Teaching data analysis through the lens of reproducibility

Mine Çetinkaya-Rundel
Duke University

RSHackathon
#1:
Convince researchers to adopt a reproducible research workflow

#2:
Train new researchers who don't have any other workflow
reproducibility often comes up in the context of published research and the need to accompany such research with the complete data and analyses, including software/code.

statistics educators who teach data analysis should be instilling best practices in students before they set out to do research.
traditional data analysis
- descriptive stats
- plots & tables
- model output

data analysis

write-up
- research question & context
- interpretations
- conclusions

write-up

lab report

lab report

👍 - familiar format
(Word / Google Doc etc.)

👎 - impossible to reproduce
- very difficult to update
- very easy for mistakes to creep in
- messy

👎 - impossible to reproduce
- very difficult to update
- very easy for mistakes to creep in
- messy
a better approach

👍 - code & output tightly connected
- easy to reproduce
- easy to collaborate
- easy to update
- standardized format
- faster

👎 - must learn syntax
- must be immaculate to compile
<table>
<thead>
<tr>
<th>Course Code</th>
<th>Course Title</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sta 101</td>
<td>Data Analysis &amp; Statistical Inference</td>
<td>Required intro stat course for (mostly) social science majors (Docker containers)</td>
</tr>
<tr>
<td>Sta 112FS</td>
<td>Better Living with Data Science</td>
<td>First-year seminar course for undergrads interested in quantitative fields</td>
</tr>
<tr>
<td></td>
<td>Coursera: Data Analysis &amp; Statistical Inference</td>
<td>MOOC on intro stats with peer assessment</td>
</tr>
<tr>
<td></td>
<td>Sta 470/851: Statistical Consulting Workshop</td>
<td>Elective for upper-class undergrads and grads in statistical science</td>
</tr>
</tbody>
</table>
reproducible science ... and better teaching instills early the idea that all analysis must be done in a reproducible framework.

... eases collaboration on labs and projects and removes ambiguity from grading.