Gmail **Fred Boehm <frederick.boehm@gmail.com>**

Journal of Statistics Education - Decision on Manuscript ID UJSE-2019-0159

2 messages

Journal of Statistics Education <onbehalfof@manuscriptcentral.com> Mon, Mar 9, 2020 at 9:49 AM
Reply-To: jo.hardin@pomona.edu
To: frederick.boehm@gmail.com
Cc: jeff.witmer@oberlin.edu

09-Mar-2020

Dear Dr Boehm:

Your manuscript entitled "A framework for mentored data science research" which you submitted to the Journal of Statistics Education, has been reviewed. The reviewer comments are included at the bottom of this letter.

I regret to inform you that the reviewers have raised serious concerns, and therefore your paper cannot be accepted for publication in Journal of Statistics Education. However since the reviewers do find some merit in the paper, I would be willing to reconsider if you wish to undertake major revisions and re-submit, addressing the reviewers' concerns.

In particular, we would like to encourage you to submit your paper back to JSE as a "Datasets and Stories" submission. A DSS paper has: a description of the pedagogical uses of multivariate dataset(s) (including discussions of achievable learning outcomes, potential pitfalls, helpful teaching tips), with the associated dataset(s) available for download with the manuscript for instructional use in classes and further analysis.

The analysis portion is quite interesting, and we agree that the project would have been engaging for an undergraduate student. However, the paper needs considerable work before publishing in JSE. The reviewers have given extensive suggestions, and we encourage you to comprehensively consider their feedback in your review. As they mention, there is very little discussion about "mentoring" and it isn't clear how your work is any different from a senior thesis project or an independent study. Instead, there is a lot of information about the data analysis structure which is why the paper might be better suited for a DSS contribution.

I've cc'ed JSE's editor, Jeff Witmer, on this email. Please feel free to reach out to him with any additional questions you might have on how to turn your submission into a DSS paper.

Please note that resubmitting your manuscript does not guarantee eventual acceptance, and that your resubmission will be subject to re-review before a decision is rendered.

You will be unable to make your revisions on the originally submitted version of your manuscript. Instead, revise your manuscript using a word processing program and save it on your computer.

To start your resubmission, please click on the link below:

*** PLEASE NOTE: This is a two-step process. After clicking on the link, you will be directed to a webpage to confirm. ***

https://mc.manuscriptcentral.com/ujse?URL_MASK=4cd304bf27b7479e8bec61449464ec88

This link will remain active until you have submitted your revised manuscript. If you begin a resubmission and intend to finish it at a later time, please note that your draft will appear in ‘Resubmitted Manuscripts in Draft” queue in your Author Center.

Because we are trying to facilitate timely publication of manuscripts submitted to Journal of Statistics Education, your revised manuscript should be uploaded by 09-Mar-2021. If it is not possible for you to submit your revision by this date, we will consider your paper as a new submission.

I look forward to a resubmission.

Sincerely,
Dr Johanna Hardin
Editor, Journal of Statistics Education

jo.hardin@pomona.edu

Reviewer(s)' and Associate Editors' Comments to Author:

Reviewer: 1

Comments to the Author
See attached.

Reviewer: 2

Comments to the Author
General Comments:

The manuscript “A framework for mentored data science research” provides a nice example of guiding undergraduates through a research project using Nolan and Lang’s recommended skill sets and the topic is of interest to JSE’s audience. For example, the manuscript includes interesting ideas for supplementing twitter data with other sources, as well as framing multiple research questions (eg for multiple student projects) from a single dataset. These topics are certainly of interest to supervisors of undergraduate research.

However, the manuscript could be greatly improved through reorganization. The paper is at times a call for reproducible methods in undergraduate education/research, an overview of the challenges of using Twitter data, and advice and lessons learned for mentors of undergraduate research. These topics are presented in such a way that a coherent message of the manuscript is hard to find. I think this article could be publishable as a case study of undergraduate research, and focus on advice and lessons learned for supervisors of undergraduate research — which is what the abstract seems to outline — but will require substantial reworking of the content.

These organizational issues are especially present in the “Methods” section. It is unclear to me whether the methods of the paper are the framework steps presented (Question formulation, Data acquisition, Data analysis and visualization, Presentation and communication), or Nolan and Temple Lang’s goals for students (Broaden statistical computing to include emerging areas, Deepen computational reasoning skills, Combine computational topics with data analysis in the practice of statistics, Develop and practice skills in reproducible research to promote open science). The “Results” section is quite short compared to the “Methods” and “Discussion” sections, and these sections could be reconfigured to better emphasize the author’s contributions.

The novelty of the work, and how it fits in with the broader literature (such as Nolan and Temple Lang’s paper) needs to be better highlighted.

