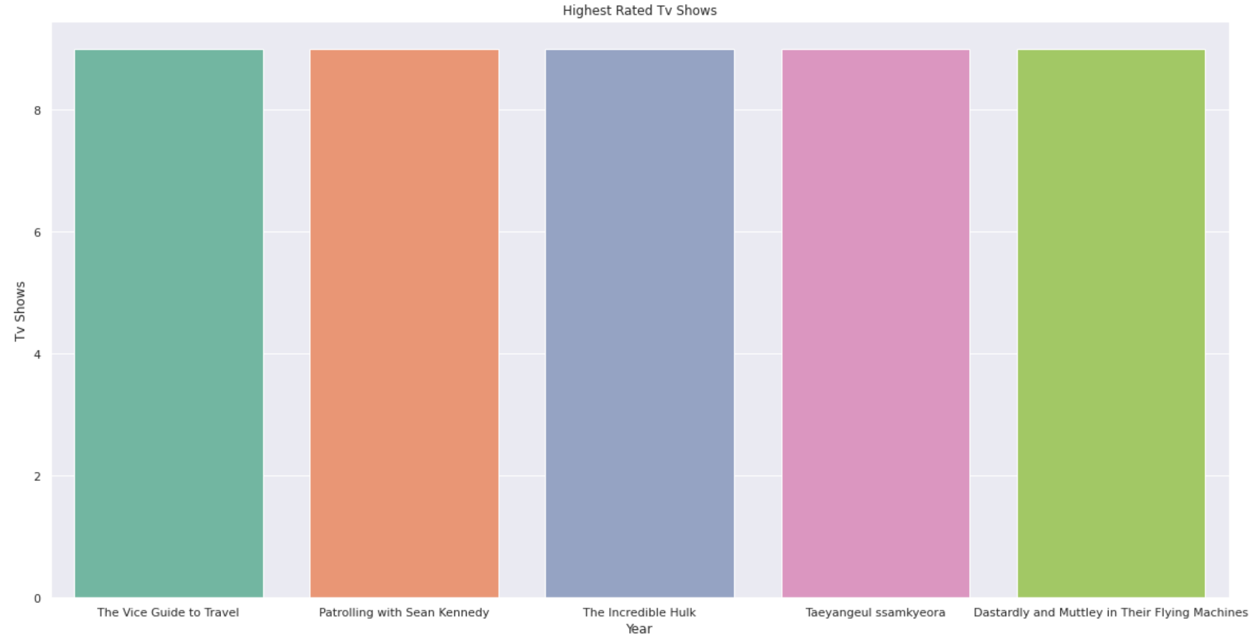

Figure 4. 5 Highest Tv Shows Rated.

We found the least 5 rated Tv Shows :The Vice Guide to Travel, Patrolling with Sean Kennedy, The Incredible Hulk, Taeyangeul Ssamkyeora, Dastardly and Muttley in their flying machines. We searched for a reference to compare our results with and we searched on IMDB website to compare and found similar results.

Number of Spoiled Reviews

1) The Count of Spoilers in Reviews:

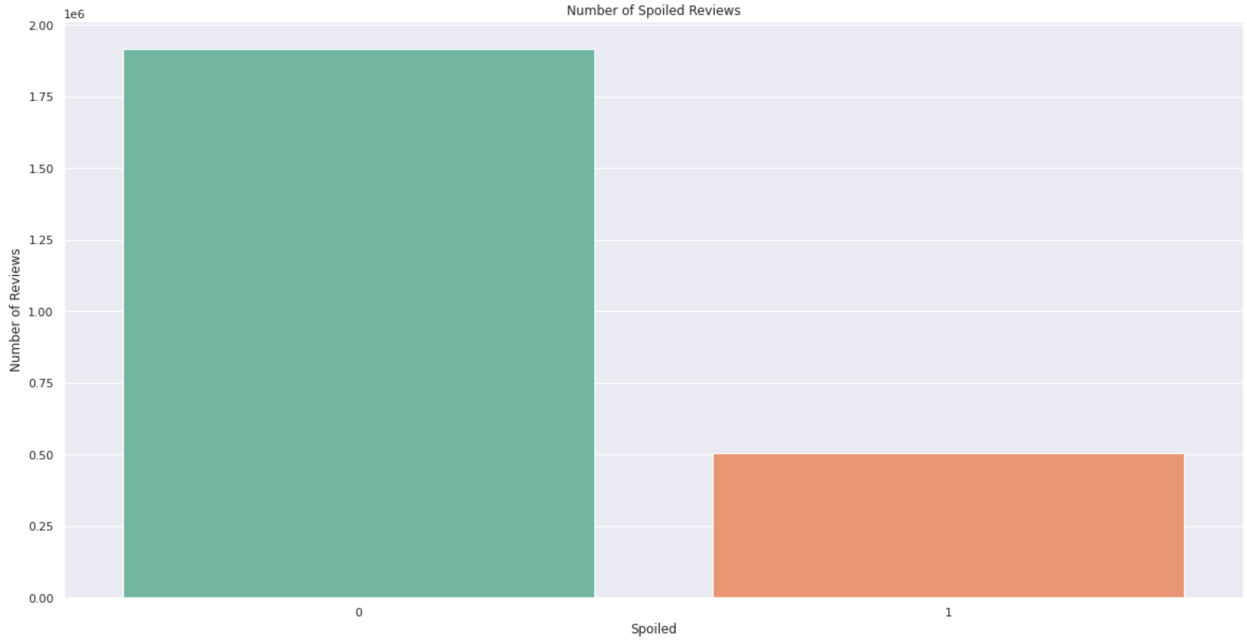

Figure 5. Countplot of Spoilers

We found that the average Spoilers is around 25% of the comments, therefore it might affect the other people before watching the actual Movie.

The highest number of helpful Reviews

1) Top 5 Helpful Tv Shows Reviews:

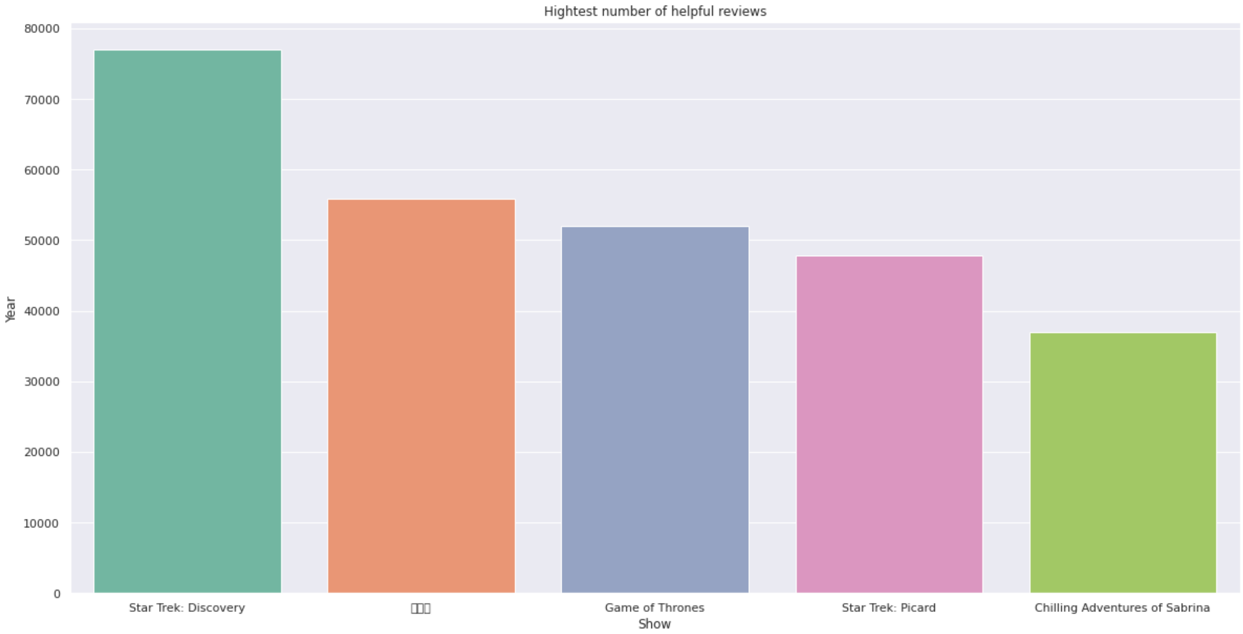

Figure 6. Top 5 Helpful Tv Shows Reviews.

