A Data Science Approach to Gender Representation in Film

Elizabeth Combs (eac721)

Lauren D'Arinzo (lhd258)

Angela Marie (Amber) Teng (at2507) – responsible

"Regardless of the content, a film has the power to shape perceptions of moviegoers on a range of subjects from love and marriage to the work of the government"

-cf. Franklin 2006 ; Kolker 1999 ; Ortega-Liston 2000 ; Riggle, Ellis, and Crawford 1996

Abstract: The Bechdel test is a blunt, basic measure of gender equality in fictional media. From a data science perspective, we wondered if and how we could automate the process of scoring a film based on the Bechdel criteria. This project aims to complete a descriptive analysis to assess the potential of creating such a method. We review the possibility through Tableau visualization dashboards to compare the number of movies scored and the ratio of passing movies over time. These visualizations are available online at our website. We subsequently examine the use of word count as a potential proxy to quantify the words spoken by women in a film. In the conclusion, we address the modeling and ethical implications of quantifying cultural variables for feature engineering.

Background

Description of Problem:

An integral part of data science is the communication of results; visualizations make it easier to present and discuss data, detect patterns, and identify significant points in a dataset. As data scientists, it is important that we make these visualizations accessible and interpretable by our population of interest.

In the digital age, our group sought to better understand how women are represented in films, particularly those rated by both users and critics. For this exploration, we utilized the Bechdel Test, which is a measure of the representation of women in fiction. The test is comprised of three criteria: a movie or work of fiction has to (1) have at least two women in it, (2) who talk to each other, and (3) have a conversation that revolves around a topic other than a man. The test has historically been used as an indicator for the "active presence of women" in a work of fiction, and in the field of media studies, to shed light on gender inequality in works of fiction [1]. However, from the website documentation, existing methods of the scoring regimen appear to be largely manual -- you have to watch the movie to do it. Consequently, we were motivated to consider ways that the Bechdel Test could be made more efficient through automation, while using compelling visualizations along the way to support our resulting conclusions.

Results

First Visualization

We viewed the data to brainstorm how we could visualize the data to explore automation for the Bechdel test using data science techniques. We then examined a few descriptive characteristics in the data by building plots in matplotlib and seaborn. This exploration allowed to start answering the following questions: (1) How does the score of the Bechdel Test relate to IMDb User Rating? (2) How does the distribution of Bechdel passing movies change over time? (3) How much does content play a role vs. word count? After data processing, we experimented with visualizations.

Tableau Dashboards

To create compelling visualizations for deployment, we again returned to our dataset for additional processing. Next, we iteratively generated clean, aggregate datasets that could be used for Tableau interactive dashboards. All three dashboards have time on the X-axis in order to examine trends over time. The first visualization is an area plot that utilizes dual Y-axes, comparing the distribution of movies that have been rated on the Bechdel scale to total number movies released. The second dashboard compares the ratio of movies that pass the test with the proportion of words belonging to female characters, aggregated across movies. The last dashboard shows the proportion of movies that pass the Bechdel test conditioned on the genre of the movie. In all three dashboards we implemented filtering widgets so that the user can interact with the figure and examine particular time periods or genres. All line graphs show the moving average by decade in order to smooth the trend line for easier interpretation.

Website Deployment

To deploy our visualizations and share our data-driven insights with the public, we created a website using HTML and CSS to tell a story and embed our Tableau dashboards. We used bootstrap themes to create a framework, and then customized our HTML index file to show our story in a cohesive manner, particularly using sectioning among the continuous scroll. We also added a menu bar with additional resources for users who are interested in learning more about our data science process. Then, we deployed our website on getforge.io, uploading our files as a compressed zip folder. Finally, we created a Binder to allow users to actively interact with our code in the form of a Jupyter Notebook and provided a link to our GitHub repository.

Discussion and Conclusion:

Despite the upward trend over time in Figure 1, the maximum ratio is less than 40%, supporting the argument that there is a need to make the scoring process more efficient. If we are to consider the representation of women in media as reflective of perceptions of women by society, it is important that the distribution of Bechdel scores is representative of all movies in order to prevent selection bias.

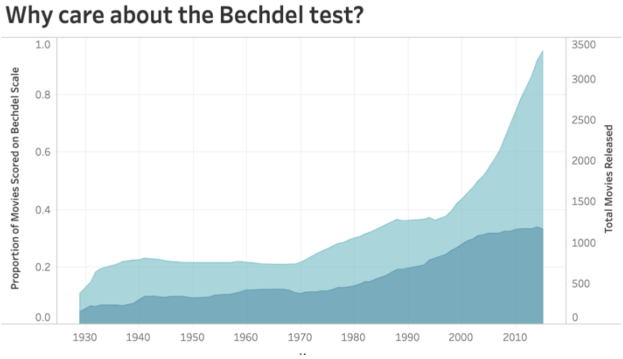
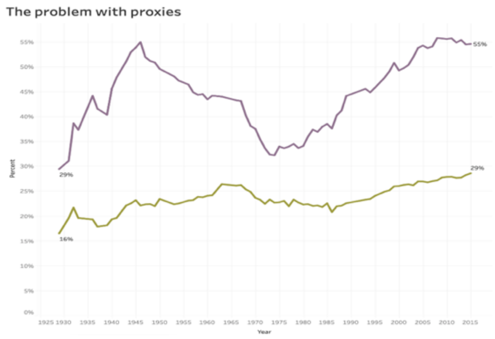