Specific Comments:

Pg 2 line 43: When was the framework implemented?
Page 3 overview table: Which steps were performed by students and which by the mentor? Is this breakdown consistent with previous literature on data science research?
Page 3 overview table: Reproducibility was emphasized as a fourth main goal earlier on the page, but it’s unclear where this emphasis occurs in the framework overview table
Page 4 Line 77-79: This point about translating between scientific and statistical questions is very important, and something students often struggle with. It would be nice to have more description of how students practiced this skill with supervisors or independently.
Page 4 Line 85: The linked package on GitHub states that tweets were downloaded using the streamR package, while the manuscript says the twitteR package was used - it looks like twitteR is in a deprecation period in favor of rtweet. Clarifying or correcting text would be helpful here
Page 4 Line 91: Which sources? What types of data?
Page 5 Line 96-99: Projects have not been introduced yet, so “event detection project” and “sentiment analysis project” are not very informative. Perhaps describe each research project in detail within each subsection of “Framework Implementation”
Page 5-6 Line 117-130: This section was very helpful in understanding the projects and it would help if the description of projects occurred earlier in the paper (similar to comment (7))
Page 7 Line 152: How does the JSON parser in the package compare to the parsers in CRAN packages for interacting with Twitter’s API (eg rtweet, streamR)?
Page 7 Line 174: if Dirichlet is capitalized, Bayesian should be too
Page 8-9: The “results” section is very brief. Is there a way to reorganize some of the “methods” section to frame as results?
Page 10 Line 227: This line seems inconsistent with the emphasis on reproducible research mentioned on Page 3 goals,

Page 8, and Page 9 “Scholarly outcomes”
Page 10: Table has a reference/bibtex error in title
Page 10 Line 238-246: Additional reasoning behind (a) the categories chosen, and (b) how topics were categorized is needed.
General: There are many sections, subsections, and subsubsections, which would be easier to keep track of as a reader if they were numbered

Reviewer: 3

Comments to the Author
The authors describe a framework for mentoring undergraduate data science projects and provide useful resources for other educators to mentor tweet-based projects. I think the manuscript can make more of an impact by expanding discussion of the ways in which the authors grew as mentors, more details on how often the students met with the mentors and what type of mentoring advice was provided, and a description of the technical statistics skills the authors believe would be a pre-requisite for other students engaging in similar projects.

Specifically, I believe the manuscript would be strengthened by addressing these issues:
1. Explicitly adding to the Framework (or next iteration of the framework) "Answer Question" or something like that rather than the less specific "Interpret" or "Share findings." I appreciate that the students were encouraged to clarify the importance of the research question. I would recommend explicit encouragement to then deliver the answer to the research question. For research to be reproducible, the initial questions must be answered, not cast aside (because of null or uninteresting findings) in favor of another question that "revealed" itself during EDA. Hence, answering the research questions is an important step. Similarly, to enhance reproducibility, students should be encouraged to pre-register their research questions and methods, perhaps through a venue such as the Open Science Framework.

2. What statistics skills did the students have (in general) and what skills would be needed for future students to engage in such projects? Could this type of experience work for first-year undergraduates? Only seniors? Only honors students?

3. In what ways did the authors grow as mentors? How often did students meet with mentors? What types of mentoring advice was given? How would the mentors improve on their actions in future iterations?

4. Discussing the limitations of having projects "data-science-student-generated." In my experience, data science undergraduates rarely have the background and knowledge in a domain of application to generate "real scientific research questions" as was claimed on p9. A possible extension of this framework would be to have students work with (collaborate with) actual domain experts who have real questions based in sociology, political science, etc. domain expertise. Not to say that no data science student can generate a research-worthy question on her own, just that by collaborating with experts who actually know where the gaps in research exist guarantees that the questions asked are worth answering and worth spending 12 months on. Otherwise, one runs the risk of doing the project just for practice and conveying the notion that data science is "cool" or "fun" but not actually useful or especially worthwhile. The alternative of answering questions that a domain expert has deemed to be interesting enables the students to learn the data science skills by applying them to answer an important question. If, in fact, the questions the students in this case study answered were "real scientific research questions," why weren't the results published? It seems to me that the contributions to science were not based on the findings of the students, but rather on how much the students learned throughout their projects. This will enable them to build on these skills to do similar (and more advanced) projects in the future, but to what end? To continue to build their data science skills or to make actual contributions to research, business, or policy?

Other (minor) issues:
p5 l7/l98 run-on sentence

p7l43/l173 inconsistent capitalization of Dirichlet (previously).

l174 capitalize Bayesian

p9l11/l207: In what ways did the authors grow as mentors?

p10l19/l235 "determine" "plan"

p10l55/l245: The Figure 1 table is too long and should be reshaped into three columns under the 3 headers of Enduring..., Important..., and Worth...

