| Criteria | Very good | Sufficient | Needs Improvement |
|---|---|---|---|
| **1. Motivation (section 1 of your documentation, 5%)** | | | |
| Clear explanation of the value in collecting the data (“task in mind”), either in the context of a specific research question or business problem. The data collection potentially generates insights into new phenomena, increases the managerial relevance of empirical work, helps to develop new models, or is an efficient way for collecting valuable information. There is clearly value to the larger research community in using the data. | Clear and well-justified motivation with strong links to the research problem. The data collection offers valuable insights and contributions | Adequate explanation of the motivation, but could use more detail on how it ties to the research problem and potential insights | Unclear or weak motivation. The value of data collection is not sufficiently explained |
| A wide range of relevant websites and APIs pertaining to the data context are assessed. The use of different extractions methods and alternatives to web scraping are considered and the data context is sufficiently scoped to ensure validity and to identify other relevant information that may be valuable. It is clear why the data was ultimately collected from the focal website/API, and not from others (i.e., the website/API emerges as the one that fits best in terms of research fit and resource use) | Thorough comparison of websites and APIs with clear justification of the data source, including extraction methods and fit. | Reasonable comparison, but lacks depth in assessing extraction methods or rationale for the chosen source. | Minimal or unclear assessment of websites and APIs. The choice of data source is not well-justified. |
| The team provides a rich set of contextually relevant information. | The data context is thoroughly mapped, providing an in-depth understanding of the underlying data structure. The potential influence of algorithms and platform updates on data validity is addressed. | Data context is reasonably mapped, but could benefit from exploring the potential influence of platform algorithms and/or changes to user interface. | Data context mapping is limited. Influences of platform changes or algorithms not considered. |
| **2. Data Extraction Plan (section 2 of your documentation, 10%)** | | | |
| The risk of algorithmic interference is taken into account and dealt with accordingly. Furthermore, possible changes to the contents of the website or data aggregator that may influence the results are considered and metadata is collected, if applicable. | Potential for algorithmic inference and their impact on data collection is considered and thoroughly addressed. | There is a basic recognition of algorithmic interference, but a more detailed explanation of the potential changes to the data collection process and how they would be addressed would strengthen the argument. | Algorithmic interference is not sufficiently addressed and/or the suggested solution(s) to overcome it is not robust. |
| The seed selection is valid and clearly explained. Potential linkages to external data sets are made explicit (e.g., by means of links to external websites or sources that explain more about the used identifiers). | Valid seed selection. Potential links to external sources are clearly defined and explained. | The seed selection is discussed but requires a more robust justification. Additionally, the potential connections to external sources should be elaborated further. If the data used is self-contained, this should be clearly stated | The seed selection and potential external sources are not well explained. It would be helpful to identify and clarify the connections between your selected seeds and other available data sources |
| The frequency at which the data is the collected and the limitations to this are made explicit. If it is opted to collect data more than once, teams used automatic scheduling to ensure valid and consistent results. | The frequency of data collection and its associated limitations are clearly outlined. A robust automatic scheduling approach is implemented for instances where data is collected multiple times. | Data collection frequency is mentioned, but the rationale could be strengthened and potential limitations should be made explicit. | The frequency of data collection is not adequately explained. |
| The design decision lead to a tradeoff between validity, technical feasibility and exposure legal/ethical risks. The consequences to these are carefully described when making decisions on one of the previous steps (i.e., which information to extract, which seeds to select and at what frequency). | The trade-offs between validity, technical feasibility, and legal/ethical risks are carefully considered, with well-reasoned solutions effectively addressing these challenges. | The trade-offs between validity, technical feasibility, and legal/ethical risks are acknowledged, but the solutions to address them needs to be more robust and thoroughly developed. | There is little to no discussion of the trade-offs in design decisions. Consider outlining how you balanced validity with technical feasibility and legal/ethical concerns. |
| Teams explicitly address potential confidentiality or sensitivitiy of the data. | The potential confidentiality or sensitivity of the data are appropriately addressed, with clear measures for data protection and ethical handling of sensitive information. | Confidentiality and data sensitivity issues are mentioned, but the measures to circumvent them needs to be more robust. | There is insufficient consideration of confidentiality and data sensitivity. |

| Criteria | Very good | Sufficient | Needs Improvement |
|---|---|---|---|
| **3. Data Extraction Process (section 3 of your documentation, 10%)** | | | |