We found that the 5 most helpful reviews for Tv shows are: Star Trek which indicates its popularity and with a highly helpful community, as well as Game of Thrones Show.

2) Top 5 Helpful Movies Reviews:

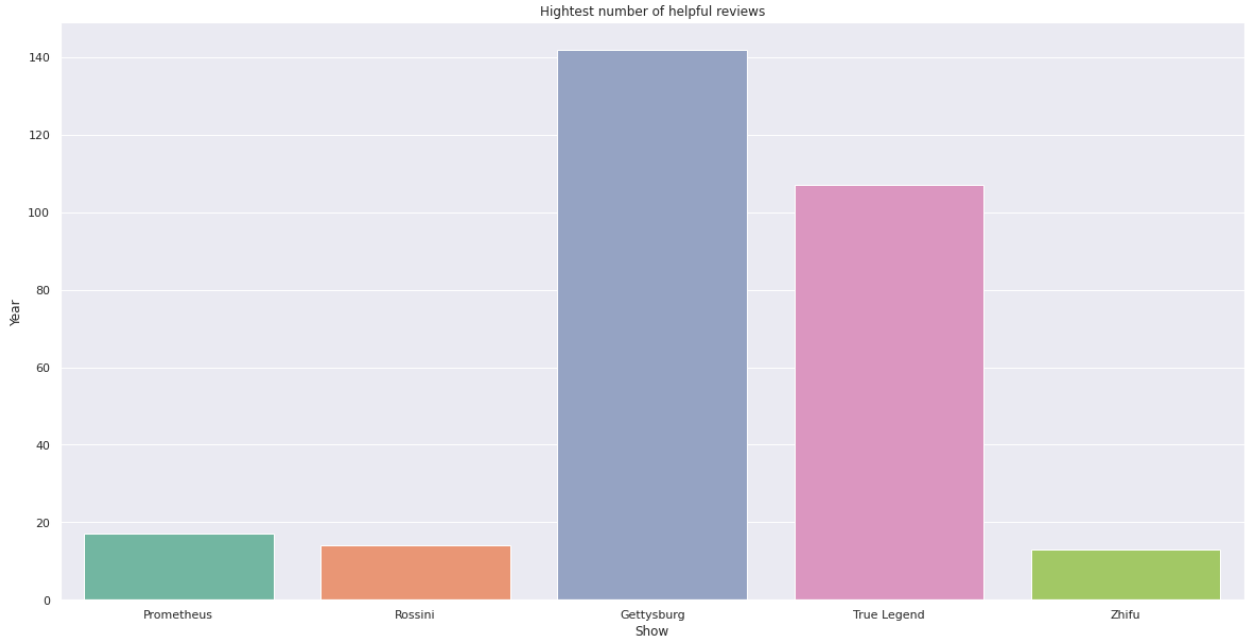

Figure 7. Top 5 Helpful Movies Reviews.

We found that the 5 most helpful reviews for Tv shows are: Gettysburg Show and True Legend Which could indicate its popularity