Reviewer: 4

Comments to the Author
Summary: This paper describes a framework for mentoring undergraduate student research in data science, primarily illustrated by the authors’ experience mentoring two particular students.

Major Issues:

I'm not sure this is enough of a contribution, but I'll leave that up to the editors to decide. To me, this doesn't present anything novel. I know that students need to do question generation, data acquisition, analysis, and then presentation. I didn't come away from this paper with anything concrete that I could take back to my own teaching and mentoring. However, a case study may be appropriate for this venue. If accepted, I would encourage the title to be adjusted to "A case study of mentored data science research using Twitter data" or similar.

Minor Issues:

1. The introduction needs to start with more of an overview. The abstract doesn't get to serve as your introduction, so you can "repeat" things like the two honors students, and the overall framework idea. I would have liked more concrete details about the project you did.

- I wondered whether students were paid or received course credit for this experience, or if they were doing it extracurricularly. The paper says these projects took place over a 12-month period. How often did the mentors meet with the students, and for how long? How many hours a week did students work on this project? Etc.

- You use the word "encouraged" several times (about poster sessions and written senior theses). I was left wondering why you didn't not require those. Perhaps that was because students were not doing this for credit or money, so you only had an informal relationship with them, but that would be good to explain. Did students choose to take the opportunities you encouraged them to do?

- If there was a grading component to this, how did you determine grades? If not, was there a formal process for setting up this mentoring relationship, or did it develop naturally?

2. I found the introduction to be both overblown and unrelated to the project. It references "big data," which I consider to be any dataset too large to be stored on your computer. While Twitter data in the whole is big data, taking a subset of twitter from the streaming API does not result in big data. (I don't think, you could make this more concrete with description of how many tweets were sampled, and what the overall file sizes were.) A better introduction would be something like, "Data science is rising in importance in many fields. To prepare students for jobs after graduation, we…"

3. The introduction makes it sound as if the mentors chose to use twitter data, and gathered the data themselves. Later in the paper it becomes clear that students must have been doing some of the data generation themselves, especially because they wrote an R package, so the introduction could be re-worked to make that clearer. Throughout the paper, you use "we" in a way that makes it sound like it is just about the faculty. Perhaps this paper was coauthored with the undergraduate researchers (I can't see authors in the blinded version), but if not, you should probably be using "they" to refer to the students. For example, you say "we used the linux tool," which makes it sound like the faculty were doing that work.

4. The paper would be strengthened with more references. Is there something you can cite about the Teaching Statistics professional development class? Was it at a conference? Through your university? Who taught it? How long was it? What takeaways did you learn?

5. On pages 2-3, the reference to Nolan and Temple Lang seems to be more about big picture ideas for statistics faculty, rather than skill sets for students. Do we want students to "broaden statistical computing to include emerging areas"? Probably not, but we want them to apply their computational knowledge to emerging areas. We as the faculty are the ones doing the broadening. Perhaps you could re-phrase your goals in a more student-facing way, and reference the Nolan and Temple Lang as an inspiration.

Typos/copyediting:

p. 1 "We design" might be better as "We present"
p. 2 "work place" should be "workplace"
p. 2 "Teaching Statistics" should have comma inside quotation marks
p. 2 "Nolan and Temple Lang summarizes" should be "summarize"
p. 3 there is an errant "t" as a subtitle for the Framework Overview
p. 6 "students learned an aspect" should be "students learned aspects"
p.6 could the references to the twitter docs be put into end-of-text citations or footnotes, to avoid interrupting the flow of reading?
p. 8 why is R Core Team cited in the last sentence?
p. 13 "developer" should have comma inside quotation marks

Editor's Comments to Author

Associate Editor
Comments to the Author:
See report attached.

2 attachments

Comments to the Author.pdf
47K

UJSE-2019-0159_AE_Letter.pdf
77K

Fred Boehm <frederick.boehm@gmail.com> Mon, Mar 9, 2020 at 9:55 AM
To: BRET M HANLON <bmhanlon@wisc.edu>

Bret - We got some really constructive feedback on the JSE submission, although it was rejected. They recommend revisions and a resubmission under a distinct format that they're calling datasets and stories. In the next few days, I'll read carefully through the comments and begin revisions. I'm eager to undertake revisions and resubmit in the next few months.

Thanks again!

fred

Fred Boehm
My pronouns are he, him, and his
https://fboehm.us/
[Quoted text hidden]

2 attachments

Comments to the Author.pdf
47K

UJSE-2019-0159_AE_Letter.pdf
77K