Phase-1 Project Report

Review

Deepali Sharma ()

Overview

- Business Goal
- How to Approach the Problem?
- Main Steps in Data Analysis
- Data Selection
- Recommendations
- Future Work

Business Goal

- Business Problem:
  - Microsoft sees all the big companies creating original video content and wants to create a new movie studio
- Objective of the Problem:
  - Provide three recommendations to Microsoft for the type of movies that it should invest or make.

How to Approach the Problem ?

- There are a few metrics one can chose to decide the type of movies
  - **Profitability**: the very first and most obvious is to look at the absolute profit of movies.
    - this can be studied as a function of **genre**, **director**, **writer**, **movie length** and **release time** of year
  - **Content**: Based on **movie ratings**, **critical reviews**, **director**, **writer** etc etc. **Caveat**: may not be profitable

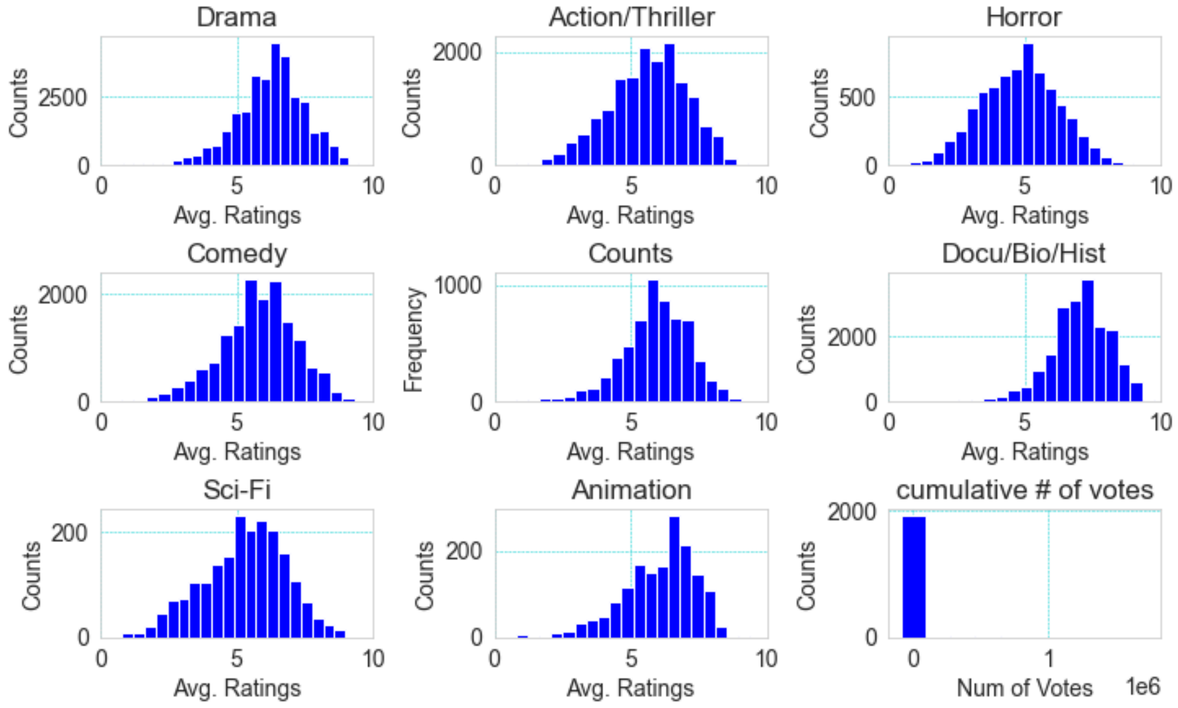
Main Steps in Data Analysis

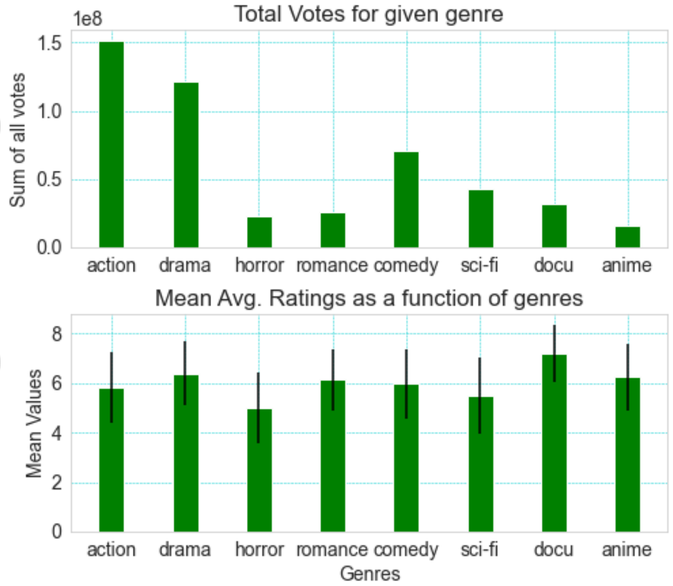
- Read Data tables (SQL, CSV/TSV)
- Clean them —- remove duplicates, nulls and merge entries for a given movie that had multiple directors or any other parameter
- Find out the important variables (move id, director, names, ratings, votes etc)
- Find out what variables to use for the merging of tables (used movie name for IMDB (sql) and BoxOffice Mojo (csv)) and check if the release dates matched.
- Tried using release date to merge tables from other data tables where movie names weren't available. This resulted in some wrong data getting merged together. So abandoned the tables that didn't have movie names.
- Convert the variables that should be numeric to numeric and dates to date time objects.
- Define some functions that could do the repetitive jobs.

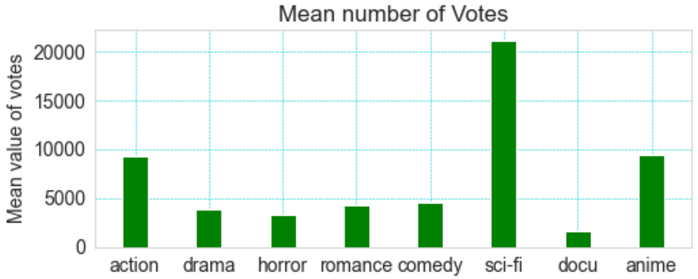
Data Selection

- After looking in detail at the various data tables, I decided to use the following datasets
  - From IMDB (SQL) and from Box Office Mojo (csv). These two datasets after merging provided me the info about directors, profits, genres, avg. ratings, votes.
- Rest of the tables Rotton tomates/TheNumbers/TheMovieDB either lacked the info about movie names or gave too few entries after merging with the other tables. Merging using release dates didn't work out and to decode the info about genre ids or movie ids (to know genre or name) required to go to websites and download more data(I am already behind deadline :(.)
- One could still extract some blind analysis for genres vs ratings etc from these tables.

Profit or Avg. Ratings ?



