Title of your data set  
*Data documentation template adapted by Hannes Datta.   
Originally based on Gebru, Morgenstern, Vecchione, Vaughan,   
Wallach, Daumeé, and Crawford. (2018). Datasheets for Datasets.*

# 1. Motivation[[1]](#footnote-1)

***1.1*** *What primary research question, business problem, or knowledge gap motivated the creation of this dataset (“data context”)? How does the dataset offer insights into new phenomena, contribute to developing new models, or streamline gathering essential information? Why is this dataset valuable to the broader research community or industry stakeholders?*

***1.2*** *Please compare the various websites and APIs you assessed relevant to your data context. A table may help to compare data sources. Why did you choose your specific data source? Discuss the research fit, extraction method (e.g., web scraping vs. APIs), efficiency of resource use, and any other factors that made it emerge as the best choice. For tips, see challenges 1.1 and 1.2 in Boegershausen et al. 2022.*

***1.3*** *Please identify potentially relevant contextual information (Boegershausen et al. 2022, challenge 1.3). Provide an overview here and save any relevant information in the digital submission of your project.*

***1.4*** *Who created this dataset (e.g., which team). Mention you are students of the Marketing Analytics program at Tilburg University.*

# 2. Data Extraction Plan

***2.1*** *Please describe your data extraction plan in such a way that another researcher or team could replicate your data collection process. Which information to extract from which pages, how to sample, at which frequency to extract the data, and how to process the data during the collection. In your description, explain how you tackled the various validity, legal/ethical, and technical challenges. See Boegershausen et al. (2022, challenges 2.1-2.4) for tips.*

***2.2*** *Does the dataset contain data that might be considered confidential or sensitive (e.g., personal data such as usernames or IP addresses, data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, financial or health data, biometric or genetic data, social security numbers, passwords, etc.)? If so, please provide a description.*

***2.3*** *If the dataset relates to people, is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.*

# 3. Data Extraction Process

***3.1*** *When was the data collected?*

***3.2*** *Please describe any technical challenges you encountered while scaling your data collection. How did you resolve them? Please provide a clear explanation of the debugging process (see Boegershausen et al. 2022, challenge 3.1).*

***3.3*** *What measures or monitoring systems were in place to ensure and validate the quality of the extracted data? Can you describe how these monitoring systems functioned? (see Boegershausen et al. 2022, challenge 3.2).*

***3.4*** *Can you specify the infrastructure you used for the deployment and execution of your data collection?*

# 4. Preprocessing, cleaning, labeling

***4.1*** *After collecting the data, did you perform any data processing? If yes, please provide specific examples and explain the reasoning behind each step.*

***4.2*** *Were any measures implemented to ensure privacy, such as anonymizing user data? Please describe the methods used.*

***4.3*** *How did you address and clean out any implausible or erroneous observations in the dataset?*

***4.4*** *Did you modify the data structure for long-term storage, like rearranging the dataset or renaming columns for clarity? If so, provide details on these changes and their rationale.*

***4.5*** *What potential threats or biases could arise from your pre-processing steps? Please elaborate on any risks associated with the modifications made to the data and how they might impact the dataset's integrity or utility.*

# 5. Data inspection

***5.1*** *Please provide a variety of meaningful summary statistics and plots. For example, consider means/SDs for continuous variables, frequency distributions for categorical variables or – in the case of plots – bar charts, line plots, or histograms. This part of the documentation is intended to illustrate the richness of the collected data.*

***5.2*** *Is any information missing from individual instances? If so, please describe why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information but might include, e.g., redacted text.*

# 6. Uses

***6.1*** *Has the dataset been used for any tasks already? If so, please provide a description.*

***6.2*** *What (other) tasks / research projects could the dataset be used for? Provide a set of potential research questions or ideas for research projects.*

***6.3*** *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?*

***6.4*** *Are there tasks for which the dataset should not be used? If so, please provide a description.*

1. \* https://arxiv.org/abs/1803.09010 [↑](#footnote-ref-1)