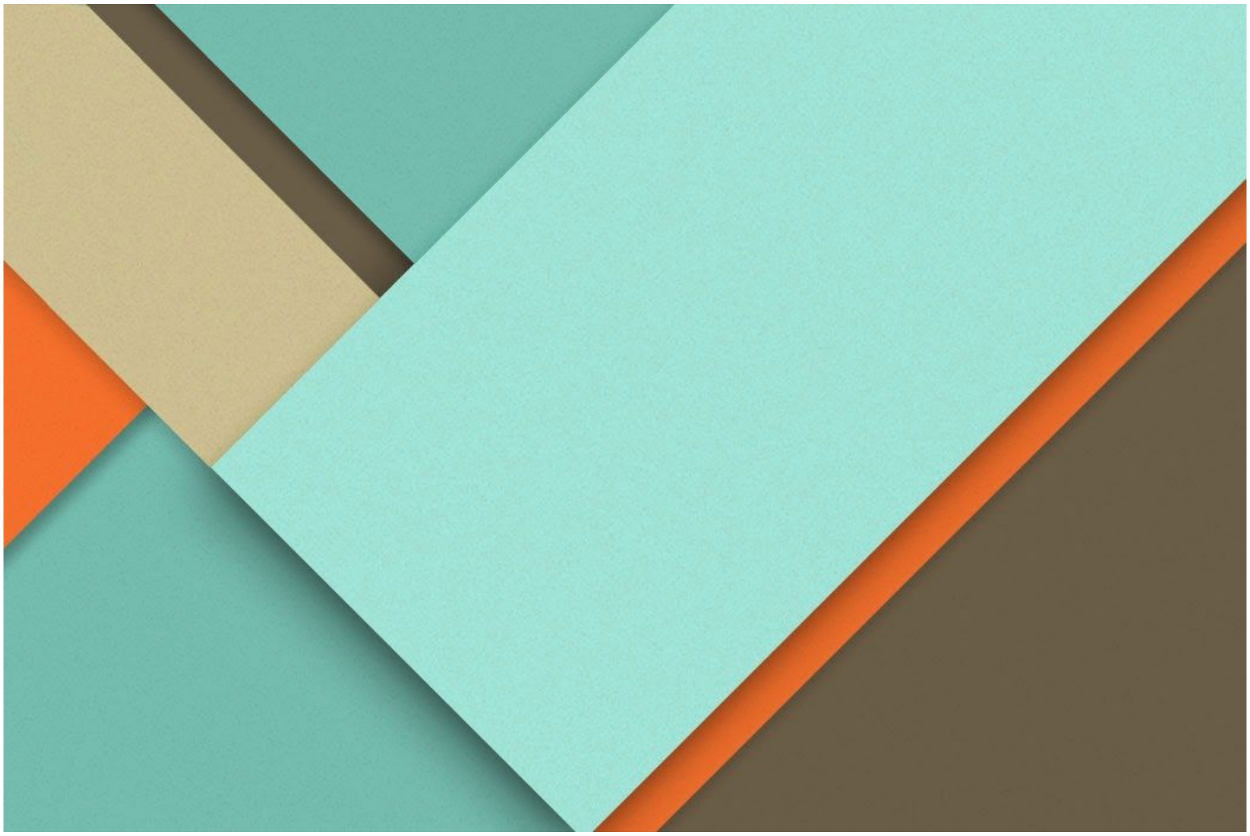

Midterm Project Proposal

11.05.2020

Wendy Liang
Boston University, MSSP
MA678 Applied Statistical Modeling

Overview

In this project, I will try to gain some insights into the movie and TV industry.

Personal Statement

I want to be a data analyst in the internet industry when I graduate. Solid data processing and analytics ability, modeling ability and insight into practical issues are needed to do well in this job.

I am always a big fan of movies and TV shows, so I choose it as the topic of my mid term project. This project enables me to walk through the data analysis process completely, from getting datasets to finishing a final report. In my vision, I will do EDA, classification, regression and prediction to those movie & TV show dataset (Or even text analysis If I could only grasp). All these processes will enable me to improve my analysis and modeling ability. In particular, I might use SQL to organize data in the first step, which I think is usual for a data analyst in an internet company.

Question

1. What are the characteristics of the released film and TV shows? (including country, director, case and so on)
2. What kind of films / TV shows tend to get higher ratings?
3. What are the most popular or successful actors and directors?
4. Can I predict the box office revenue of film (worldwide or national) through other variables?
5. What are the differences between several streaming media like Netflix and Hulu? What streaming media is best for me?
6. Does Netflix really focus more on TV shows tham film?

P.S. These questions are relatively macro and I will add detailed questions to each of them later.

Data Sources

TMDB

TMDB 5000 Movie Dataset

MovieLens

movielens dataset download

BoxOfficeMojo

R has special package to access the data

https://www.boxofficemojo.com/yearly/

OMDB

R has special package to use the API to access the data

omdb api

Other Dataset from Kaggle

Netflix Dataset

Movies on Netflix, Prime Video, Hulu and Disney+

TV shows on Netflix, Prime Video, Hulu and Disney+

The 45,000 Movies Dataset from tmdb and movielens

Timelines

I. EDA ---- 11.05 ~ 11.13

II. Data Processing ---- 11.14 ~ 11.20

III. Modeling and Validation ---- 11.21 ~ 11.27

IV. Write Up ---- 11.28 ~ 11.30