The Most Reviews Movies & Tv Shows

1) <u>Counts of the most Top 5 Reviewed Movies & Series</u>

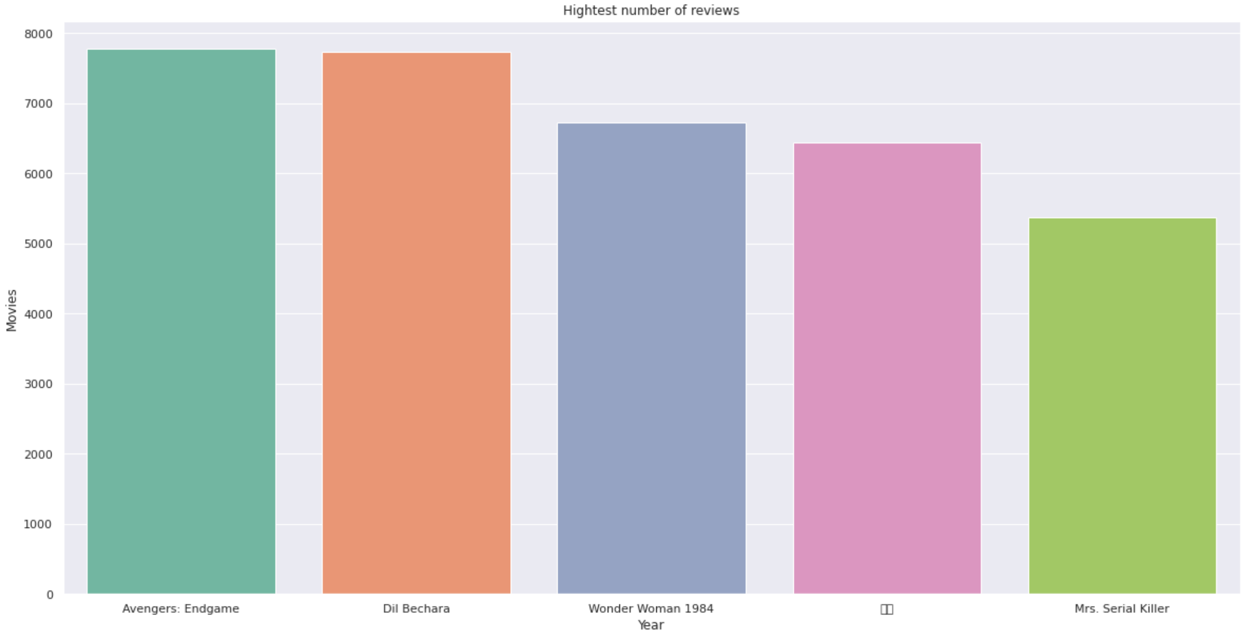

Figure 8. Top 5 Reviewed movies

We found that there is No Correlation between number of reviews and the average rating of a movie, as we can’t find any of these top reviewed movies, they’re not in top rated movies, as Avengers: Endgame has ~8/10 rating.

2) Counts of the most Top 5 Reviewed Series

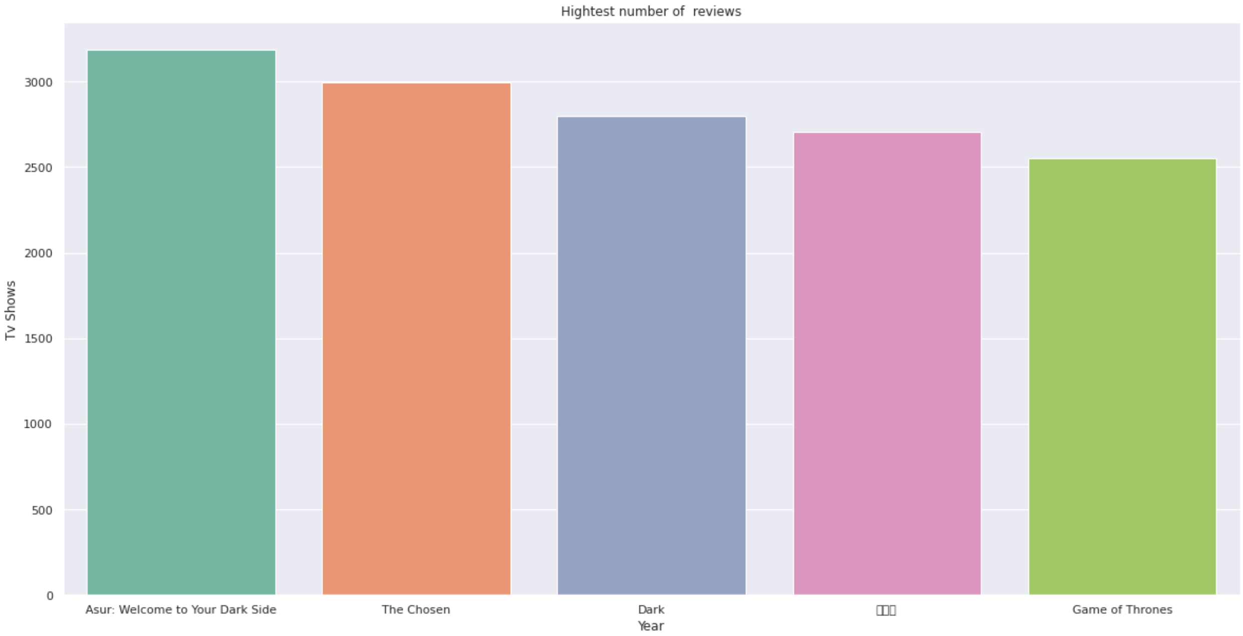

Figure 9. Top 5 Reviewed Series

We found that the most reviewed series were Asur and the chosen with Dark and Game of thrones present which could be an indicator that as the propurity increases the number of reviews for a series increases as well.

