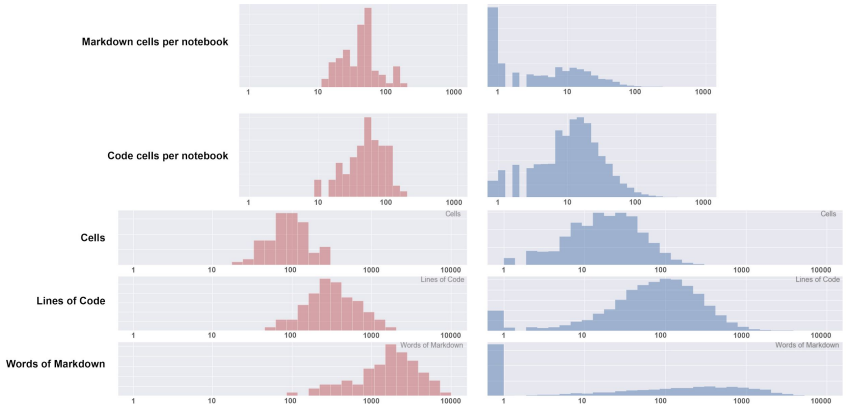
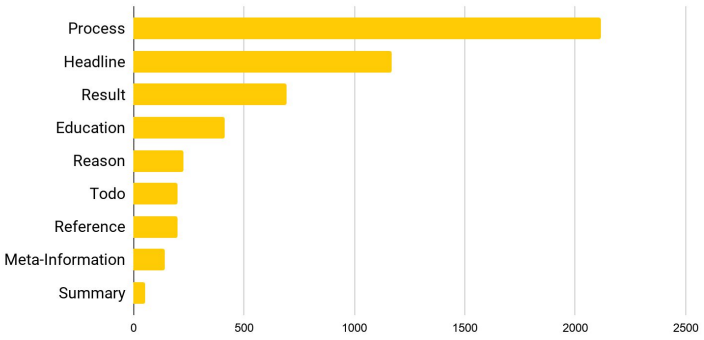
What Makes a Well-Documented Notebook? A Case Study of Data Scientists' Documentation Practices in Kaggle

We sampled 80 highly-voted Kaggle notebooks and conducted a qualitative content analysis to explore their documentation practices.

Documentation Coverage

We replicated the notebook level descriptive analysis by Rule et al.'18. The left side (in red) represents the 80 highly-voted computational notebooks from Kaggle and the right side (in blue) represents the 1 million computational notebooks on Github. The Kaggle Corpus is better documented compared to the Github Corpus.





A Broad Range of Topics

We coded nine categories of documentation in the Markdown cells. For example, "process" describes what the adjacent code cell is doing; "headline" is the styled headline for organizing the notebook into separate sections; "result" describes the outputs from code execution; "reason" explains results or critical decisions.

Interplaying with Data Science Stages

We coded Markdown cells based on where they belong in the data science workflow. We found that the data science problems on Kaggle competitions contain clearly project goals and datasets. In total, we identified four stages and 13 tasks.

4 Stages

- Environment Configuration
- Data Preparation and Exploration
- Feature Engineering and Selection
- Model Building and Selection

Presenter April Yi Wang

Authors April Yi Wang[1] • Dakuo Wang[2] • Jaimie Drozdal[3] • Xuye Liu[3]
Soya Park[4] • Steve Oney[1] • Christopher Brooks[1]

Affiliations [1]University of Michigan • [2]IBM Research
[3]Rensselaer Polytechnic Institute • [4]MIT CSAIL