Datasheet for the ASOS Digital Experiments Dataset

C. H. Bryan Liu [⋆†] Ângelo Cardoso [†] Paul Couturier [⋆] Emma J. McCoy [⋆]
[⋆] Imperial College London & [†] ASOS.com, UK

Oct 2021 (Version 1)

This document describes the motivation, composition, collection process, recommended uses, etc. of the ASOS Digital Experiments Dataset, a public dataset that supports the end-to-end design and running of Online Controlled Experiments (OCEs) with adaptive stopping. The document is based on "Datasheets for Datasets" by Gebru et al. (2018).

The dataset is released in conjunction with the paper "Datasets for Online Controlled Experiments" by Liu et al. (2021).

Motivation

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.
The dataset is released to address the gap in publicly available Online Controlled Experiment (OCE) dataset that can support research on the end-to-end design and running of OCEs with adaptive stopping. For a detailed motivation of the problem, please refer to the "Datasets for Online Controlled Experiments" paper.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?
This dataset is created by the AI & Data Science Platform at ASOS.com, in collaboration with the EPSRC Centre for Doctoral Training in Modern Statistics and Statistical Machine Learning at Imperial College London and University of Oxford (StatML.IO).

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.
The dataset, as part of the wider research work on surveying and building a taxonomy for OCE datasets, is part funded by the EPSRC Centre for Doctoral Training in Modern Statistics and Statistical Machine Learning at Imperial College London and University of Oxford (StatML.IO) and ASOS.com.

Any other comments?
No.

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.
The dataset contains one relational table, with each row containing a set of business performance measurements for a particular experiment, treatment, and decision metric, covering a specific time period of the said experiment. Please refer to the schema hosted on the Wiki page on the OSF project for further information.

How many instances are there in total (of each type, if appropriate)?
There are 24,153 rows in the dataset, recording the results of 78 online controlled experiments. Each experiment has 2–5 variants (including one variant acting as the control), with a total of 99 treatments being experimented. Four decision metrics, specific to the business unit which ran the experiments, are captured with a daily or bi-daily frequency.

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 representa-

tiveness 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 records the results of experiments run by a specific business unit in ASOS.com, a global online fashion retail platform mainly targeting 20-somethings based in North America, Europe, and Oceania. The experiments are run at various times over a 2-year period (2019 to 2020). Within the said scope the dataset contains all possible rows for all valid experiments. Rows associated to experiments that the team had to abandon due to incorrect setup on the experimentation platform have been removed as they do not carry any learning value.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.
Each row consists of six entries:

- Number of users in the control group
- Sample mean of the responses (under the corresponding decision metric) from users in the control group
- Sample variance of the responses from users in the control group
- Number of users in the treatment group
- Sample mean of the responses from users in the treatment group
- Sample variance of the responses from users in the treatment group

Please refer to the dataset schema for further information. All responses are aggregated up to the control/treatment group level via the use of summary statistics. The dataset does not identify or record any individual responses from users.

Is there a label or target associated with each instance? If so, please provide a description.
The six entries mentioned above act as the "label/target" for one to run statistical tests on.

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.
A small number of experiment/treatments (5 out of 99 treatments under Metric IDs 2, 3, and 4) are missing the variance entries due to an issue in capturing the underlying data.

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.
The rows are linked via the experiment ID, variant ID, and (decision) metric ID. The experiment ID uniquely identifies an experiment. The metric ID uniquely identifies a metric. The combination of the experiment ID and the variant ID uniquely identifies a treatment—The same variant ID can be seen in multiple experiments and hence can not be used to identify a treatment.

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.
N/A for the running of OCEs as it is the norm to apply a statistical test to all available data. We also have no specific data split recommendations when it comes to learning the statistical test hyperparameters from the data.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.
There are no duplicated rows in this dataset. However, common pitfalls in the running and analysis of online controlled experiments like sample ratio mismatch, novelty effect and telemetry loss may lead to unexpected data points. Some possible examples include:

- The treatment mean entry for Experiment '162a38', Metric 1, taken 38 days since the start of the experiment demonstrates a spike as compared to that of the control mean and treatment in surrounding time points, which could be a result of telemetry loss.
- The treatment mean entries for Experiment '54a85a' are increasing faster than the control mean entries of the same experiment, indicating a possible primacy effect in play.

Note we are unable to definitively tell whether the said pitfall applies as we are observing a real-life system without a ground truth. We intentionally leave these data in as we believe they are reflective of what an OCE practicioner will face in their day job. We hope the dataset will attract methods that address the underlying problem(s).

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 existed 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, as designed.

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

No. All entries in the table that are considered business sensitive have been anonymized (e.g. experiment ID) or removed (e.g. name of experiment, start date/time of experiment) to prevent third parties linking the entry back to a particular experiment being done on the digital platforms.

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

No.

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

The dataset records the result of business-side measurements. The measurements based on the activities of hundreds of thousands or millions of users and aggregated over all of them. There is no notion of an individual in the dataset. For prudence, we also provide a response to the following questions.

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.

No. The dataset does not identify any subpopulations based on protected characteristics. The dataset does record the aggregated activity attained by the treatment and control groups, which users are randomly assigned into.

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.

No. It is impossible to identify an individual from this data as each instance aggregates the activity of at least a hundred thousand individuals.

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.

No.

Any other comments?

No.

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (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.

The dataset is aggregated from a raw clickstream data, which records the browsing behavior of individual users. The raw clickstream data is *not* released along with this dataset.

