CS5830/6830 Project 3 Report

Hari Chandana Kotnani and Jacob Johns

Introduction

The TV industry is a $300B market that is expected to grow for years to come. Our analyses, which were conducted using Python and the Pandas data science package, provide valuable insights to TV producers. Specifically, we analyzed data from the top 1000 TV shows on IMDb to identify past and present market trends, examine the effect of the Covid-19 pandemic on the TV industry, and gauge public perception of different shows. This information will enable producers to make informed decisions about which shows to accept, continue, or cut, and to optimize current appeals or strategize by utilizing underutilized market segments. Overall, our analysis provides a comprehensive understanding of the TV industry, which can help studios to succeed in a highly competitive market.

Presentation Link GitHub Link

Dataset

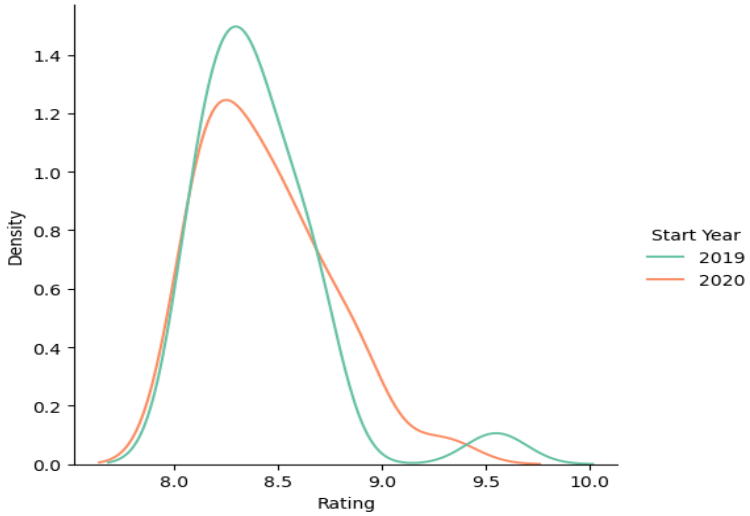
Our data is scraped from IMDB's list of the top 1000 TV shows based on user rankings which includes the average user rating, the number of votes, and genre of each show, as well as its ranking and maturity certificate. It is important to note that unlike other web services, IMDB uses public (ie. not a professional critic) ratings. Thus, the data capture the general public's preferences, rather than a small group of individuals. Since only the top 1000 TV shows are reported, the lowest average rating is 8.1. For reproducibility, the data is stored in a csv file, but code is included to update it as needed.

Analysis Technique

In this project, there are four analysis techniques to gain insights into the television industry. To measure the impact of the Covid-19 pandemic, we compared the user ratings of shows made in 2019 and 2020 in a distribution plot. The potential growth and high return on investment in the TV industry are best seen by an ever-increasing line plot. In order to show the power that a fanbase has, bar graphs are juxtaposed. A figure of box plots on adult vs. non-adult content is also presented, where the differences between the two are highlighted by radical size differences. These analyses were facilitated by functions from the Pandas library and plotting tools from the Matlab and Seaborn libraries. The chosen techniques were appropriate for the data and domain as they allowed us to effectively visualize and compare different aspects of the television industry, including the impact of the pandemic, fanbase power, and content differences.

