Introduction to data science

Patrick Shafto

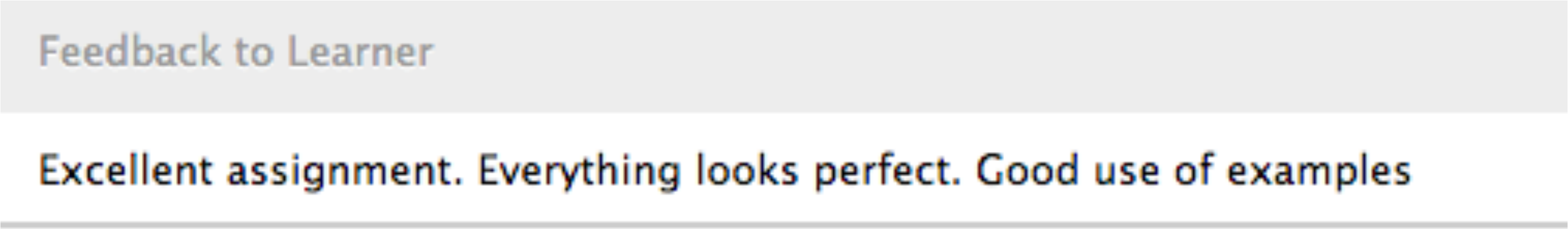
Department of Math and Computer Science

Plan for today

- HW 3
- Asking and answering questions (Part 1 of N)

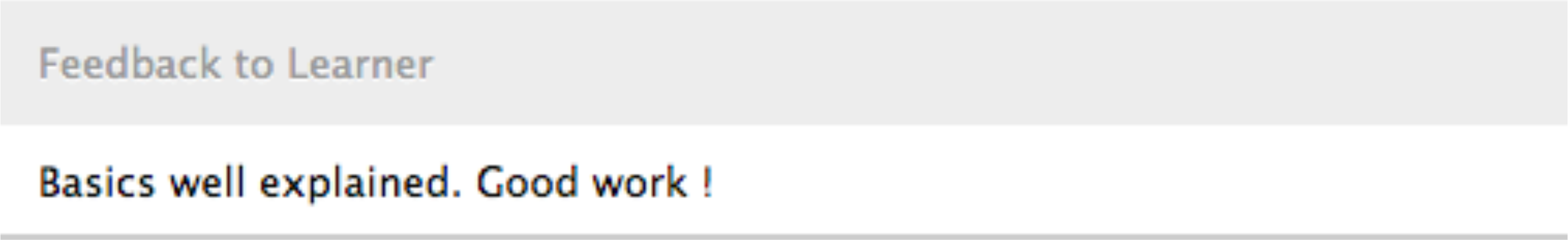
HW 3

- Write a short tutorial on numpy and scipy
- Cover basic functionality with worked examples
- Due Sunday by 11:59 pm
- Also read the Betchel test notebook and come prepared to answer questions
- Evaluations due Tuesday by 11:59pm
    - Use the rubric!
    - Justify your scores!



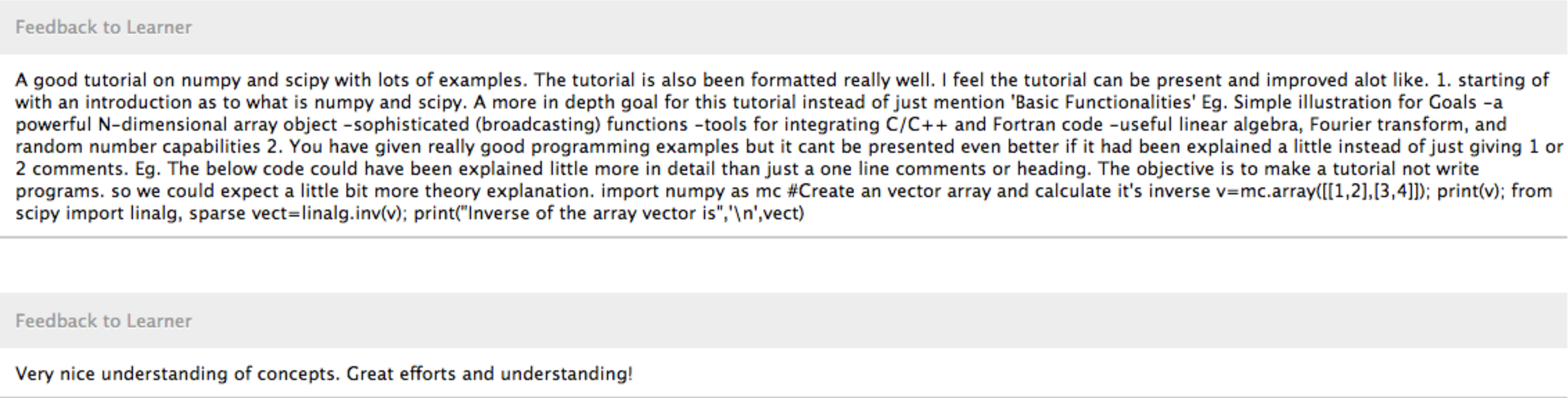
Feedback to Learner

Excellent assignment. Everything looks perfect. Good use of examples



Feedback to Learner

Basics well explained. Good work !