| The technical extraction plan has been described in a way that the data collection could be replicated. This encompasses providing a solid argumentation on why a particular data extraction technology used (e.g., selenium versus Beautifulsoup for websites, a package versus self-coded requests for APIs). If teams came across technical issues when scaling the data collection, the debugging stage is clearly explained. | The technical extraction plan is exceptionally clear and allows for full replication. The team provides a robust and well-argued rationale for choosing either web scraping (e.g., Selenium vs. BeautifulSoup) or API integration (e.g., self-coded requests vs. existing packages). For teams using APIs, considerations of rate limits, authentication, and data structuring are thoroughly explained. For web scraping teams, technical aspects like page structure changes and potential obstacles (e.g., captchas) are well-managed and described. Any technical issues encountered during scaling, including handling API limitations or overcoming web scraping obstacles, are clearly documented, and the debugging process is comprehensively explained. | The technical extraction plan is adequately explained and provides enough detail for the process to likely be replicated. The team offers some reasoning for selecting either web scraping (Selenium vs. BeautifulSoup) or API integration, though more depth would improve clarity. For API-based projects, explanations of key factors like rate limiting or authentication processes are present but could be more detailed. For web scraping, the handling of dynamic page elements or other technical hurdles is mentioned but somewhat superficial. Debugging steps are described but lack detail on how issues were resolved during scaling, whether related to API rate limits or web scraping challenges like blocking mechanisms. While the plan is functional, it would benefit from more detailed technical explanations. | The technical extraction plan lacks sufficient detail for reliable replication. The justification for choosing web scraping or API integration is weak or missing, with little to no discussion of why the specific method (e.g., Selenium, BeautifulSoup, or an API package) was selected. For API-based projects, critical details such as handling rate limits, authentication, or response structure are either unclear or absent. Similarly, for web scraping, the plan does not adequately address challenges like handling dynamic content or page structure changes. The debugging process is poorly explained, with little detail on how issues, whether related to API limitations or web scraping scaling, were resolved. Significant improvements are needed to justify the technical approach and clarify the handling of technical challenges. |
| Users of the data learn about the technical hurdles that needed to be overcome, and which monitoring was in place to guarantee (and validate) data quality. | The team clearly explained the technical hurdles and provided detailed insights into how they overcame them. Effective monitoring was in place to ensure data quality, with clear validation methods described. Users of the data can easily understand the challenges and how data integrity was maintained. | The team identified the key technical hurdles and gave a reasonable explanation of how they were addressed. Monitoring was in place, but the methods for ensuring and validating data quality could be more clearly explained. Users will gain some understanding of the challenges, though more detail would be beneficial. | The explanation of technical hurdles is unclear or missing, and the monitoring process is inadequately described. There is little information on how data quality was ensured or validated. Users of the data would struggle to understand how challenges were managed and data integrity maintained. |
| Details are given on how (deployment infrastructure) and when the data collection was executed (e.g., by meaningful summaries of the timestamps in log files), and where the final data set was stored during the collection. | The team provides clear details on the deployment infrastructure and a well-structured summary of when the data collection occurred, supported by meaningful timestamp summaries. The final data set’s storage location during collection is clearly explained, ensuring full transparency of the process. | The team gives a basic explanation of the deployment infrastructure and some information on when data collection was conducted. Timestamps are provided but lack depth in summarization. The storage of the final data set is mentioned but could be more clearly detailed. | The deployment infrastructure is unclear or insufficiently explained, and little to no information is provided on the timing of data collection. Timestamps are either missing or poorly summarized. The location of the data set during collection is not adequately detailed. |

| Criteria | Very good | Sufficient | Needs Improvement |
|---|---|---|---|
| **4. Preprocessing (section 4 of your documentation, 5%)** | | | |
| Any pre-processing on the fly has been motivated and explained, using a few specific examples. Any further pre-processing after the collection has been described (e.g., such as to anonymize users for privacy concerns, to identify and clean out implausible observations, or to improve data structure for long-term storage, such as rearranging the data structure, relabeling columns into more meaningful and clear variable names). Potential threats that may result from this pre-processing are brought up and elaborated on. | The pre-processing steps are well-motivated, clearly explained with specific examples, and thoroughly address tasks like anonymization, cleaning, and improving data structure. Potential threats are thoughtfully identified and elaborated upon. | The pre-processing steps are adequately explained, though additional examples or detail would strengthen the justification. Key tasks are addressed, but some aspects and potential risks could be expanded upon. | The pre-processing steps are insufficiently explained or lack motivation. Key tasks, such as cleaning or anonymization, are missing or unclear, and potential threats are not adequately addressed. |
