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

Datasheets for Datasets  
*Adapted from:* *Gebru, Morgenstern, Vecchione, Vaughan,   
Wallach, Daumeé, and Crawford. (2018). Datasheets for Datasets.*

Derive

***1.1*** *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*

This dataset was created to gain insight into the electronics industry, specifically focusing on the offerings of Mediamarkt. Thanks to the wide range of products on offer at Mediamarkt, various analyses can be carried out on, for example, prices per brand and the number of different types of appliances. The idea was born when someone told that he was looking for a new phone, but it should not cost too much. So, Mediamarkt was visited, and the scraping started. With this dataset you have the possibility to easily compare different devices on price, reviews, brand, and different kind of specifications like Wi-Fi and Bluetooth.

***1.2*** *Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?*

This dataset was developed by group 7 of the Online Data Collection & Management course. This course was taken at Tilburg University during the Marketing Analytics master's program.

***1.3*** *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*

There was no funder for the creation of this dataset.

# 2. Composition

The distribution in electronic devices and brands on Mediamarkt.com  
*Adapted from:* *Gebru, Morgenstern, Vecchione, Vaughan,   
Wallach, Daumeé, and Crawford. (2018). Datasheets for Datasets.\**

Derive

***2.1*** *What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of in- stances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

The dataset exists out of 4 product categories: smartphones, televisions, tablets, and laptops. All brands are included in the scrapping process, including Apple. This makes it possible to analyze which brands are generally more expensive, but it also makes it possible to analyze which brand is the most expensive or the best by category.

***2.2*** *How many instances are there in total (of each type, if appropriate)?*

|  |  |
| --- | --- |
| Product category | Number of products |
| Televisions | 254 |
| Smartphones | 472 |
| Tablets | 264 |
| Laptops | 608 |

***2.3*** *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).*

The dataset contains all products in earlier mentioned categories. But Mediamarkt of course does have more product categories than the four mentioned. The reason why these categories were chosen is because they all fall under the category of image & sound and telephony & navigation. And the brands in these categories are present in all categories. That is useful for the analyses.

***2.4*** *What data does each instance consist of? “Raw” data (e.g., unprocessed text or images) or features? In either case, please provide a de- scription.*

|  |  |
| --- | --- |
| *Name of variable* | *Description* |
| Device type | One of the four product categories. |
| Name | Name of product. |
| Instock | If the product is in stock or not. |
| Rating | Average review rating. |
| Nr\_reviews | Number of reviews a product has. |
| Date\_and\_time | This is the date and time that product was scraped. of course, the offer of mediamarkt changes often, therefore the date is added to make it clear when the dataset was created. |
| attributes | This variable exists out of sub variables. These sub variables are all the specifications a device has. |

These sub attributes are the following:

|  |  |
| --- | --- |
| *Sub attribute* | *Description* |
| Type apparaat | Device type (Smartphone, television, laptop or tablet. |
| Besturingssysteem | This is the operating system, so for example android or IOS. |
| Processor | This is the hardware of a device. |
| Geheugencapaciteit | Memory capacity, how much GigaByte a device has. |
| Geheugenkaartformaat | Memory card format, this is for example a micro-sd card in a phone. |
| Werkgeheugen | This is the working memory a device has, it is expressed in GigaByte. |
| Sensoren | Which sensors a device has. For a smartphone, this can for example be an fingerprint scanner to unlock a phone. |
| Maximale capaciteit geheugenkaart | This is the maximum capacity of the memory card. This can be expressed in Gigabytes or in Terabytes. |
| Simkaartformaat | Sim card format |
| Artikelnummer | This is the number a product has. |

***2.5*** *Is there a label or target associated with each instance? If so, please provide a description.*

There is no label or target associated with each instance.

***2.6*** *Is any information missing from individual instances? If so, please provide a description, explaining 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.*

Some information was not available or just was not there. For the main variables, so; name, Instock, Rating, and Nr\_reviews, if the information was not available, we mentioned it in the dataset with for example   
“no name” or “no rating”.

***2.7*** *Are relationships between individual instances made explicit (e.g., users’ movie ratings, social network links)? If so, please describe how these relationships are made explicit.*

There are no relationship instances.

***2.8*** *Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*

There are no recommended data splits.