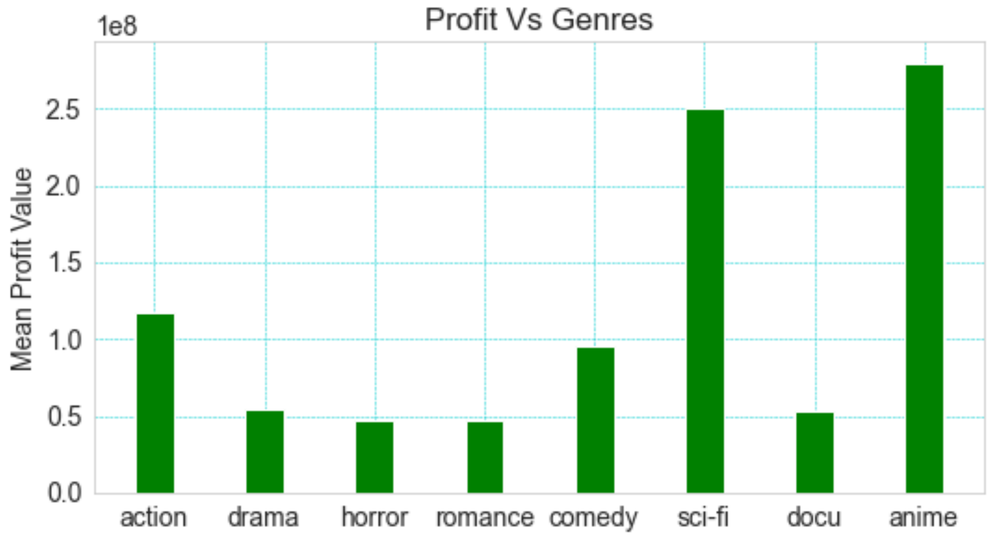


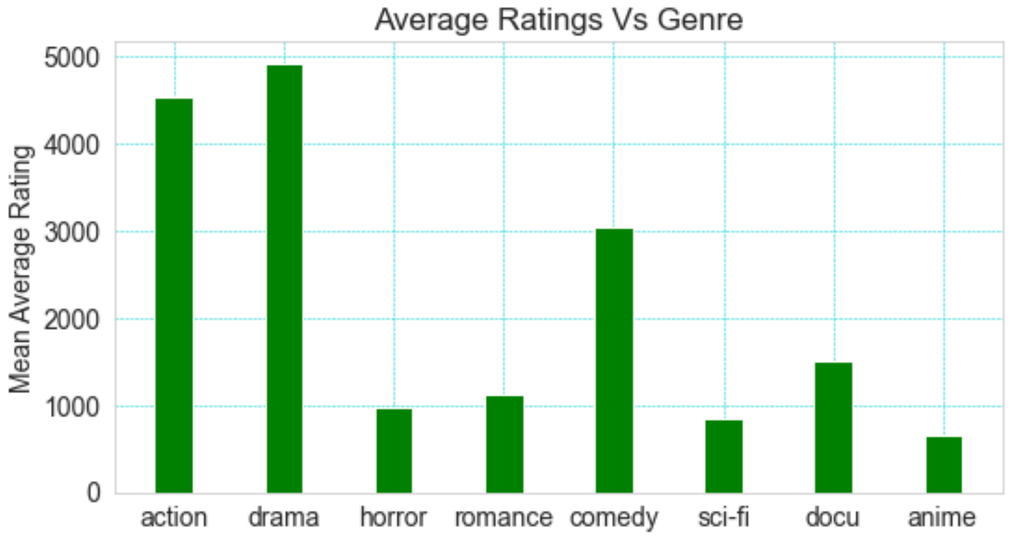


- I first started looking at average ratings as a metric to decide about the movies.
- But with the current data the average ratings distributions for different genres didn't give any concrete results. So decided to us profit instead !
- However, one should do a multi-dimensional analysis using ratings, profits, votes, popularity and that I believe would give more concrete understandings/results.

Recommendation 1 (Best Genre)

- Based on my classification of movies into different genres, I found that sci-fi/anime/action movies are more profitable.
- Sci-fi/Anime movies however tend to have low ratings!