Figure 1

Without preliminary data exploration word-count would seem like a logical proxy to capture representation for use in an algorithmic approach. However, the results of Figure 2 suggest otherwise. Since 1970, the ratio of movies passing the test has a larger slope than the ratio of female words to total words; thus, the content criterion of the Bechdel Test seems to play an important role compared to just the word count. We may be missing information if we only use word count as a proxy in our feature engineering process.

Figure 2



In Figure 3, we are able to see clear trends in the proportion of passing movies when the feature is split on genre. Romance and Horror films have had the highest ratio of passing the Bechdel test since the 1980s, while Crime and Action films have had the lowest ratio of passing.

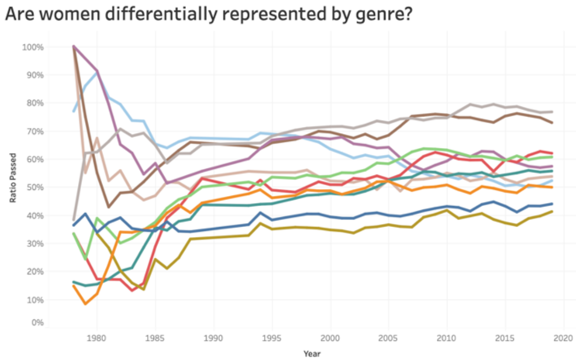


Figure 3

In this exploration of data, we have begun to see how nuanced an automation methodology would need to be to accurately capture the scoring of the Bechdel test. It would be difficult to find single proxy variables for the three criteria of the test, and feature engineering would need to include domain knowledge. This is not to say that automation is impossible; but, as data scientists, we will need to voice our concerns about using unrepresentative proxies, and we should include the voices of domain experts when modeling qualitative outcomes that may be subjective.

Finding a way to automate the Bechdel test may be a difficult undertaking; however, it could help better capture the true proportion of films passing the Bechdel test by scoring all movies in the IMDb and beyond. It is worth noting is that the Bechdel test is not the only metric for quantifying representation in media, as it only considers one identity variable -- gender. Washington Post writer Alyssa Rosenberg claims, “The Bechdel Test set a very low floor for Hollywood. We can't let it become the ceiling for progress [3].” In the past decade, many other representation tests have been formulated, considering the distribution of people who work on the film’s production, or relating other intersectional features like race, ethnicity, socioeconomic status, so that representation is inclusive to identities beyond cisgender white women. Though our analysis focuses on the Bechdel Test, our suggestions about feature engineering can help establish similar methodologies for other tests and continue to build a more equitable movie industry.

Although we were able to integrate aggregated script data, there are other features that were not available in the data sources we used. One option we would consider for future work would be web-scraping on a site like Wikipedia to get more information about each movie. Once the corresponding urls are found, we could use the library lmxl to parse the Wikipedia content into document object models in order to extract additional features.

Software Used:

- *Python*: Packages: requests, numpy, pandas, json, matplotlib, seaborn, Jupyter Notebook / Google Colab
- *Version Control*: GitHub
- *Visualization*: Tableau
- *Website Deployment*: GetForge, HTML/CSS/JavaScript + Bootstrap

Data Sources:

- Bechdel (API)
- Scripts: character_list5.csv, Character_mapping.csv
- IMDb: title.akas.tsv, title.basics.tsv, title.ratings.tsv

References:

"Bechdel Test Movie List." *Bechdel Test Movie List*, https://bechdeltest.com/.

"Bechdel Test." *Wikipedia*, Wikimedia Foundation, 19 Nov. 2019, https://en.wikipedia.org/wiki/Bechdel_test.

Garcia, David, et al. "Gender Asymmetries in Reality and Fiction: The Bechdel Test of Social Media." *Association for the Advancement of Artificial Intelligence*, 2014, https://www.sg.ethz.ch/media/publication_files/Title-Gender-Asymmetries-in-Reality-and-Fiction-The-Bechdel-Test-of-Social-Media.pdf.

Rosenberg, Alyssa. "Opinion | In 2019, It's Time to Move beyond the Bechdel Test." *The Washington Post*, WP Company, 21 Dec. 2018, https://www.washingtonpost.com/opinions/2018/12/21/its-time-move-beyond-bechdel-test/.