***2.9*** *Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they ex- isted at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

The dataset is self-contained.

***2.10*** *Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctorpatient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.*

The dataset does not contain confidential data. The data is publicly available on the Mediamarkt website.

***2.11*** *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*

The dataset does not contain data that might be offen-sive, insulting, threatening or might cause anxiety.

***2.12*** *Does the dataset relate to people? If not, you may skip the remaining questions in this section.*

The dataset does not relate to people.

***2.13*** *Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

***2.14*** *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.*

***2.15*** *Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*

# 3. Collection Process

***3.1*** *How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by sub- jects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

All data collected was directly observable on the website.

***3.2*** *What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software pro- gram, software API)? How were these mechanisms or procedures validated?*

The data was collected by a self-written webscraper in Python. Comparing the information with the selected product page validated the data collected. The selection of pages to validate was chosen randomly.

***3.3*** *If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?*

The data was not a sample of a larger set, however it does not include all data from Mediamarkt.nl. Four separate categories were specifically chosen all connected with technology.

***3.4*** *Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?*

Throughout the data collection process only the four before mentioned students were involved in the process. The students were not compensated by their work as this was done for a course at Tilburg University.

***3.5*** *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the time- frame in which the data associated with the instances was created.*

The data was scraped within 45 minutes. We do not expect Mediamarkt to change much information of the product pages within these 45 minutes of scraping. The scraper was executed multiple times and no differences were detected.

***3.6*** *Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

No ethical review process was conducted beforehand or during the collection process.

***3.7*** *Does the dataset relate to people? If not, you may skip the remaining questions in this section.*

This data set did not relate to people.

***3.8*** *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?*

***3.9*** *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or other- wise reproduce, the exact language of the notification itself.*

***3.10*** *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

***3.11*** *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

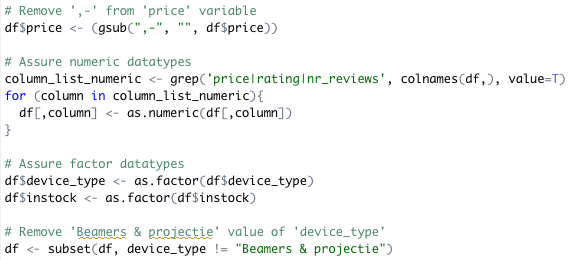
***3.12*** *Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

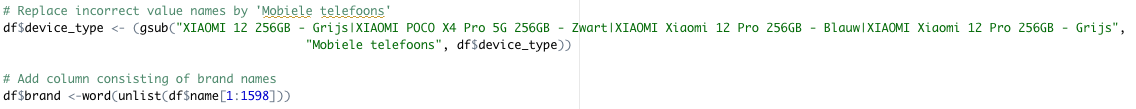
# 4. Preprocessing, cleaning, labeling

***4.1*** *Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remain- der of the questions in this section.*

During the writing of the data scraper in Python, cleaning the data was already taken into account. Only necessary data was scraped.

However, in Rstudio when data was analyzed, the “raw” data was cleaned/prepared. The following code was used:





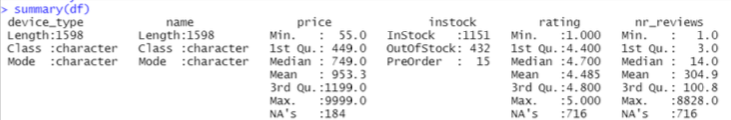
As can be seen on the screenshots, the data was first cleaned by removing unnecessary characters. After, the categories price, ratings and number of reviews were made numeric to ensure the correct data type.

Waarom beamers etc weg?

As some categories did not include the correct value name, this is added by replacing an incorrect name with “mobiele telefoons”. These belong to that same category and therefore can now be used in the analysis.

Lastly, a column was added with the brand names. This is added so the data set can easily be used to analyze a difference in brand names. The brand name was not a specification on itself when scraping so adding this column was done manually.

After cleaning the data, the following overview was created:



This overview shows the dataset contains information about 1598 devices, of which 1151 currently in stock at MediaMarkt. The price distribution was given also with a mean of 953.30 EUR. The maximum and minimum price differ much with a min of 55.00 EUR and a max price of 9999.00 EUR.