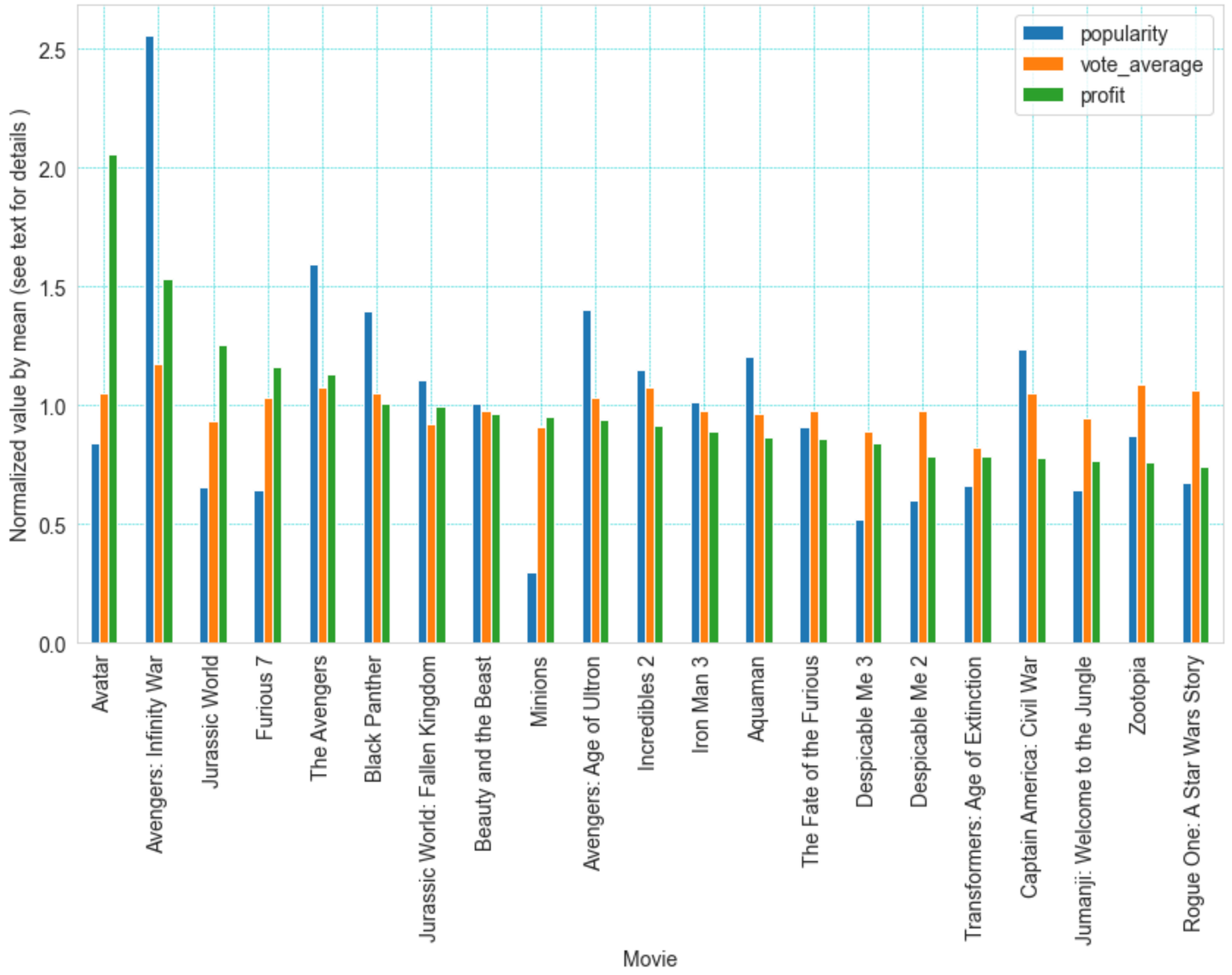




***I am not quite sure about the definition of average ratings in IMDB, is it viewer based or some combination of critics/consumers*