The Highest Rate Show in Each Year

1) Top Shows in Each Year

We found out the Top 5 Rate Show in Each Year: Cobra Kai:in 2021, Surge (II): in 2020, Adventure Boyz in 2019, Tijuana Jackson: Purpose Over Prison in 2018, and Rockaway in 2017.

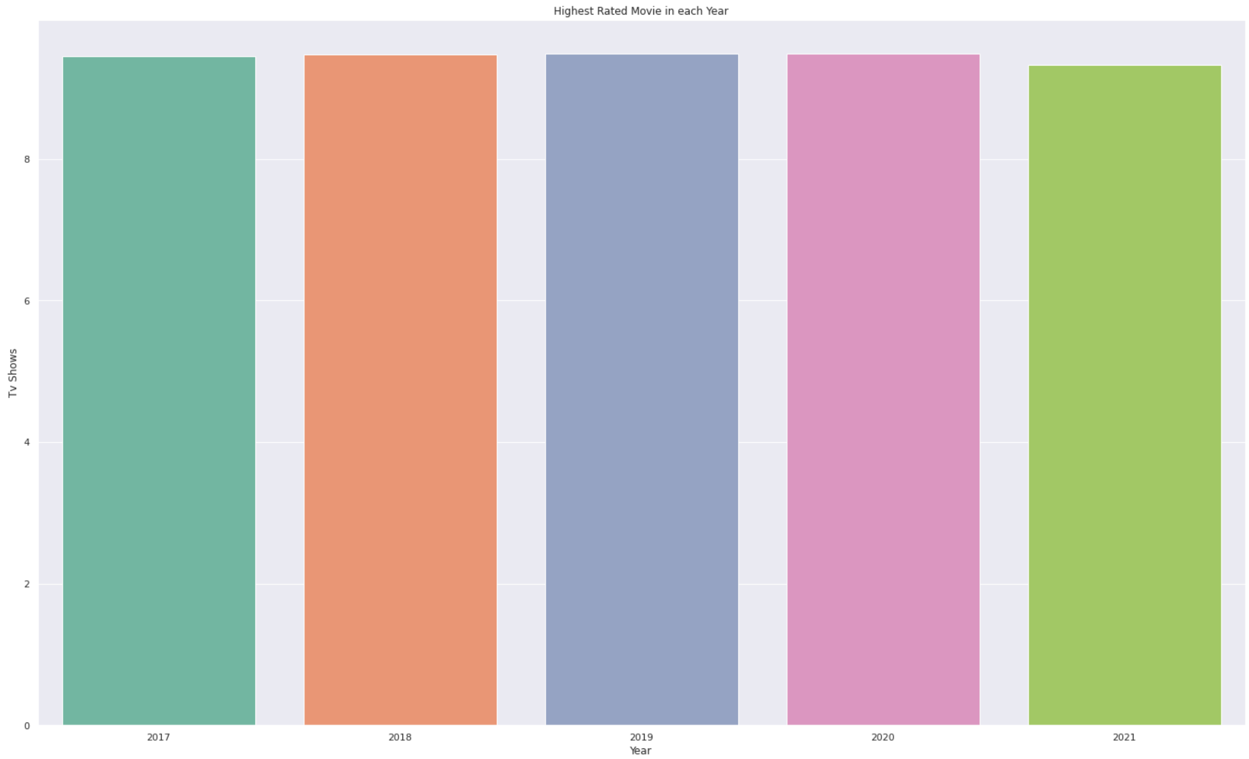

Figure 10. Highest rated Shows in each year

We can see the Highest rate was lower in 2021 than previous years, and this can be because of the high hopes after the global lockdown, or the slightly worse version of the media industry as it got effect by the lockdown at that time.