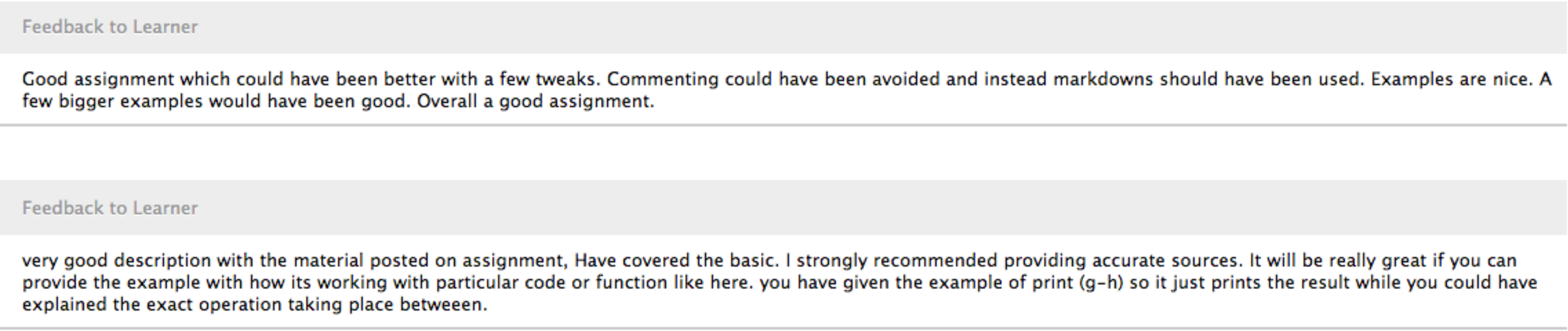


Feedback to Learner

A good tutorial on numpy and scipy with lots of examples. The tutorial is also been formatted really well. I feel the tutorial can be present and improved alot like. 1. starting of with an introduction as to what is numpy and scipy. A more in depth goal for this tutorial instead of just mention 'Basic Functionalities' Eg. Simple illustration for Goals -a powerful N-dimensional array object -sophisticated (broadcasting) functions -tools for integrating C/C++ and Fortran code -useful linear algebra, Fourier transform, and random number capabilities 2. You have given really good programming examples but it cant be presented even better if it had been explained a little instead of just giving 1 or 2 comments. Eg. The below code could have been explained little more in detail than just a one line comments or heading. The objective is to make a tutorial not write programs. so we could expect a little bit more theory explanation. import numpy as mc #Create an vector array and calculate it's inverse v=mc.array([[1,2],[3,4]]); print(v); from scipy import linalg, sparse vect=linalg.inv(v); print("Inverse of the array vector is",'\n',vect)

Feedback to Learner

Very nice understanding of concepts. Great efforts and understanding!

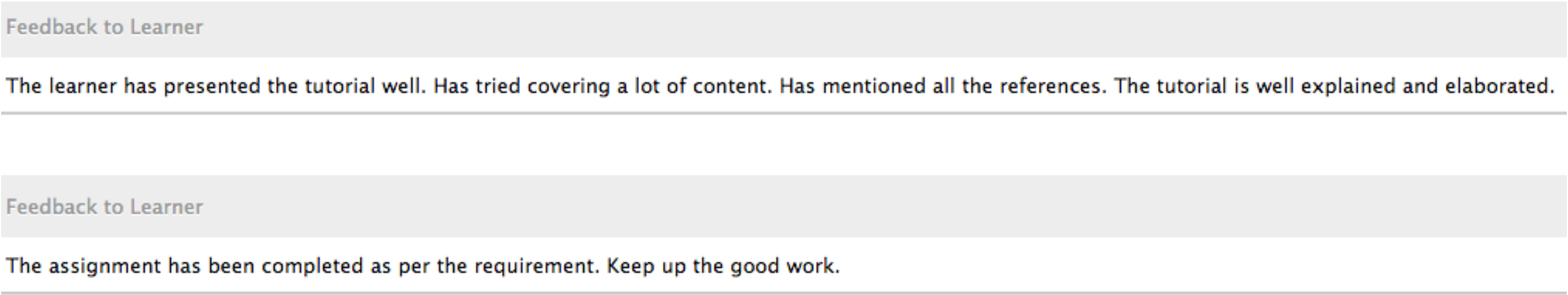


Feedback to Learner

Good assignment which could have been better with a few tweaks. Commenting could have been avoided and instead markdowns should have been used. Examples are nice. A few bigger examples would have been good. Overall a good assignment.