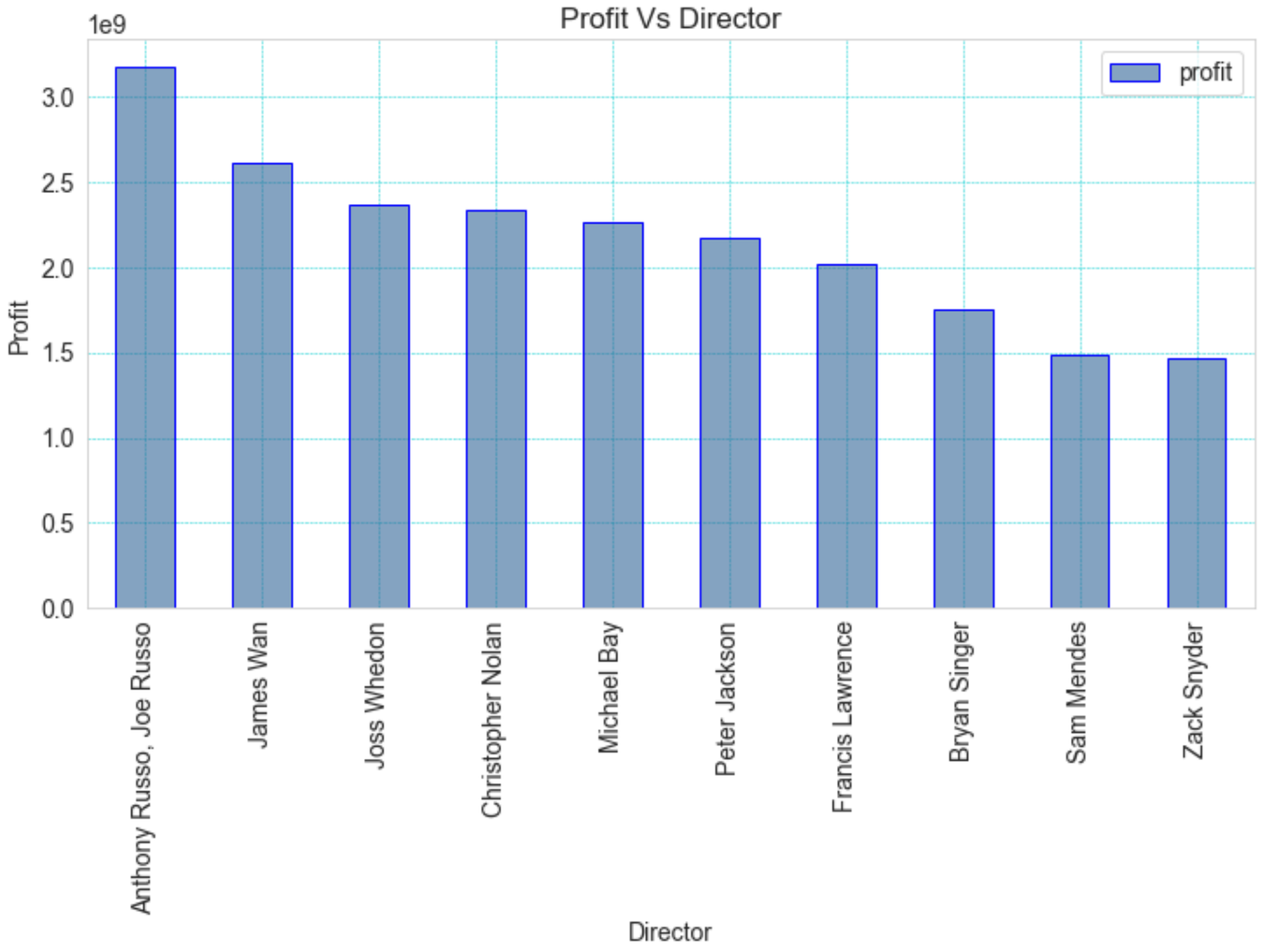
Recommendation 1 (Best Genre)

- A supporting plot to show that the sci-fi/ action/anime movies are indeed profitable.
- Plotted here is the normalized value for