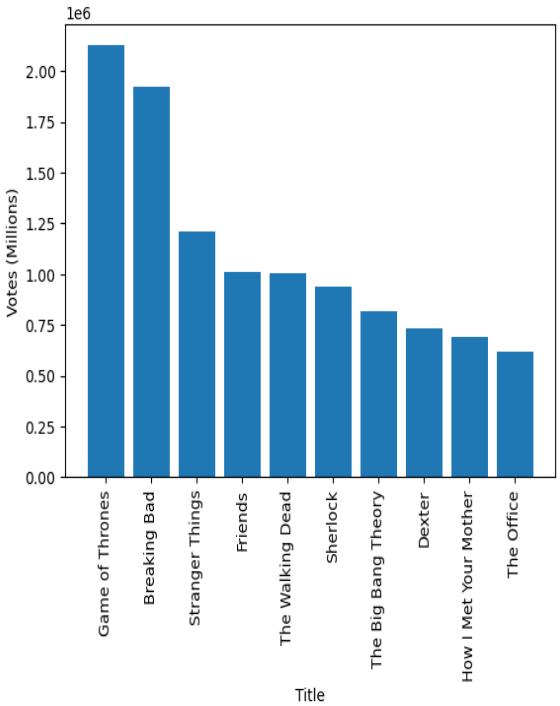
Results

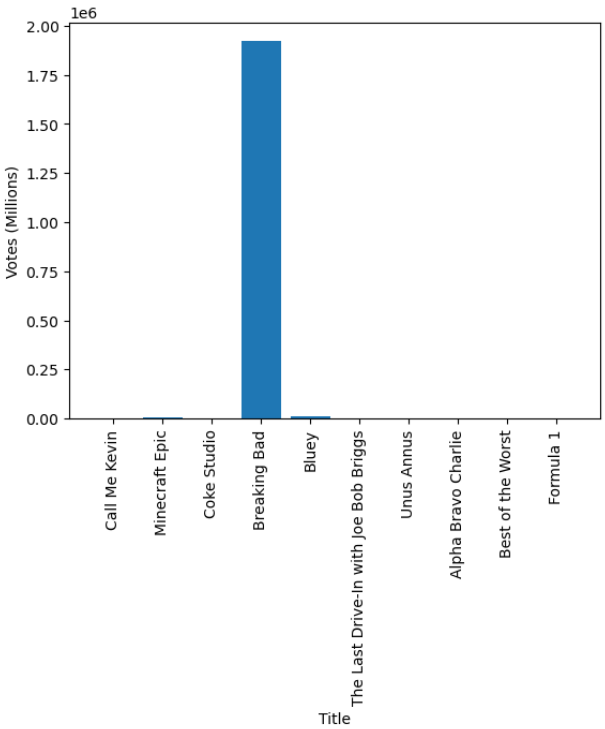
We first started looking at the effect that the Covid-19 pandemic had on the TV industry. Specifically, we wanted to know if the pandemic had affected the quality of television produced. We found no evidence to support that claim. The t-test, which analyzes the probability of whether two groups are different, was used to investigate whether there was a fundamental difference in the quality of TV production between the years 2019 and 2020, with and without the impact of the Covid-19 pandemic. The results of the t-test showed that the t-statistic was -0.71 and the corresponding p-value was 0.47, indicating that there was no statistically significant difference in TV production quality between the two years. Therefore, we can conclude that there is no strong evidence to suggest that the Covid-19 pandemic had a significant impact on the quality of television produced during the years 2019 and 2020.



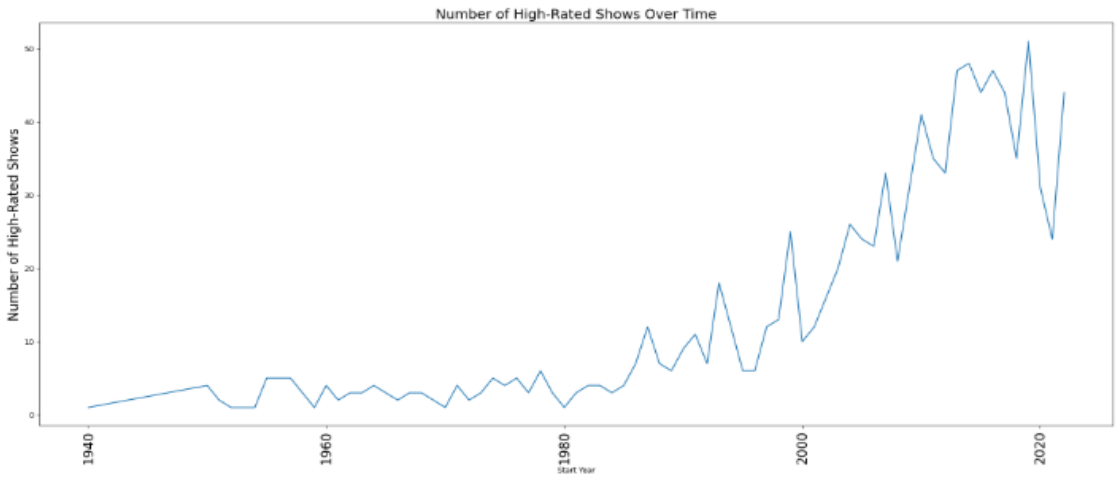
We then looked how a show's popularity differed from the user ratings. Popularity was determined by the number of votes a show received. Our analysis revealed that, out of the top 10 TV shows by user rating, only Breaking Bad had more than 4000 votes. It is the only one to appear on the top 10 TV shows by number of votes also. We surmised that the other top 9 were smaller cult classic shows, with thousands of loyal (and highly praising) viewers. This suggests that many of the other highly-rated shows have smaller, dedicated fan bases.