Feedback to Learner

very good description with the material posted on assignment, Have covered the basic. I strongly recommended providing accurate sources. It will be really great if you can provide the example with how its working with particular code or function like here. you have given the example of print (g-h) so it just prints the result while you could have explained the exact operation taking place betweeen.



Feedback to Learner

The learner has presented the tutorial well. Has tried covering a lot of content. Has mentioned all the references. The tutorial is well explained and elaborated.

Feedback to Learner

The assignment has been completed as per the requirement. Keep up the good work.

Asking and answering questions (Part 1 of N)

The need for openness in data journalism
https://nbviewer.jupyter.org/github/brianckeegan/Bechdel/blob/master/Bechdel_test.ipynb

Analyzing the structure of a report

What is the goal?

The Need for Openness in Data Journalism

Brian Keegan, Ph.D. (@bkeegan) College of Humanities and Social Sciences, Northeastern University

Do films that pass the Bechdel Test make more money for their producers? I've replicated Walt Hickey's recent article in FiveThirtyEight to find out. My results confirm his own in part, but also find notable differences that point the need for clarification at a minimum. While I am far from the first to make this argument, this case is illustrative of a larger need for journalism and other data-driven enterprises to borrow from hard-won scientific practices of sharing data and code as well as supporting the review and revision of findings. This admittedly lengthy post is a critique of not only this particular case but also an attempt to work through what open data journalism could look like.

- To that end, this article tries to do many things for many audiences which admittedly makes it hard for any single person to read. Let me try to sketch some of these out now and send you off in the right path.

- **First, I use an article Walt Hickey of FiveThirtyEight published on the relationship between the financial performance of films that the extent to which they grant their female characters substantive roles as a case to illustrate some pitfalls in both the practice and interpretation of statistical data**.

- This is a story about having good questions, ambiguous models, wrong inferences, and exciting opportunities for investigation going forward. If you don't care for code or statistics, you can start reading at "The Hook" below and stop after "The Clip" below.

- To that end, this article tries to do many things for many audiences which admittedly makes it hard for any single person to read. Let me try to sketch some of these out now and send you off in the right path.

- **Second, for those readers who are willing to pay what one might call the "Iron Price of Data Journalism", I go "soup to nuts" and attempt to replicate Hickey's findings**.

- I document all the steps I took to crawl and analyze this data to illustrate the need for better documentation of analyses and methods. This level of documentation may be excessive or it may yet be insufficient for others to replicate my own findings. But providing this code and data may expose flaws in my technical style (almost certainly), shortcomings in my interpretations (likely), and errors in my data and modeling (hopefully not). I actively invite this feedback via email, tweets, comments, or pull requests and hope to learn from it. I wish new data journalism enterprises adopted the same openness and tentativeness in their empirical claims. You should start reading at "Start Your Kernels..."

- To that end, this article tries to do many things for many audiences which admittedly makes it hard for any single person to read. Let me try to sketch some of these out now and send you off in the right path.

- **Third, I want to experiment with styles for analyzing and narrating findings that make both available in the same document.**

- The hope is that motivated users can find the detail and skimmers can learn something new or relevant while being confident they can come back and dive in deeper if they wish. Does it make sense to have the story up front and the analysis "below the fold" or to mix narrative with analysis? How much background should I presume or provide about different analytical techniques? How much time do I need to spend on tweaking a visualization? Are there better libraries or platforms for serving the needs of mixed audiences? This is a meta point as we're in it now, but it'll crop up in the conclusion.

- To that end, this article tries to do many things for many audiences which admittedly makes it hard for any single person to read. Let me try to sketch some of these out now and send you off in the right path.

