Project Title/Team Name :

- MovieDataMagic - GitHub : https://github.com/ocsc-datascience/MovieDataMagic

Team Members :

- Christian Ott
- Jane Chang
- Jack Jeng

Project Description/Outline :

- We would like a better understanding of the domestic box office (DBO) correlations among the following factors:
  a. Rotten Tomatoes Score (Critics & Audience)
  ~~b. Weather Forecasts~~
  c. Calendar Date/Four Seasons (Spring, Summer, Winter, Fall)
  d. Twitter Sentiment using Vadar limited since twitter only allows 7 days data pulls

Hypothesis :

1. If higher the RT score (seasons, twitter sentiments) then higher the box office opening.

Null Hypothesis :

2. The correlation isn't strong enough to make our assumption (>RT score == Higher Box Office Openings) conclusive.

Research Questions :

- E.g. If there is a correlation between **domestic box office opening weekend** and **Rotten Tomatoes score**, then a higher Rotten Tomatoes score (95%) correlates to blockbuster box office opening weekend. ($50.0 Million) A low Rotten Tomatoes score (20%) correlates to small box office opening weekend. ($6.0 Million)

- If there is a correlation between **domestic box office opening weekend** and **seasons** then **summer** correlate to blockbuster box office opening weekend. ($50.0 Million). **Summer** correlate to small box office opening weekend. ($6.0 Million)

- If there is a correlation between **domestic box office opening weekend** and **twitter sentiment**, then a **high positive & low negative twitter sentiment** correlate to blockbuster box office opening weekend. ($50.0 Million) **High negative sentiment & low positive sentiment** correlate to small box office opening weekend. ($3.0 Million)

- E.g. If there is NOT a correlation between **domestic box office opening weekend** and **Rotten Tomatoes score**, then a higher Rotten Tomatoes score

(95%) does NOT correlates to large blockbuster box office opening weekend. ($50 Million) A low Rotten Tomatoes score (20%) does NOT correlates to small blockbuster box office opening weekend. ($6.0 Million)

- E.g. If there is NOT a correlation between **domestic box office opening weekend** and **calendar dates**, then **calendar months May, June, July August** correlate to large blockbuster box office opening weekend. ($50.0 Million) **Calendar months May, June, July August** (35%) correlate to small box office opening weekend. ($6.0 Million)

- If there is NOT a correlation between **domestic box office opening weekend** and **Rotten Tomatoes score**, then a higher Rotten Tomatoes score (95%) does NOT correlates to large blockbuster box office opening weekend. ($50 Million) A Rotten Tomatoes score (35%) does NOT correlates to small box office opening weekend. ($3.0 Million)

Data Sets to be Used :

- We plan to use public data sets exposed to the public including:
    a. Rotten Tomatoes Score via review aggregator website RottenTomatoes.com cowned by Flixster/Warner Bros. (30%) and Fandango Media/NBCUniversal (70%)
    ~~b. Weather via OpenWeather API~~
    c. Box Office Grosses via Box Office Mojo created by Brandon Gray
    d. Release Dates via Box Office Mojo created by Brandon Gray
    e. Calendar Dates/Seasons

Rough Breakdown of Tasks :

- Here's the tasks:
    1. Research and formulate Research questions such as does domestic box office (DBO) correlate to rotten tomatoes score? Weather? Calendar dates mid june vs mid sept?
    2. Identify, procure, and store appropriate data sets
    3. Analyze data sets using Python pandas and SciPy, matplotlib, etc.
    4. Using statistical techniques identify useful insights specifically correlations (if have time)
        a. Calculate One Sample t-tests allow you to compare your sample mean to the population mean
        b. Or Calculate Independent (Two Sample) t-tests allow you to compare the means of 2 independent samples
        c. Check the p-value less than 0.05?
        d. Chi square test
    5. Assumptions

    a. Data is normally distributed
    b. Data is independent
    c. Data is randomly sampled
    d. Data is examined wide releases (3,000+ theater locations)
    e. Data seasons are defined as the following:
        i. Winter - Note: The Winter Season is defined as the first day after New Year's week or weekend through the Thursday before the first Friday in March
        ii. Spring - Note: The Spring Season is defined as the first Friday in March through the Thursday before the first Friday in May.
        iii. Summer - Note: The Summer Season is defined as the first Friday in May through Labor Day Weekend.
        iv. Fall - Note: The Fall Season is defined as the day after Labor Day Weekend through the Thursday before the first Friday in November.
        v. Holiday - Note: The Holiday Season is defined as the first Friday in November through New Year's week or weekend.

To-Do List

1) Create 5-6 data visualizations that analyzes the data
    - **a) Scatter plot**
    - b) Bar Charts
    - c) Pie Chart?
2) Christian (today):
    - a) DONE: Plots showing box office vs. ratings, both linear and log
    - b) DONE: Figure out hypothesis testing for ratings vs. box office
3)