profit(popularity, vote average) for top 20 profitable movies .
- As can be seen the plot is dominated by movies from these 3 genres!



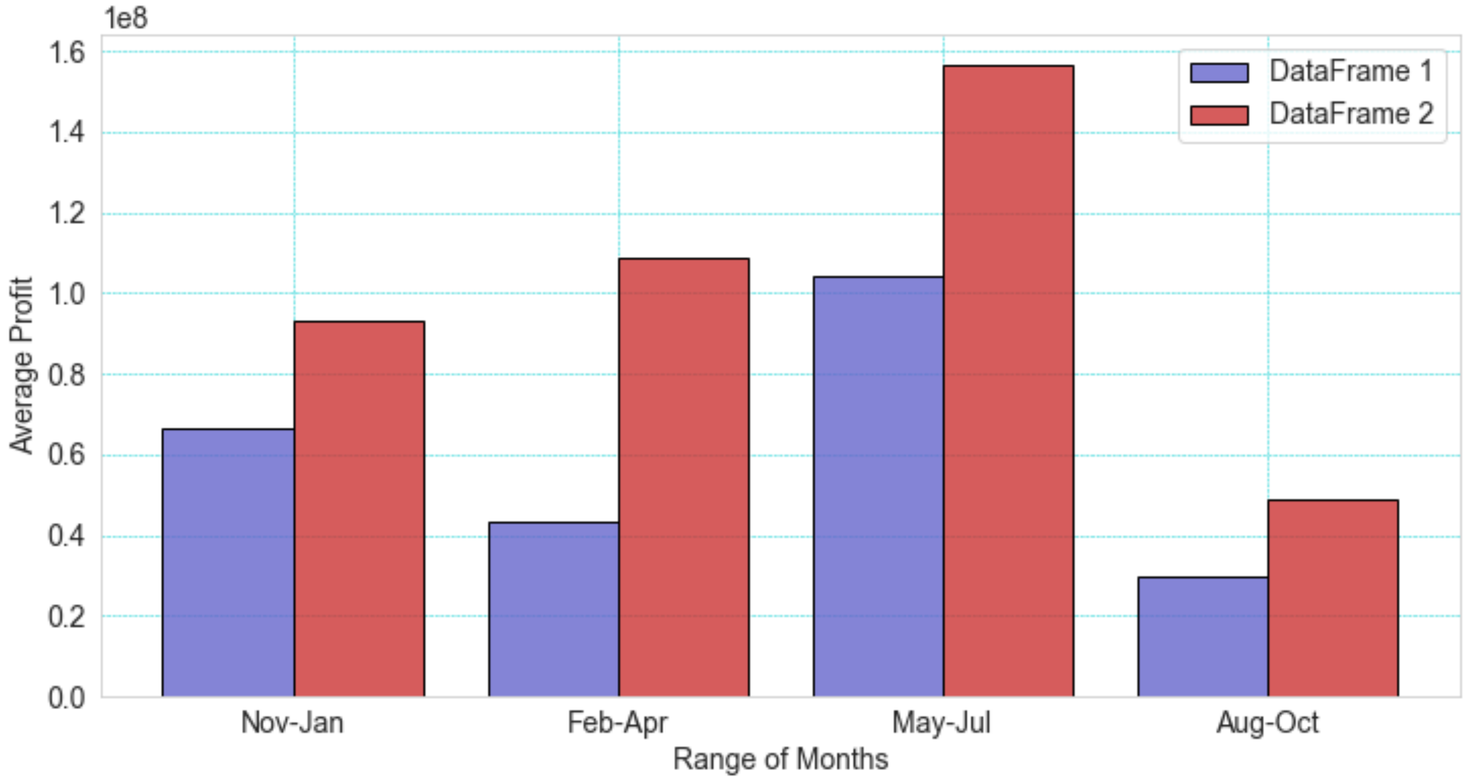
Recommendation 2 (Best Director)

