**INTRODUCTION**

A longstanding pillar of the American experience is the excessive consumption of films and exuberant sharing of criticism with anyone who will listen. In the early days of film, viewers relied on the analysis of critics to decide whether or not to spend time and money on a new flick. In the era of the internet, critics are no longer the only ones who can share judgement - sites like RottenTomatoes.com and IMDB.com aggregate ratings of both critics and the public. While many analysts are interested in how the ready proliferation of reviews affect box office performance, our team sought to explore how the reviews of the public and accredited critics differ across time, genre, MPAA rating etc.

Using Python, we proceeded to collect, clean, query, analyze and visualize data listing approximately nine-thousand movies. This process and its findings are described in detail below:

**THE DATA**

Our first order of business was to collect and clean a robust dataset that we could rely upon for multiple analyses. The data was sourced from a site called MetaCritic.com which aggregates reviews for movies, games, TV shows and music. We chose this site over several alternatives (i.e. those listed in the introduction) since it is more easily accessed from the perspective of a web scraper and it has stringent rules on which reviews it aggregates. From the site we can learn more about this data’s quality:

“Why don't you have 97 reviews for every movie like those other websites do?”

Several other websites that provide links to movie reviews have weighed the quantity vs. quality issue and come out in favor of quantity. These sites typically include links to as many reviews as there are available on the net, but the quality of many is inconsistent at best. In addition, there is such a thing as too much information, and statistically, once we include a certain number of reviews in our calculations, adding additional reviews will not change the overall METASCORE much in one direction or another.”

One of our concerns in data sourcing was scale; we wanted to ensure that we had a large enough data set to draw meaningful conclusions. Knowing that outlier exclusions and other limitations would ultimately trim down our set, we proceeded to pull every single review listed on the website.

The web-scraper was designed to adhere to the RESTful web framework and systematically work its way through the MetaCritic catalog. Its modules were designed to be easily de-bugged and loosely coupled. We relied mostly on the BeautifulSoup and Requests packages to request and process raw HTML feedback into a flat dataset. The scraper is described below:

Read Page Module: Sends a URL request and loads the returned HTML into a parseable soup format. Used multiple times throughout script.

Title Getter Module: Works through all 93 pages of the following URL http://www.metacritic.com/browse/movies/score/metascore/all/filtered?sort

a. Identifies all “a href” links on a page that contain “/movie/“ - these URLs can be used to find a movie’s review page.

b. Output all movies to a pipe delimited text file

Extract Movie Deets Module: Takes a HTML page response for a movie detail page and extracts fields of interest. Designed using multiple try-except blocks out of necessity since some fields might not exist for a given movie (i.e. a movie may not have an MPAA rating so that HTML block would not exist and throw an error.). Testing was performed to make sure the failure to collect information would feedback an “unknown” entry.

critic\_rating\_val, critic\_rating\_num, user\_rating, user\_rating\_num, mpaa\_rating, genres, movie\_date, movie\_name, description, current\_url

Movie Extraction Module: For a list of movie URLs it reads the page, extracts the details and creates a pipe delimited text file of the results.

Notify Module: Sends an email when parsing process is completed. This module is a convenience feature and goes beyond the requirements outlined in the rubric.

**PRE-PROCESSING**

The data procurement process took nearly 4 hours to run and write to text. Once this process was completed, our team began analysis by importing the file into a data frame structure using Python’s data analysis package PANDAS. We performed the following pre-processing in order to limit the data set to movies of interest and remove outliers. Naturally, since the data contains a number of qualitative, quantitative and categorical variables we queried the data before applying these preprocessing limitations:

1. Strip punctuation from the movie descriptions

2. Remove user ratings that were “Unknown” or “TBD” as these ratings did not have a sufficient number of ratings to be included in the dataset

3. Remove MPAA ratings that were “Unknown”, “Not Rated”, or related to TV-shows

4. Change the column types so that they can be used for analysis (i.e. character -> numeric)

5. Created subset files that condensed the genre and MPAA ratings into groups (i.e. R and NC-17 movies get combined to the R group)

**GOALS**

1. Understand the difference between critic and user ratings – which is higher? Our hypothesis is that user ratings will on average be higher than critic ratings.

2. Determine which genres result in high score (for critic and user ratings) and which genres result in low scores (for critic and user ratings). Our hypothesis is that family or action movies will be the higher rated genres, and horror films will be the lower rated genre.

