What makes a high rating score on IMDb? - Final Project Proposal

TEAM NAME: Team MRP(Movie Rating Predictors)

TEAM MEMBERS:

- Shuibenyang Yuan A14031016
- Meliza Ramos Suarez A12604036
- Sutianyi Wen A13992949
- Froilan Grepo A15714684
- Cristina Ahamad A13869848

DATA SCIENCE QUESTION(S) & HYPOTHESIS:

- Data Science Question:

  According to the data from "*The statistic Portal*", the global box office revenue will reach 50 billion U.S. dollar in 2020 and 19% Americans go to movie theatres once a month. Because of the film industry's prosperity in recent years, there are nearly 2000 movies made worldwide each year. As a group of movie fans, we usually find that there are many boring movies existing in the market and we often need to check the IMDb rating score before buying the ticket. But how are the rating scores calculated? What kind of movies are more likely to have higher rating scores on IMDb? And is the movie worth-watching if it has a rating over 6? By searching some related topics online, we find that Sun Chuan give some interesting insights about rating scores prediction, such as the longer the movie is ,the higher the rating will be. Therefore, we want to look into the movies dataset and get more insights about what affects the IMDb rating scores.

- Hypothesis:

  After the first meeting, we come out some factors that may influence the ratings.
  - ➔ The director of the movie
    - ◆ The more famous the director is, the higher the rating will be
    - ◆ The larger number of the directors is, the higher the rating will be
  - ➔ The color of the poster
    - ◆ The movies with light colors on posters are more likely to have a higher rating score than those with dark colors on posters.
  - ➔ Advertisement strategies
    - ◆ How do producers advertise their movies will affect the rating score.
    - ◆ High exposure to media online tend to lead a rating score.
  - ➔ The budget of the movie
    - ◆ The higher the budget is, the higher the rating score will be.
  - ➔ The genre of the movie
    - ◆ Sci-fi movies are more likely to have higher ratings than Romance movies.

ETHICAL CONSIDERATIONS:

- Permission to the dataset:
  The dataset we will use is allowed for public access. The dataset is from data.world which is an open source dataset website. The dataset is provided by Chuan Sun who scraped tons of metadata using a combination of www.the-numbers.com, IMDB.com, and a Python library called "scrapy".

- Privacy Concern:
  Although the dataset we use is completely public, we still have our privacy concerns regarding our dataset. For instance, the dataset reveals 2399 unique director names, thousands of actors/actresses and the number of likes on their Facebooks. In this case, we decide to comply with the Safe Harbor Method and anonymize our dataset.

- Potential Biases:
  One of the potential biases of the dataset is the countries of the movies. Since the dataset only contains the information about 66 countries, it can only apply to the IMDB score over those countries.
  Another potential bias of the dataset is that the IMDB scores are only by the preference of users in English-speaking countries. The IMDB score itself can only represent the rating of movies with English-speaking users.

BACKGROUND:

The dataset that we will be examining for this project consists of a collection of thousands of movies over a span of 100 years from 66 different countries. The data collected from these movies were used to determine how movie ratings are created through an abundant number of different ways such as social media, advertisement, the actors who are involved in the movie and many other things were analyzed as well. Through this dataset our group was able to analyze the different input that is taken from social media, advertisements, and all the different categories that are used to rate movies in order to come up with an overall rating for these movies.

A similar project that our group found was “IMDB Exploratory Data Analysis Project” by Ilya Ezepov. Ezepov’s dataset project is very similar to the path our group has decided to take because Ezepov also explores movie ratings through a number of different categories but instead of figuring the exact movie rating through these categories, Ezepov takes a different approach of predicting movie ratings through movie characteristics and the overall production budget. Some of Ezepov’s findings consisted of sound and color being a huge hit in the movie industry, political and economical events impact the audience's movie choices, and economic events also impact the amount of views per movie.