- This plots shows the top 10 most profitable directors plotted in order.
- Russo brothers are the most profitable director duo (Avengers , Captain America etc)



Recommendation 3 (Best time to release movie)

- This bar chart shows the average profit made by movies that are released during different time periods in a year
- The two colors represent two data frames with one of them having more statistics.
- Based on this I see that movies release over summer months from May to July are more profitable.



Conclusion

- Based on the analysis of available data
  - I found out that **sci-fi movies/action/anime** movies do well in terms of profit and also tend to be more popular.
  - Also movies released during **summer months** turn to be more profitable as people are on summer break and going to movies is part of enjoyment.
  - Russo brothers, make on average, more profitable movies.
- Microsoft therefore should definitely **consider creating sci-fi movies,** and try to get top directors (Russo brothers if possible) for direction. Once the movie is ready, it should be times to come out in summer months to maximize the profitability!

Future Work

- Definitely more data needs to be analyzed that can be downloaded from websites or API's
- One needs to do a multidimensional analysis looking at various metrics such as average ratings, profits, directors, actors, popularity, votes etc to get a complete picture.
- Also one can look at foreign market and domestic markets separately.

Thank You

- **Deepali Sharma**: email:(deeps.sharma@gmail.com, deepali@rcf.rhic.bnl.gov)
- **Linkedin:** https://www.linkedin.com/in/deepali-sharma-a83a1126/
- **GitHub**: https://github.com/deepssharma

Back-Ups

Genres Definition

- df_action_db =df_movie_db[df_movie_db['genres'].str.contains("Action|Thriller|Crime|Mystery")==True]
- df_drama_db =df_movie_db[df_movie_db['genres'].str.contains("Drama|War|Family")==True]
- df_romance_db =df_movie_db[df_movie_db['genres'].str.contains("Romance")==True]
- df_horror_db =df_movie_db[df_movie_db['genres'].str.contains("Horror")==True]
- df_comedy_db =df_movie_db[df_movie_db['genres'].str.contains("Comedy")==True]
- df_docu_db =df_movie_db[df_movie_db['genres'].str.contains("Documentary|Biography|History|News|Music")==True]
- df_scifi_db =df_movie_db[df_movie_db['genres'].str.contains("Sci-Fi")==True]
- df_anime_db =df_movie_db[df_movie_db['genres'].str.contains("Animation")==True]