Analysis of Bechdel Tests on Movies from 1970-2013

Team Members

1. Alexander Thurston – A02257888 (Undergrad) - Lead for Preprocessing & Data Collection
2. Hailey Dennis – A02242676 (Undergrad) - Lead for Statistical Analysis & Data Storage
3. Tyler Kunz – A02275136 (Undergrad) - Lead for Machine Learning

Description of the Project

This project aims to explore existing data on Bechdel Tests performed on 1,794 movies ranging from 1970-2013. Alison Bechdel created three test cases to explore the prominence of women in film (https://en.wikipedia.org/wiki/Bechdel_test). The test cases involved go as follows: at least two women are present in the movie, these women talk to each other, and they converse about something other than a man. The motivation behind these tests is to measure gender inequality in the entertainment industry. One matter of significance behind this is the idea that film representation both influences and is influenced by our society. The data we are using provides a score based on which of the test criterion each movie passes, and whether they pass or fail overall.

With the data gathered, we will find patterns and comparisons among budgets, sales, time periods, awards, etc. with respect to the results of these tests. Specifics on the data will be noted in the section below. This will allow us to draw conclusions relating the success of a movie to the roles of women portrayed in them.

Each member of the project will lead efforts in their respective specialties to prepare and explore the dataset. We will all actively participate in each step of the project and document findings to share with each other in our meetings. So far our research goals include implementing thorough data collection and preparation techniques, statistical analysis, market basket analysis, and machine learning topics/models as we learn them in class.

In summarizing our findings, we will use various representations such as raw statistics, graphs, and possibly even 3D data visualization techniques on top of our report and presentation. Our hope is that this will lead us to understand better patterns in the way gender interacts with popular movies.

Dataset

Five-Thirty-Eight Bechdel Test Dataset -
https://github.com/fivethirtyeight/data/blob/master/bechdel/movies.csv

TMDB API - https://developers.themoviedb.org/3/getting-started/introduction

We will use the tabular dataset from the Five-Thirty-Eight dataset which consists of 1,794 movies we will append additional features from TMDB. Preprocessing will consist of feature extraction supported by initial statistical analysis of the dataset as well as any necessary changes needed to the format of the data in order to work with our model.

The initial features we will focus on will be the Bechdel Test criterion failed/passed, movie budget, gross domestic revenue, international gross revenue, year, and genre. Throughout the course of the project, we may add features such as awards, nominations, cast, etc.

Implementation Plan

Checkpoint 1: March 24

Before the progress report deadline, we plan on completing all preprocessing and data collection. Preprocessing includes selecting useful features from the main dataset and gathering supplemental data via the TMDB API. The data will be cleaned and stored in an SQL database and a 1-2 page progress report will be written. If time allows, some basic data analysis will be performed. Alex will lead this segment, but all other group members will be contributing.

Checkpoint 2: April 14

The halfway point between the project report and the project presentation is approximately April 14th and is when we will complete the statistical analysis portion of our project. We will have calculated descriptive statistics for the data and generated visuals displaying trends in the collected data. Hailey will be in charge of this portion of the project.

Presentation Deadline: April 30

By the presentation deadline, we will have completed the machine learning segment of this project, which will be led by Tyler. Multiple regression models and clustering algorithms will be trained to predict whether a movie passes the Bechdel Test. Results will be graphed and displayed. We will make a poster for the final presentations.

Report Deadline: May 2

Following the class presentations we will write a report on our findings.

Role of Members

Alex will gather and analyze related data from GitHub and the TMDB (The Movie Database) API. Conduct feature selection for the dataset to narrow down redundant correlations. Ensure the quality of data through preprocessing techniques such as determining covariance, random sampling of data, data discretization, and feature aggregation.

Hailey will focus on statistical analysis and data storage. This will include useful summarizations and visual representations according to relevant findings. She will be responsible for storing data from the original data source along with any additional data found via TMDB, as needed. We are planning on using MySQL to store and query items. This will be especially useful in initial data exploration that can be built upon later. She will also lead any necessary efforts with detecting anomalies if applicable.

Tyler will lead the machine learning portion of the project, which will involve generating predictive models and looking for patterns based on the dataset. Regression analysis and other models will be used to determine trends related to the Bechdel Test. TMDB data will also be incorporated in order to determine what variables correlate to a movie’s Bechdel evaluation.