| The files have a correct data structure, and variables are of the correct type (e.g., numbers as integers or floats, not as strings; time stamps properly formatted, or Unixtime used). Application of data enrichment and feature engineering strategies (e.g., to derive new variables from the data, where necessary). Data has been normalized (i.e., preferably multiple tables that can be joined together, rather than a wide table that contains many duplicates on some of the variables). If imputation is used, it is indicated which values have been imputed (e.g., interpolated; for example: followers (without missing), and followers_inputed as a TRUE/FALSE variable, indicating which ones were imputed). Finally, the data set is provided in CSV files, including column names, proper use of delimiters (e.g., a "," may be inappropriate for textual data involving commas). No row names/index column. | The data files have a correct structure, with variables of appropriate types and timestamps properly formatted. Data enrichment and feature engineering are effectively applied, and normalization is well-implemented with joinable tables. Imputation, where used, is clearly indicated with appropriate markers. The dataset is provided in clean CSV format, with proper delimiters and no unnecessary row names or index columns. | The data files are correctly structured, but there are minor issues with variable types or formatting (e.g., timestamps or delimiters). Some enrichment or normalization is present, but it could be more comprehensive. Imputation, if used, is mentioned but lacks clear markers. The dataset is provided in an acceptable CSV format, but small improvements in formatting could enhance usability. | The data files lack proper structure, with incorrect variable types or poorly formatted timestamps. Enrichment, feature engineering, and normalization are minimal or missing. Imputation, if used, is not clearly indicated. The dataset format has issues, such as improper delimiters, row names, or index columns, requiring significant revisions for usability. |
| **5. Data inspection (section 5 of your documentation, 15%)** | | | |
| The collected data is accompanied by meaningful summary statistics (e.g., the number of units per entity, means/SD for continuous variables, and frequency distributions per variable, for each entity). | The collected data is accompanied by comprehensive and meaningful summary statistics, including counts per entity, means and standard deviations for continuous variables, and frequency distributions for categorical variables. These summaries provide clear and valuable insights into the dataset. | The collected data includes basic summary statistics, such as counts, means, and frequency distributions, but some details are missing or could be expanded for greater clarity. | The collected data lacks sufficient summary statistics. Key details, such as counts, means, standard deviations, or frequency distributions, are missing, making it difficult to assess the dataset's overall structure and content. |
| Missingness has been investigated (e.g., for individual entities, but also for the collected variables). | Missingness has been thoroughly investigated, with detailed analysis at both the entity and variable levels. The results are clearly documented and provide valuable insights into potential data gaps. | Missingness has been investigated to some extent, but the analysis lacks depth or is limited to either entities or variables. Additional detail would improve understanding of the data gaps. | Missingness has not been adequately investigated. Key details about missing data at the entity and variable levels are absent, leaving potential gaps unaddressed. |
| Any redundancies, errors, or sources of noise have been clearly described. Identified subpopulations are labeled, so that users of the data can more easily get started using the data. | Redundancies, errors, and sources of noise are clearly identified and described, with well-documented steps to address them. Subpopulations are effectively labeled, making the dataset user-friendly and easy to navigate. | Some redundancies, errors, or noise are described, but the explanation could be more comprehensive. Subpopulations are labeled, but additional clarity or detail would enhance usability. | Redundancies, errors, or sources of noise are not adequately described, and subpopulations are either unlabeled or insufficiently documented, limiting the dataset's usability. |
| **6. Uses (section 6 of your documentation, 5%)** | | | |
| Users of the data learn about tasks the data set could be used for. I.e., from the description, it is clear how the composition of the data set or the way it was preprocessed might affect future use. A clear indication is given for what the data should not be used for, e.g., relating to any of the legal or ethical concerns identified before. | The dataset description provides clear and comprehensive guidance on potential tasks the data can be used for, with specific examples. The impact of the dataset's composition and preprocessing on its future use is thoroughly explained. Clear and explicit indications are provided for inappropriate uses, including any relevant legal or ethical concerns. | The dataset description outlines potential tasks it could be used for, but the examples or explanations about the impact of its composition and preprocessing are somewhat limited. Some guidance on inappropriate uses is provided, but it could be more detailed or specific. | The dataset description lacks sufficient information about potential tasks it could be used for, with little to no explanation of how its composition or preprocessing might affect future use. There is no or minimal guidance on inappropriate uses or related legal and ethical concerns. |