In addition, another similar project that we found was “Movie Exploratory Analysis Using IMDB Datasets” by Yash Sharma. Sharma’s project explores the ideas of which country produces the most movies and what kind of movies are produced the most. Sharma’s project examines a similar dataset as the one our group has decided to analyze but takes a different approach while analyzing it. Some of the insights that Sharma was able to collect through the dataset used for his project consisted of Drama, Comedy, and Thriller are the top genres and that the US has the most thriving movie industry. Both of these projects are very similar to the approach our group has decided to take and can be a great resource for the creation of our final project.

References Link:
IMDB Exploratory Data Analysis Project - Ilya Ezepov:
http://rstudio-pubs-static.s3.amazonaws.com/52740_40aabe898b7a46b99c2b3f4ca3042e8a.html
Movie Exploratory Analysis Using IMDB Datasets - Yash Sharma:
https://rpubs.com/yash91sharma/dw_project_ys

DATA:

- Dataset Name: IMDB 5000 Movie Dataset
- Dataset Link: https://data.world/popculture/imdb-5000-movie-dataset
- Number of Observations: 5043
- Number of Features: 29
- Dataset Description:
    - ➔ This dataset shows there are many influential factors towards the IMDB rating of each movie. With 29 variables and 5043 describing (almost) every aspect of IMDB rating of different movies spanning across 100 years and 66 countries. There are 2399 unique director names, and thousands of actors/actresses.

Team Expectations Agreement

Read over the COGS108 Team Policies individually. Then, include your group's expectations of one another for successful completion of your COGS108 project below. Discuss and agree on what all of your expectations are. Discuss how your team will communicate throughout the quarter and consider how you will communicate respectfully should conflicts arise. By including each member's name above and by adding their name to the Gradescope submission, you are indicating that you have read the COGS108 Team Policies, accept your team's expectations below, and have every intention to fulfill them.

These expectations are for your team's use and benefit—they won't be graded for their details. Goals should be realistic: "No group member will never miss a meeting and everyone will always show up early" is probably unrealistic, but "Group members will attend almost every meeting and will communicate their absence at least a day in advance of the group meeting" and "When group members are unable to attend a meeting, they will submit their notes and progress ahead of the group meeting" are realistic expectations. Expectations for deadlines, how you'll work together, meeting attendance and participation, and project completion should all be considered and details included below.

INCLUDE YOUR TEAM'S EXPECTATIONS HERE

- Show up to set meetings
- Constant communication with each other
- Set/meet deadlines for individual assignments
- Help each other out when stuck
- Constructive criticism
- Meeting every other Monday at 6pm

Project Timeline Proposal

Include actual dates and times for due dates and meetings below, not just what week they'll be completed

	Draft Text?	**Write Code?**	**Proposed due date**	**Discuss at team meeting**	**Edit?**
Initial team meeting	NA	NA	NA	week 2	NA
Background Research	Meliza R.	NA	week 3	week 4 (4/29) Monday @ 6pm	Sutianyi Wen
Question & Hypothesis	Sutianyi Wen	NA	week 3	week 4	Meliza R.
Ethical Considerations	Froilan Grepo	NA	Week 3	Week 4	Meliza R.
Dataset	Shuibenyang Yuan	Shuibenyang Yuan	week 3	week 4	Sutianyi Wen
Data Wrangling	Shuibenyang Yuan	Shuibenyang Yuan	week 3	week 4	Sutianyi Wen
Descriptive	Froilan Grepo	Froilan Grepo	week 5	week 6 (4/13) Monday @ 6pm	Shuibenyang Yuan
Exploratory	Shuibenyang Yuan	Shuibenyang Yuan	week 5	week 6	Sutainyi Wen
Analysis - Part I	Cristina Ahamad	Sutianyi Wen	week 6	week 7	Shuibenyang Yuan
Analysis - Part II	Cristina Ahamad	Cristina Ahamad	week 6	week 7	Shuibenyang Yuan
Analysis - Part III	Sutianyi Wen	Sutianyi Wen	week 6	week 7	Cristina Ahamad
Summarize Results	Meliza R.	NA	week 7	week 8 (4/27) Monday @ 6pm	Cristina Ahamad
Conclusions	Meliza R.	NA	week 7	week 8	Cristina Ahamad