Jacob DiStefano – jdiste5
Levi Hererra – levih2

Project Proposal:

1. **Project goal, the question you are trying to answer. If you have a hypothesis clearly state it. If you want to make a prediction describe what you are trying to predict, how do you plan to measure your success, etc.**

   In this project we would like to analyze ratings from rottentomatoes.com. We would like to see if there is any relation between ratings and other features such as; streaming service, runtime, ratings, genre, actors, and director.
   We plan to measure these relationships by creating plots and attempting to find relationships, if they exist. We will use pandas and seaborn for data preparation and visualization.
   We may also try to analzye rating for an individual actor or director, and see how their work is rated between genres or over time.

2. **How do you plan to collect your data?**

   We plan to use BeautifulSoup for web scraping. We've looked into using the official rottentomatoes.com API, but it turns out they don't support unofficial projects and the API costs $60,000/year to use.

3. **What challenges are you anticipating?**

   There are a lot of movies on Rotten Tomatoes, and they all aren't contained on a central list. So it may be difficult to get all of the information we need.

4. **What importance your project will have on the related field (basically justify why you picked what you picked)**

   This project may be useful to filmmakers to determine what drives ratings, or at least Rotten Tomatoes ratings. It will also be useful to critics to show them their biases and maybe improve fairness in film critiquing.

Progress Report:

1. **An introduction part to your data:**
   1. **Data spec: describe your data. Include the format and any assumptions about your data, size of the dataset**

This dataset is in csv format, each row is an individual film from rottentomatoes.com. It contains the following information for each film; Tomatometer Score, Audience Score, Release Date, Runtime, Rating, Genre, Cast. Studio, and Director

2. **A link to your full data in downloadable form, you can keep your data on Google Drive, Box, DropBox, GitHub, or personal website**

   The dataset can be found on the project github repo here **vvvvvvv**

   https://github.com/jldistefano/rt_rating_analysis/blob/master/scores/all_scores.csv

3. **A sample of your data (n = 10 – 50)**

title	tomato_score	audience_score	release_date	runtime	rating	genre
Many Adventures of Winnie the Pooh (1977)	100	88	1977-03-10 16:00:0 0-08:00	74	G	['Animation', 'Comedy', 'Kids & Family', 'Musical & Performing Arts']
Toy Story 2 (1999)	100	86	1999-11-23 16:00:0 0-08:00	92	G	['Animation', 'Comedy', 'Kids & Family']
The Odd Couple (1968)	100	89	1967-12-31 16:00:0 0-08:00	105	PG	['Classics', 'Comedy', 'Drama']
Old Yeller (1957)	100	79	1957-12-24 16:00:0 0-08:00	84	G	['Action & Adventure', 'Classics', 'Drama', 'Kids & Family', 'Western']
On a Clear Day You Can See Forever (1970)	100	75	1969-12-31 16:00:0 0-08:00	129	G	['Classics', 'Comedy', 'Drama', 'Musical & Performing Arts', 'Science Fiction & Fantasy']
Darby O'Gill and the Little People (1959)	100	77	1959-06-25 17:00:0 0-07:00	93	G	['Kids & Family', 'Science Fiction & Fantasy']
The Yearling (1946)	100	77	1946-12-17 16:00:0 0-08:00	128	G	['Classics', 'Kids & Family']
My Darling Clementine (1946)	100	86	1946-12-02 16:00:0 0-08:00	97	G	['Action & Adventure', 'Classics', 'Drama', 'Western']
Toy Story (1995)	100	92	1995-11-21 16:00:0 0-08:00	80	G	['Animation', 'Comedy', 'Kids & Family']
That's Entertainment (1973)	100	85	1973-12-30 16:00:0 0-08:00	132	G	['Classics', 'Documentary', 'Musical & Performing Arts', 'Television']

The entire table cannot fit on the page, but I assure you there are more fields.

2. **A report of your data collection process**

    1. **How did you collect your data**

    We first went to the “Browse Movies” page on rottentomatoes.com, and iterated througha all of the pages, collecting the individual page urls at each page. Next iterate through each of the urls, collecting all of the data from each individual page, when the page information was collected, it was then written to a csv file.

    2. **How did you clean your data**

    We cleaned the data as it was being collected, this included removing rating descriptions, saying why the movie got the rating it did. We also removed formatting whitespace on the genre, rating, and release date fields. I had to convert the release date from a string to a datetime object, and determine whether the movie was released in theaters at all.

    3. **Mention any difficulties you faced in the beginning steps**

    The dataset is very large, ~21,000 movies, so the script took a very long time to run and collect all of the movies. I had several test runs on smaller datasets to ensure that the data was being processed correctly. However many pages didn’t load, or the information was incorrectly gathered, causing the movie to be excluded from the dataset.

    After collecting all of the data, it turned out that I didn’t clean the ratings field thoroughly enough, so there are some movies which still have parenthesis or spaces in the ratings field. Which means that some ratings are split between several uncleaned types.

3. **A summary of challenges and observations you have made so far.**

    Rottentomatoes has about 21,000 movies, and it took a very long time to load all of the movies. The first time I ran the scraping script it stopped loading films with a tomato score less than 59. I suspect this is either because the initial movie list had that restriction, or there were too many urls in the python list, and it couldn’t hold any more. So I had to restart the script with a maximum tomato score of 59 in order to get all the data. In total it took about 2.5 hours for the script to collect all of the movie information.

    I haven’t looked at the data too much, but from what I can tell, the movie scores are basically the same between streaming services, so we likely won’t be using streaming service as a variable, as it isn’t very interesting.

4. **A brief mention of your next steps and what you plan to do with your data as you move into the analysis (If you are already in the analysis phase you can mention that as well)**

We will now look at the other features in the dataset, such as runtime and rating, and see how they correlate to rating. We will also have to find a way to efficiently parse the director, cast, and genre fields, as they are stored as python lists.

5. Group member duties

Jacob DiStefano – Project Captain, Data Collection, Data Cleaning, Data Analysis
Levi Hererra – Data Analysis, Data Visualization