1. \*\*Automation and Provenance Tracking\*\*:

- Dr. Keegan's attempt to reproduce Walt Hickey's analysis likely faced challenges in terms of automation and provenance tracking if Hickey's original work was not well-documented or automated. If Hickey did not provide clear scripts or records of the data analysis process, it would have been difficult for Dr. Keegan to automate the reproduction effectively. For the most part, Hickey did. But there are a bunch of places where Hickey made some assumptions without proof and used data sources that aren’t open source. Keegan struggled there.

- Dr. Keegan should have documented any challenges faced during the reproduction process and highlighted any discrepancies between their results and Hickey's original results to maintain transparency. He did a great job here.

In the absence of knowing the precise methods used but making reasnable assumptions of what was done, I was able to replicate some of his findings, but not others because specific decisions had to be made about the data or modeling that dramatically change the results of the statistical models. However, the article provides no specifics so we're left to wonder when and where these findings hold, which points to the need for openness in sharing data and code. Specifically, while Hickey found that women's representation in movies had no significant relationship on revenue, I found a positive and significant relationship.

But the questions and hypotheses Hickey posed about systematic biases in Hollywood were also the right ones. With a reanalysis using different methods as well as adding in new data, I found statistically significant differences in popular ratings also exist. These differences persist after controlling for each other and in the face of other potential explanations about differences arising because of genres, MPAA ratings, time, and other effects.

The article doesn't actually say how these variables are constructed or where they come from, so I could be wrong but the data and tradition definitions of these variables present obvious candidates involving different relationships between the same two variables Budget (Expenses) and Income (Revenue).

Hickey's article uses data about international receipts, which doesn't appear to be available publicly from this website, so I cannot replicate these analyses.

Here I tried to write some code to get historical CPI data from the [BLS API](http://www.bls.gov/developers/home.htm), but I couldn't get it to work. Instead, I downloaded the January monthly data from 1913 through 2014 for Series ID CUUR0000SA0 which is the "CPI - All Urban Consumers" and saved it as a CSV. The inflator dictionary provides a mapping that lets me to adjust the dollars reported in previous years for inflation: a dollar in 1913 bought 23.8 times more goods and services than a dollar in 2014.

The article claims to use two specific outcomes: **return on investment** and **gross profits** which are different ways of accounting for the relationship between income and costs. I don't claim to be an accounting or business expert, but "gross profit" is traditionally defined as "income minus costs" while "return on investment" (RoI) is traditionally defined as profits divided by assets. Thus we need at least three variables: income, costs, and assets.

However, only two of these are available in the movie financial data we obtained from [The-Numbers.com](http://www.the-numbers.com/): income and costs. We can calculate profits easily, but it's unclear what a movie's assets are here unless we use costs again. Again, Hickey may have done something different for make this variable, but we don't know so we can't replicate.

2. \*\*Availability of Software and Data\*\*:

- Availability of software and data can be a significant hurdle for reproducibility. If Hickey used proprietary software or data sources that were not made available, Dr. Keegan might have faced difficulties in reproducing the exact analysis. Issue with data source. Keegan has done an excellent job highlighting how he went about it.

- Dr. Keegan should have clearly stated whether they had access to the same software and data sources used by Hickey. If not, they should have explained how they obtained similar data or used alternative methods to address this challenge. He has done so excellently.

3. \*\*Open Reporting of Results\*\*:

- Open reporting of results is a critical aspect of reproducibility. Dr. Keegan should have provided detailed documentation of their reproduction process, including any modifications made to Hickey's methodology or code. He has provided everything, going as far as making everything open-source available on Github. In fact, he encourages others to reuse his code to make better findings. He is a strong advocate of transparent data journalism.

- If Dr. Keegan's analysis was well-documented and openly reported, it would enhance the reproducibility of their work. Any deviations or discrepancies between their results and Hickey's should have been clearly explained. He has explained it well.

While I would have trouble reproducing Hickeys work, Keegans I can easily reproduce because he has adhered to all the principles.