What stands out for now if the information about the ratings. The mean number of ratings is 304.9, while the maximum number of ratings is 8828.0. For now this is much higher than we had expected.

The mean rating is interesting as well, with a 4.485. This shows (in this summary) that consumers are satisfied with their product.

This summary is a solid foundation to start and to compare further analysis with.

***4.2*** *Was the “raw” data saved in addition to the prepro- cessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the “raw” data.*

Our python coding script, XXXXFINALNAMEXXX, includes code that generates json and csv files when run. These files include all data scraped from the Mediamarkt and are the “raw” data files before importing into R. The raw datasets can also be found in the folder “data” for more easy access.

***4.3*** *Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.*

We used a self-made scraper to obtain the data from the website Mediamarkt.com. This scraper was written in Python using Jupyter notebook. Jupyter notebook was installed through Anaconda.

During the writing of the code, preprocessing, cleaning and labeling the data was already taken into account. No other software was used for this purpose. After extracting the data and rewriting the data in csv and json, the set was imported in Rstudio to analyze.

Both programs, Jupyter notebook and Rstudio, are publicly available to download on the following websites:

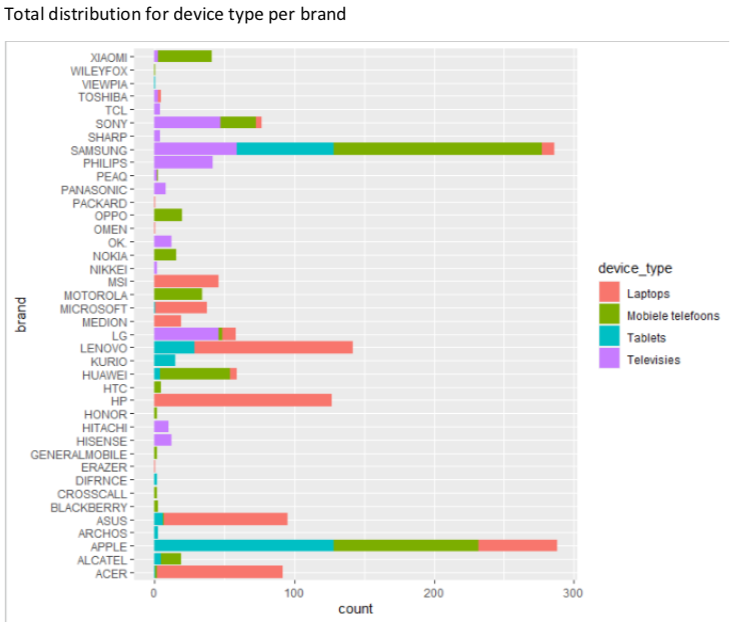
- Jupyter notebook: <https://tilburgsciencehub.com/building-blocks/configure-your-computer/statistics-and-computation/python/>

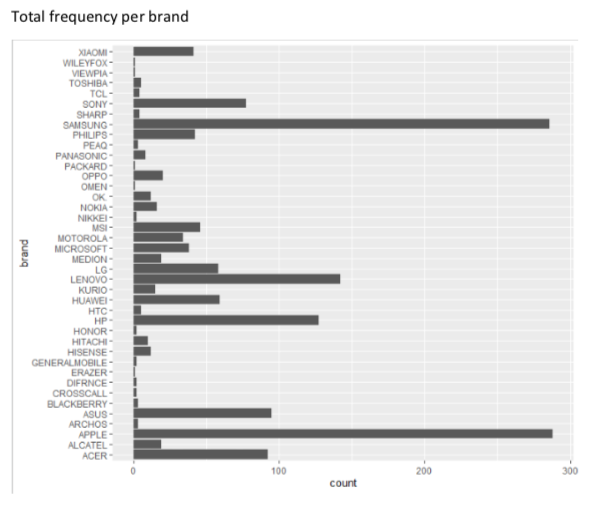
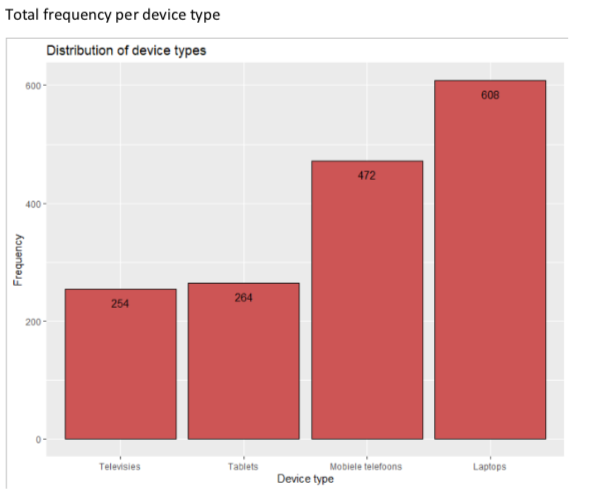
- Rstudio: <https://tilburgsciencehub.com/building-blocks/configure-your-computer/statistics-and-computation/r/>

