Data Storytelling Midterm

Mitchell Beckner

3/9/2020

Author

Mitchell Beckner

Links

Ghost Deck PDF on GitHub: https://github.com/mitchb63/MovieLens_Analysis/blob/master/Beckner_MovieLens_Analysis_ghost.pdf

Ghost Deck .pptx on GitHub: https://github.com/mitchb63/MovieLens_Analysis/blob/master/Beckner_MovieLens_Analysis_ghost.pptx

Data Cleaning and EDA in R: https://github.com/mitchb63/MovieLens_Analysis/blob/master/movie_analysis.R

Summary

This project was created as part of Udacity's Data Visualization Nanodegree program. The dataset contains intormation for more than 45,000 movies in the MovieLens Dataset. The dataset consists of movies released on or before July 2017. variables include cast, crew, plot keywords, budget, revenue, posters, release dates, languages, production companies, countries, TMDB vote counts and vote averages.

The objective of this project was to create a problem statement and PowerPoint ghost deck outlining a proposed analysis project. The problem statement was to be specific, measurable, and actionable, and the ghost deck should be logically structured and represent a clearly defined analysis based on a mutually exclusive, collectively exhaustive set of hypotheses.

The problem defined for this project is:

**What are the key components that differentiate the top 10% of films from the remainder when considering profit margin?

The hypotheses chosen as the main structure for the analysis were:

- Story factors influence profit margin
- Production factors influence profit margin
- Public opinion influences profit margin

An issue tree summarizing the analysis can be seen in Figure 1.

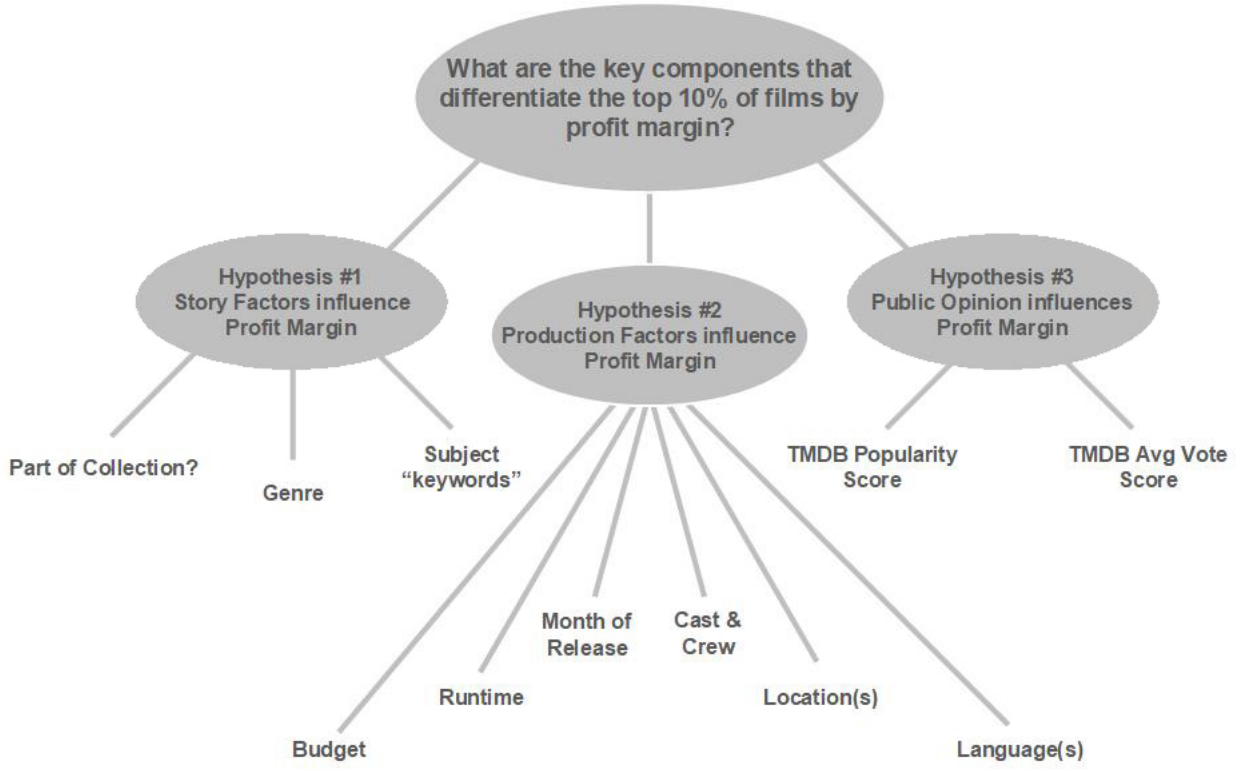

Figure 1: MovieLens Analysis Issue Tree

Data Cleaning and EDA

The original files required some transformation and data cleaning. This was done using a combination of Excel and R. New variables for profit and profit margin were created as profit margin is to be the primary measure of a film's success. Once these variables were created, the dataset was filtered to remove any films that did not include the financial information required for the anaylsis. This took the available data from over 45,000 movies to under 5,300. The 90th percentile based on profit margin was then determined for the remaining films and a new variable was created assigning each film to either the top 10% or bottom 90% group.

Very basic EDA was then performed in R just to get a rough beginning idea of what might be found when exploring the proposed hypotheses. Examples of this early analysis can be seen in the pdf and PowerPoint files linked to above.

References

Anonymous. (1970, January 1). McKinsey Presentations - How to Apply Ghost (aka Shell and Skeleton) Decks and Pages. Retrieved March 7, 2020, from http://workingwithmckinsey.blogspot.com/2013/07/McKinsey-presentations-ghost-decks.html

Banik, R. (2017, November 10). The Movies Dataset. Retrieved March 1, 2020, from https://www.kaggle.com/rounakbanik/the-movies-dataset#movies_metadata.csv