This information is highly useful to viewers who are seeking recommendations of high-quality but perhaps lesser-known shows to watch. Our findings also suggest that user ratings may be a better indicator of quality than popularity alone. Many of the best-rated shows have relatively few votes compared to more popular but lower-rated shows, indicating that the opinions of highly engaged viewers may be more valuable than those of more casual viewers.

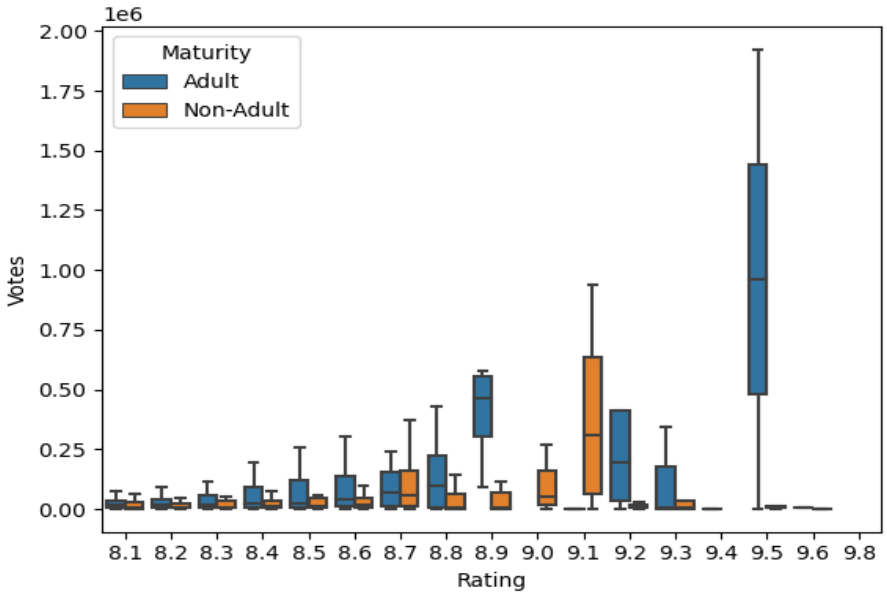




Upon looking at the top 10 shows by voter popularity, we realized that all except Friends had been made in the 21st century. Upon looking at the number of these top 1000 TV shows, we saw that there are more and more as time goes on. Part of this can be attributed to the rise in the number of shows being made, but it is nevertheless interesting to speculate that the future of TV shows has no signs of slowing down. This information helps to paint a picture of the evolution of the TV industry and its growing popularity, potentially indicating that the future of TV shows is bright and that there may be continued growth in the industry.



The analysis presented here is based on the hypothesis that the presence of adult content in a television show can be a contributing factor to its success. Noting that Breaking Bad by measure of popularity and rating was very successful, we thought that part of that success could be attributed to the adult content of the show. Accordingly, we looked at the distribution of each rating level based on if the show had adult content or not. In nearly all rating levels, adult shows were wildly more popular than their kid friendly counterparts. This may be because adults are more likely to vote on IMDB (and it is in fact against their terms and conditions to have anyone under the age of 13 to vote).



Technical

In order to properly analyze this data, we first had to obtain it. This was done by scraping it, processing the HTML with another python package named BeautifulSoup, and then a lot of string formatting, mostly to strip whitespace and punctuation, finished up by type casting. There were about 75 entries that lacked a rating certificate, so they were included in all analyses except the final one. To prepare the dataset for analysis, we conducted additional data wrangling, such as removing duplicates and filling missing values, to ensure that the data was consistent and accurate. We used a variety of analysis techniques, from distribution plots to line plots, and bar plots to box plots. Each one was chosen based on the data which we were analyzing. For example, we used a distribution plot to highlight the difference (or rather lack thereof) in the distribution of ratings pre and post pandemic. A line plot clearly demonstrates the increase of the number of highly rated films as time progresses.

During the analysis process, we encountered some challenges, such as dealing with outliers and missing data. We addressed these issues by adjusting our analysis methods and techniques. The project evolved from social networking to brand sentiment analysis until it finally settled here. As we studied the data, we found relatively little to report on. Aside from the increase in the number of highly rated shows created, air dates weren't found to have much connection with anything else. Instead of only taking the top 1000 TV shows, it may have been more beneficial to analyze the last 1000 shows made. Overall, the chosen analysis techniques were appropriate for the dataset and research goals, providing valuable insights into the top 1000 TV shows of all time.