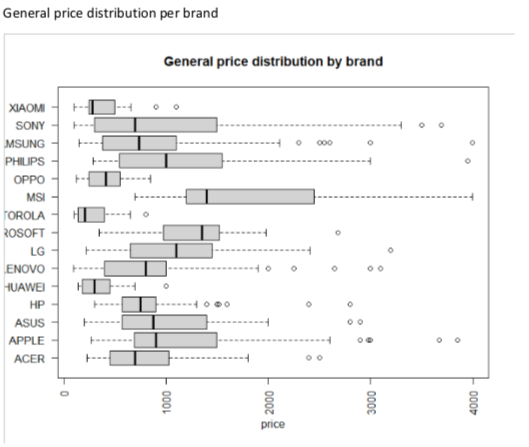
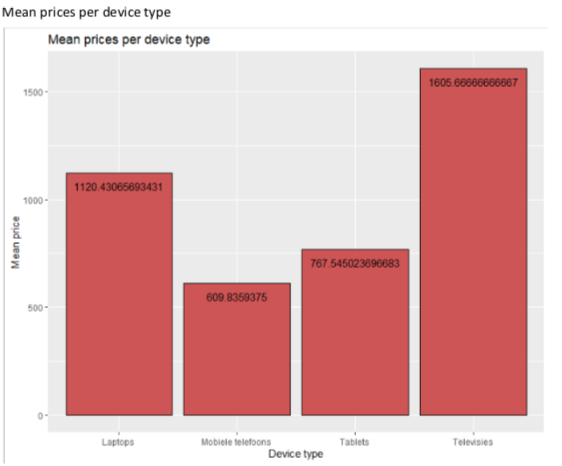
# 5. Uses