3. Determine whether or not scores differed for different MPAA ratings. Our hypothesis is that PG13 movies will be the higher rated MPAA group, and that R movies will be the lower rated MPAA group.

4. Out of curiosity, the last question we had was what were the top 5 films for the past 20 years – and were they the movies we expected?

**RESULTS**

The first query that we were interested in running was a simple line-plot that shows the critic and user scores over our date range. The date range of our raw data goes from 1953 to 2016, but we limited the range of our analysis to the last 20 years (1996 to 2016) in an effort to include recent movies with high numbers of reviews. The result of this line plot showed that user scores have decreased over the last 20 years, while critic scores have fluctuated between average ratings of 5.16 (year 2000) and 6.02 (year 1996).

This led us to wonder about the distribution of critic and user ratings – and whether they differed. Our team used the KDE plot in the PANDAS package to plot the density plots for critic ratings and user ratings. What we found is that user ratings are extremely left skewed, with a large portion of the ratings falling between 6 and 9 – and that critic ratings are more symmetrical, with the mean falling around 6.

The next query that our team looked into was the average ratings by genre. The initial data that we scraped had a high number of distinct genres, so for the purpose of this analysis our team narrowed down the genres to only 6 options – action, comedy, drama, romance, family, and horror. We found that 92% of movies fell into one of these 6 categories, with films typically falling into more than one. Once we had this subset we calculated the average rating for movies that fell into these categories and found that Drama films are consistently rated the highest (with an average critic rating of 5.84 and an average user rating of 7.09) and Horror films are consistently rated the lowest (with an average critic rating of 4.68 and an average user rating of 6.36).

At this point, the team wanted to understand the breakdown of movies by MPAA rating. In order to do this, we grouped certain ratings together (i.e. R and NC-17 get grouped as R; G, GP, and PG get grouped as G, etc.) Once we had these groupings we calculated the average rating for movies that fell into each category and found that X-rated films had the highest ratings (with an average critic rating of 6.72 and an average user rating of 8.22) and PG13-rated films had the lowest ratings (with an average critic rating of 5.17 and an average user rating of 6.72). The X-rated subgroup only contained 5 X-rated films, which is a much smaller sample than the PG13-rated group (1609 films) as well as the others, which skewed the average rating. One thing that should be noted is that these X-rated films are not the type of films you expect them to be – they are actually just the X-rated versions of certain R-rated films (i.e. Scarface, A Clockwork Orange, etc.)

The last question our group was curious about was the top 5 movies for each year – again, the date range we use for year is 1996 to 2016 to include recent movies with high numbers of reviews. The results of that table are in the appendix.

**CONCLUSIONS**

Below are the conclusions we found for the four goals we outlined in a previous section.

1. As noted in the results section, the distributions of critic and user ratings show that users tend to rate films higher than critics. We also found that over the past 20 years the user scores have been decreasing over time, while critic scores have remained somewhat stable. We believe the decrease seen in user score is potentially due to the fact that critic scores for older movies are recorded at the time (shortly after release), where user scores could be recorded much later – to the point where movies could mature and be considered as ‘classics’.

2. Our hypothesis was slightly correct in that horror movies were the lowest rated genre, but we were surprised to find that drama movies were the highest rated genre as we expected family or action to take that spot. We believe this is due to the fact that each movie can be tagged in multiple genre categories, and drama is a genre category that is applicable to a majority of movies.

3. Shockingly, PG13 films were the lowest rated MPAA group instead of the highest – which we initially hypothesized. Additionally, the highest rated MPAA group was X-rated films. However, due to their small sample size (5 movies) we believe this to be an inaccurate representation of their average due to the fact that each movie has a significant weight on the average.

4. We have answered our question by producing a list of the top 5 films by year (for the last 20 years) in the Appendix section of this report. In the event of a rainy day, we now know where to look to find a highly-rated classic movie to watch!

For further research on this topic we would want to be able to include a larger range of variables to allow for a greater understanding of what influences critics’ and users’ ratings. Variables which we would be interested in finding information on include: the director, lead actor and actress, filming location, if the film was a sequel or prequel, and the production and distribution companies of the film. We would also want to look at if there was a correlation between the ratings of the films and whether the film won or was nominated for any major awards. A final piece of information we would wish to work with is whether the film was originally shown at a film festival or if it was released directly to theaters.

As far as advancing our program, we would have like to work with better graphing libraries which we were unable to work with within the time frame we had. One example of these is the Bokeh library. We feel like this would have been a useful library to use due to the more aesthetically pleasing to the viewer.

**Appendix**