- **Fourth, I want to experiment with technologies for supporting collaboration in data journalism by adopting best practices from open collaborations in free software, Wikipedia, and others.**

- For example, this blog post is not written in a traditional content-management system like WordPress, but is an interactive "notebook" that you can download and execute the code to verify that it works. Furthermore, I'm also "hosting" this data on GitHub so that others can easily access the writeup, code, and data, to see how it's changed over time (and has it ever...), and to suggest changes that I should incorporate. These can be frustrating tools with demoralizing learning curves, but these are incredibly powerful once apprenticed. Moreover, there are amazing resources and communities who exist to support newcomers and new tools are being released to flatten these learning curves. If data journalists joined data scientists and data analysts in sharing their work, it would contribute to an incredible knowledge commons of examples and cases that is lowering the bars for others who want to learn. This is also a meta point since it exists outside of this story, but I'll also come back to it in the conclusion.

Analyzing the structure of a report

What is the goal?

- Walk Hickey published an article on April 1 on FiveThirtyEight, titled The Dollar-And-Cents Case Against Hollywood's Exclusion of Women. The article examines the relationship between movies' finances and their portrayals of women using a well-known heuristic call the Bechdel test. The test has 3 simple requirements: a movie passes the Bechdel test if there are (1) two women in it, (2) who talk to each other, (3) about something besides a man.

- Hickey's article makes two central claims:
  - We found that the median budget of movies that passed the test...was substantially lower than the median budget of all films in the sample.
  - We found evidence that films that feature meaningful interactions between women may in fact have a better return on investment, overall, than films that don't.

- In the image below, we see that movies that have non-trivial women's roles get 24% lower budgets, make 55% more revenue, get better reviews from critics, and face harsher criticism from IMDB users. Bars that are faded out mean my models are less confident about these findings being non-random (higher p-values) while bars that are darker mean my models are more confident that this is a significant finding (lower p-values).

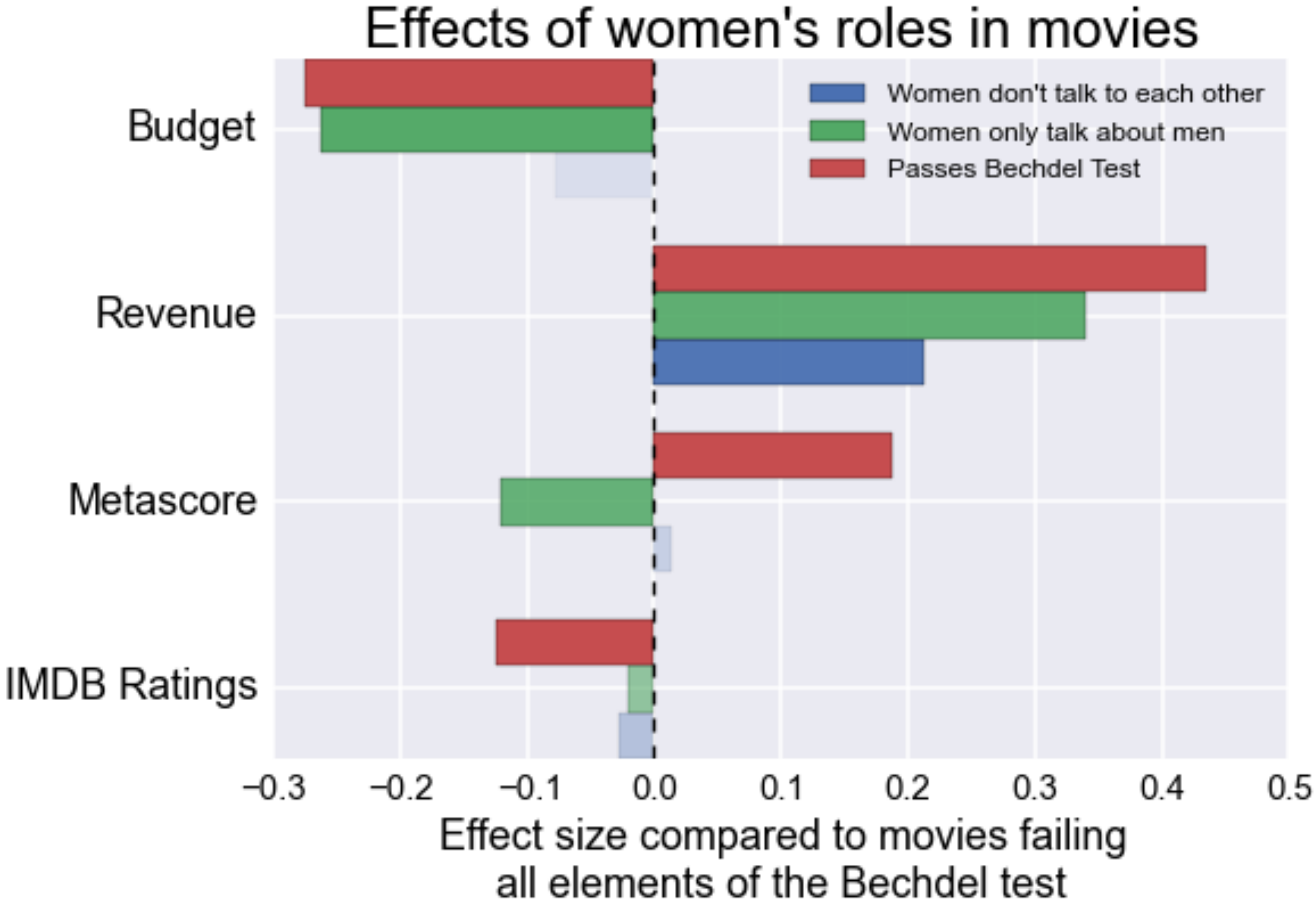
- Movies passing the Bechdel test (the red bars):

