2022 Kaggle DS & ML Survey

Methodology:

- Survey data collection
    - The full list of questions that were asked (along with the provided answer choices) can be found in the file: kaggle_survey_2022_answer_choices.pdf. The file contains footnotes that describe exactly which questions were asked to which respondents. Respondents with the most experience were asked the most questions. For example, students and unemployed persons were not asked questions about their employer. Likewise, respondents that do not write code were not asked questions about writing code. For more detail, see the footnotes associated with each question in the document kaggle_survey_2022_answer_choices.pdf.
    - The survey was live from 08/16/2022 to 09/16/2022. We allowed respondents to complete the survey at any time during that window. An invitation to participate in the survey was sent to the entire Kaggle community (anyone opted-in to the Kaggle Email List) via email. The survey was also promoted on the Kaggle website via popup "nudges" and on the Kaggle Twitter channel.
- Survey data processing
    - The survey responses can be found in the file kaggle_survey_2022_responses.csv. Responses to multiple choice questions (only a single choice can be selected) were recorded in individual columns. Responses to multiple selection questions (multiple choices can be selected) were split into multiple columns (with one column per answer choice). The data released under a CC 2.0 license: https://creativecommons.org/licenses/by/2.0/
    - To ensure response quality, we excluded respondents that were flagged by our survey system as “Spam” or "Duplicate. We also dropped responses from respondents that spent less than 2 minutes completing the survey, and dropped responses from respondents that selected fewer than 15 answer choices in total.
    - To protect the respondents’ privacy, free-form text responses were not included in the public survey dataset, and the order of the rows was shuffled (responses are not displayed in chronological order). If a country or territory received less than 50 respondents, we grouped them into a group named “Other”, again for the purpose of de-identification.