Number of Shows released Each Year

1) Number of Movies in Each Year

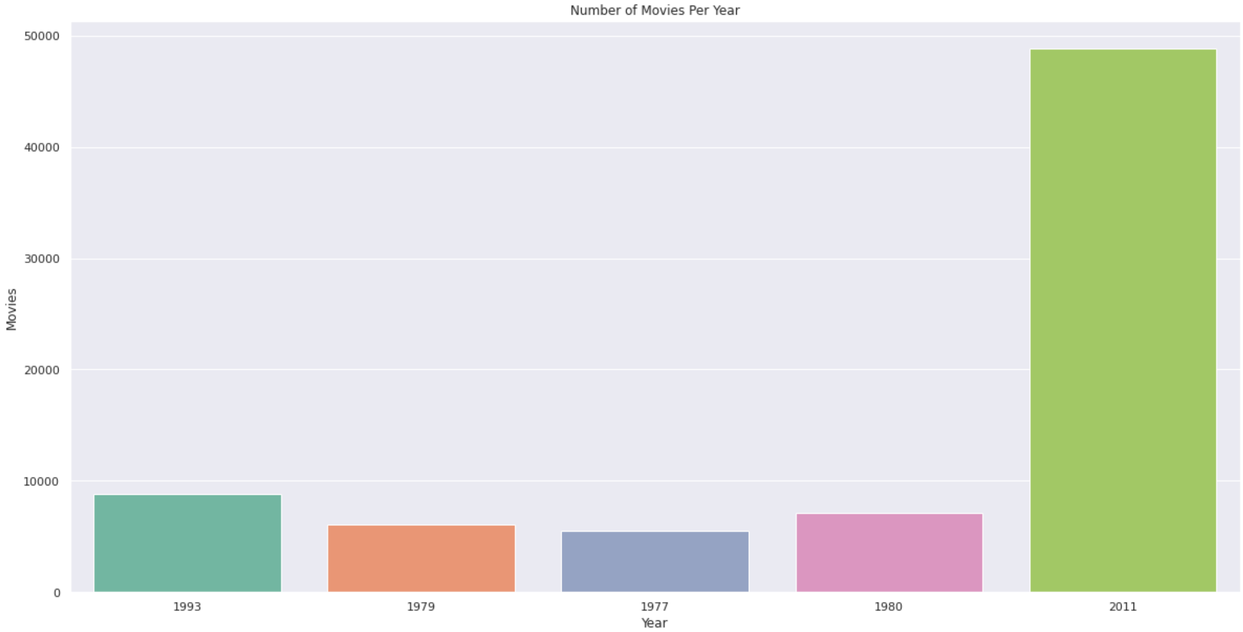

Figure 11. Number of movies per year

We can indicate from the graph that as the years increase the number of movies being made increases which could be as the number of people interested in movies increases and the money that is invested increases as well.