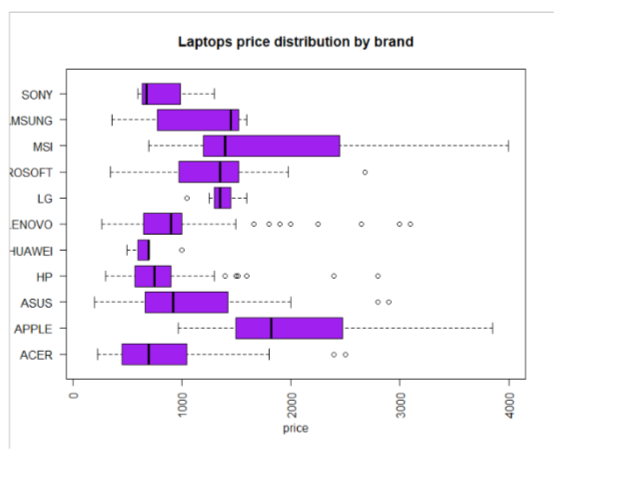
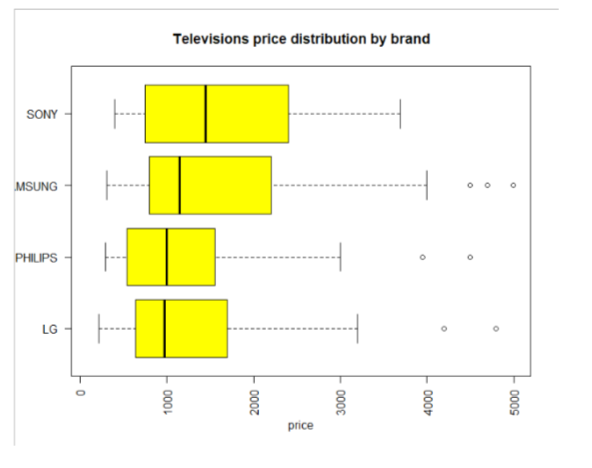
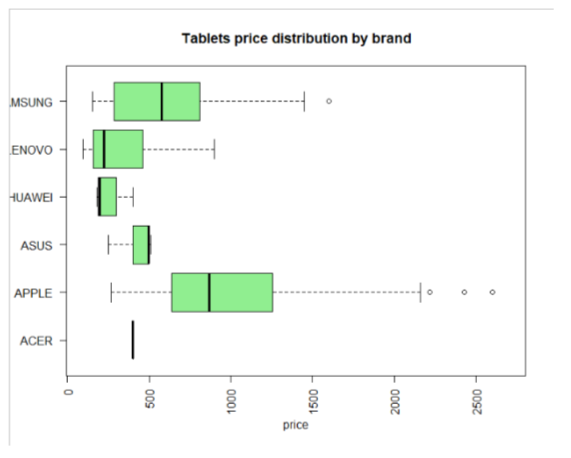
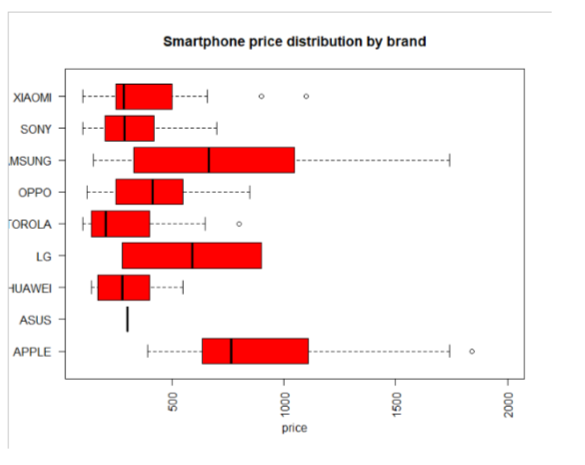
***5.1*** *Has the dataset been used for any tasks already? If so, please provide a description.*

This dataset has been used to conduct a statistical analysis in Rstudio. The following output was generated:









***5.2*** *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*

This dataset has not been used yet other than to generate the input mentioned in chapter 5.1. Therefore, no repository can be linked.

***5.3*** *What (other) tasks could the dataset be used for?*

This data set can be used for multiple purposes. An example could be that a competitor of Mediamarkt.com would like to know the customer ratings of products so they can choose what to sell (or not to sell). The data set can also for example be used to identify differences between the scraped categories.

As the information scraped includes most of the aspects from the product pages, the dataset can be used so you do not have to go through all pages instead. The dataset is therefore valuable for companies, but if made publicly available to consumers as well. Consumers can use this set to analyze the differences between products so they can make a well-informed choice.

***5.4*** *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?*

If our data scraper is used in the future, this should not have a great impact. Our scraper obtains information directly from Mediamarkt.com, so if there are changes made on the website this will also change the dataset slightly. For example, if a price of a product is adjusted this will be shown in the data as well.

The only time misinformation can be spread is if people use our prepared data file (in csv) to retrieve data, while Mediamarkt.com changed these aspects on their website.

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

The information is public information scraped from Mediamarkt.com to be more easily accessible to analyze. However, this dataset is very specific so it only reflects information of Mediamarkt.com. Therefore, this set can only be used to analyze Mediamarkt.com specifically and this dataset does not reflect information outside this scope.

# 6. Distribution

***6.1*** *Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*

***6.2*** *How will the dataset be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*

***6.3*** *When will the dataset be distributed?*

***6.4*** *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*

***6.5*** *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*

***6.6*** *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*

# 7. Maintenance

***7.1*** *Who will be supporting/hosting/maintaining the dataset?*

***7.2*** *How can the owner/curator/manager of the dataset be contacted (e.g., email address)?*

***7.3*** *Is there an erratum? If so, please provide a link or other access point.*

***7.4*** *Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?*

***7.5*** *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*

***7.6*** *Will older versions of the dataset continue to be sup- ported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.*

***7.7*** *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.*

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