- ...receive budgets that are 24% smaller

- ...make 55% more revenue

- ...are awarded 1.8 more Metacritic points by professional reviewers

- ...are awarded 0.12 fewer stars by IMDB's amateur reviewers


Effects of women's roles in movies
Women don't talk to each other
Women only talk about men
Passes Bechdel Test
Budget
Revenue
Metascore
IMDB Ratings
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4
0.5
Effect size compared to movies failing
all elements of the Bechdel test

Analyzing the structure of a report

What is the goal?

What are the data?

What are the data

What are the data? Through this process of obtaining the data, it becomes clear that the analyst quickly has to make choices about the types of data to be obtained. Does the data come from Table X or Table Y? Does budget data include marketing expenditures or not? Are these inflation-corrected real dollars or nominal dollars? They should be the same, but there are things missing in one that aren't missing in the other. These decisions are not documented in the article nor are these data made available, both of which makes it hard to determine what exactly is being measured.

What are the data

What are the variables? The article also creates variables such as "return on investment" and "gross profits" from other variables in the data. These data can be highly-skewed or contain missing data which can break some statistical assumptions -- how were these dealt with? 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). In creating a new variable that combines two other variables, the behavior of this new variable is intrinsically related to the other variables. This can become problematic very quickly, as we will see in the next step.

What are the data

Revenue data from The-Numbers.com

Inflation data from BLS

Bechdel Test data from BechdelTest.com's API

Additional data from OMDBAPI

Analyzing the structure of a report

What is the goal?

What are the data?

What is the analysis/logic?

What are the analyses

What are the models? The article then performs analyses on these new and old variables using methods that assume they should be independent. As the previous bullet emphasizes, "return on investment" and "gross profits" are financial performance outcomes that already capture a relationship between Budget and Income. The model could measure these outcomes as a function of the Bechdel score alone. The model could also estimate basic Income as a function of Bechdel plus throw in Budget as a control. The model might still also could try to estimate the combined financial performance as a function of Bechdel controlling for Budget again, even though is already in the outcome. The write-up makes it seem like the last model is the one used, which is problematic.

What are the analyses

Bechdel Test over time
Budgets differ
Earnings differ
A different model

Analyzing the structure of a report

What is the goal?

What are the data?

What is the analysis/logic?

What are the conclusions?

What are the conclusions

Hickey found that the number of movies passing the Bechdel test has been increasing over time. I found the same, but the effect had slowed in recent years. Simplistic extrapolations suggest it may take until the end of the century until the average movie passes the Bechdel test or that we run the risk of regressing towards more backwards portrayals of women in film.

What are the conclusions

Hickey found that Bechdel movies had lower median budgets. I also found that passing the Bechdel test was correlated with significantly lower budgets.

What are the conclusions

Hickey found that Bechdel movies had no effect on the film's ROI, controlling for budget. I also found that there was no significant relationship between ROI and movies passing the Bechdel test.

What are the conclusions

Hickey found that Bechdel movies had no effect on the film's gross profits, controlling for budget. I also found that there was no significant relationship between Profit and movies passing the Bechdel test.

What are the conclusions

I noted that ROI and Profit are problematic variables for the modeling Hickey described, so I used a simpler model of Revenue alone but still controlling for budget. This model provided a better fit to the data and also found that Bechdel movies had significantly higher revenue.

Analyzing the structure of a report

What is the goal?

What are the data?

What is the analysis/logic?

What are the conclusions?

Now review the report in detail…

HW 4

- Write a short tutorial on pandas
- Cover basic functionality with worked examples
- Due Sunday by 11:59 pm
- Evaluations due Tuesday by 11:59pm
    - Use the rubric!
    - Justify your scores!