2) Number of Tv Shows in Each Year

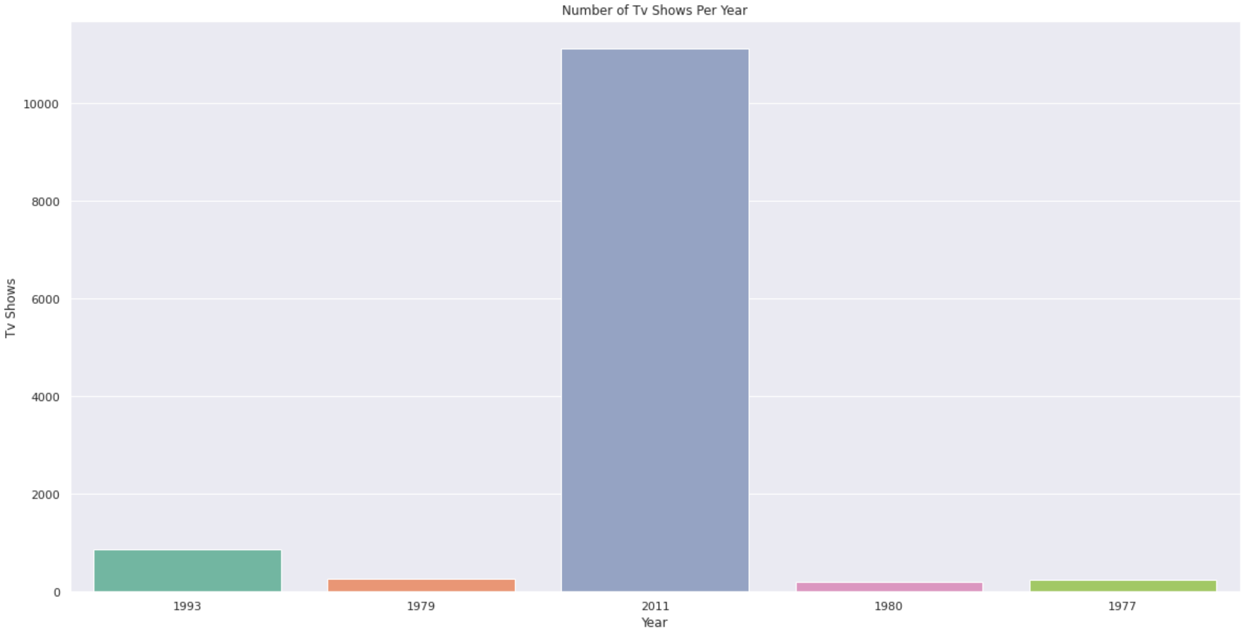

Figure 12. Number of Tv Shows per year

We can indicate from the graph that as the years increase the number of Tv Shows being made increases which could be as the number of people interested in Tv Shows increases and the money that is invested increases as well.

Machine Learning Models

Machine learning models can be useful with many applications with advantage of IMDB Dataset, Such as:

1. Sentiment Analysis: Machine learning models can be used to analyze the sentiment of IMDB movie reviews and identify emotion-based trends among different films and genres.

2. Recommendation System: Machine learning models could be used to generate personalized recommendations for users based on their ratings and reviews of films on IMDB.

3. Movie Search Engine Optimization: Machine learning models can be used to optimize the search results on IMDB related to a user's query in order to increase website visits and engagement with the platform.

4. Box Office Prediction: Machine learning models can be used to predict the box office success of films by analyzing data from IMDB reviews, ratings, financial information and other sources across different countries/regions.

In this part, we used SparkML to apply some ML model on the dataset in order to predict some important information. We firstly chose the columns of interest to be predicted after training the model on the given dataset. As shown above, the dataset has 2 interesting features: Rating, and Spoiler_tag. We decided to set the previously mentioned columns as the output of the model. We used the Logistic Regression Model, the Naive Bayes Model, and the Support Vector Classifier (SVC) to predict the spoiler_tag given a particular input review. Accordingly, we tried to perform sentiment analysis by using a set of labeled data on movie/series reviews.

Results of the classification models are summarized in the following table:

Model	Logistic Regression	Naïve Bayes	Support Vector Machine
Accuracy	0.7919	0.7840	0.8011
F1-score	0.7181	0.7821	0.7851

Table 1. Results of the models.

We got the following results which indicate the edge of SVM on other Models in classifying the sentiment analysis of the review, therefore it would be the most suitable model option for us.

In order to predict the value (1-10) of the rating given a specific review, we used Linear Regression, Decision Tree Regressor, and GBT Regressor. After training the Linear Regression Model, we obtained an R-squared value of nearly 0, which indicates no linear relationship between the review and ratings. So, we tried the other models.

Results of the regression models are summarized in the following table:

Model	Decision Tree Regressor	GBT Regressor
RMSE	0.21	0.17

Table 2. Results of the models.

Conclusion:

Machine learning and data analysis can be helpful in further applications as:

1. Automated movie recommendation: to identify similar movies based on movie attributes such as director, cast and/or genre. This type of automated suggestion system would enable IMDB users to discover new films they may enjoy.

2. Movie ratings prediction: Regression algorithms like linear regression and logistic regression can be used to predict ratings for upcoming movies, based on users' history of giving ratings for other IMDb movies. Such predictions will help film releasing companies to set their expectations and plan the publicity accordingly.

3. AI-assisted movie reviewing: By analyzing comments from other reviewers and past content from movie critics, an AI-powered model can suggest high quality reviews which contain useful insights about a given film that complement specific insights from humans writing the copy.

Therefore, with more data, analysis and model optimization, we can perform more powerful and further applications.