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

The aggregation is done via a series of proprietary data processing pipelines within ASOS.com.

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

No sampling was carried out when aggregating the numbers. This is to ensure the calculated decision metrics are accurate.

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)?

The data collection is done via digital platforms built by ASOS.com and related third-parties. The plat-

forms are general-purpose and not built specifically to create this dataset.

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 timeframe in which the data associated with the instances was created.
The dataset describes results from OCEs that were run in years 2019 and 2020. The results are cleaned and curated in May 2021.

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.
Each OCE run within the business unit is subject to a review process, which among many things, include ethical considerations.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.
The dataset concerns business-side measurements, yet the nature of online controlled experiments means the dataset is derived from the activities of millions of users to the digital platforms. For prudence, we will provide a response to the remaining questions in this section.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?
The data is ultimately based upon users' browsing behavior, which is recorded on digital platforms built by ASOS.com and related third-parties.

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 otherwise reproduce, the exact language of the notification itself.
Yes. Users are notified of the data collection policy[1] and privacy policy[2] via a banner upon visiting the website. The privacy policy describes how the information will be used under the section "How we use your information":

- "We also anonymise and aggregate personal information (so that it does not identify you) and use it for purposes including testing our IT systems, research, data analysis, improving our site and app, and developing new products and services."
- "We use the data we collect to help us provide you with the best service, the best shopping experience and to show you the latest and greatest products and services that we think you will love."

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.
Yes. Users have to consent to the data collection in order to be able to use the service.

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).
Yes. Users are made aware of ASOS.com's Cookies Notice[1] and Privacy Policy.[2] The latter provides a mechanism for users to revoke their consent in the future / for certain use subject to the relevant Data Protection laws.

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.
The Data Protection Officer at ASOS.com has reviewed the dataset and its schema, and given each measurement is reflective of the collective activity of hundreds of thousands to millions of users, determined there are no data protection risk from the publication of this dataset.

Any other comments?
No.

Preprocessing/cleaning/labeling

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 remainder of the questions in this section.
The measurements contained in this dataset is aggregated from a raw clickstream dataset via a series of

[1] https://www.asos.com/discover/marketing-terms-and-conditions/privacy-policy-cookies/
[2] https://www.asos.com/privacy-policy/

proprietary data pipelines. Some manual data cleaning is carried out to remove invalid experiments that do not carry any learning value.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.
The raw clickstream data that this dataset is derived from is proprietary to ASOS.com. The data is stored subject to applicable laws and regulations but is not (and will not be) publicly available.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.
No. The data processing pipelines that curate this dataset are proprietary to ASOS.com.

Any other comments?
No.

Uses

Has the dataset been used for any tasks already? If so, please provide a description.
We demonstrated in the "Datasets for Online Controlled Experiments" paper that the dataset can support the development in OCE methods via the following:

1. Running of meta-analyses
2. Design and running of experiments with adaptive stopping
3. Acting as a quasi-benchmark for adaptive stopping methods

For details, please refer to Section 5.1 of the "Datasets for Online Controlled Experiments" paper.

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.
No. Although we invite individuals who have benefited from the use of this dataset to cite either the dataset or the broader work which also provides a survey and taxonomy on publicly available OCE datasets. The resultant citation network may provide a glimpse on the dataset's impact.

What (other) tasks could the dataset be used for?
Being both a multi-experiment and time series dataset, we believe the dataset can also be useful for:

1. Bias (of the estimator) detection across time; and
2. Learning of correlation structure across experiments, metrics, variants and time

See Section 5.1 of the "Datasets for Online Controlled Experiments" paper for a more detailed discussion.

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?
We urge potential users of this dataset to exercise caution when attempting to generalize the learnings. Examples of generalizing the learnings include the use of this dataset as a full performance benchmark and directly applying the value of hyperparameter(s) obtained while training a model on this dataset to another dataset.
Please refer to Section 5.2 of the "Datasets for Online Controlled Experiments" paper for a detailed discussion on this matter.

Are there tasks for which the dataset should not be used? If so, please provide a description.
The dataset is made available with the intent to support development in the statistical methods required to run OCEs. The experiment results shown in the dataset is not representative of ASOS.com's overall business operations, product development, or experimentation program operations, and no conclusion of such should be drawn from this dataset.

Any other comments?
No.

Distribution

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.
Yes. The dataset is open-sourced under a CC-By Attribution 4.0 International (CC BY 4.0) License.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub) Does the dataset have a digital object identifier (DOI)?
The dataset can be downloaded from Open Science Framework (OSF) via the link https://

osf.io/64jsb/. The DOI for this dataset is 10.17605/OSF.IO/64JSB.

When will the dataset be distributed?
The dataset is currently available to the public.

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.
Yes, the dataset is distributed under a CC-By Attribution 4.0 International (CC BY 4.0) License.

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.
No.

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.
No.

Any other comments?
No.

Maintenance

Who will be supporting/hosting/maintaining the dataset?
The primary contributor, C. H. Bryan Liu, will be the main contact for support, hosting, and maintenance of the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
The primary contributor, C. H. Bryan Liu, can be contacted via the email address bryan.liu12 (at) imperial.ac.uk.

Is there an erratum? If so, please provide a link or other access point.
No. In the case where one is necessary, we will update the datasheet to reflect the change.

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)?
We commit to update the dataset in the case where major error(s) in the measurements are found. We have no plans to add new rows describing the result of more recent experiments in the short term.

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.
N/A. The rows concern business measurements and does not record the activity of an individual directly.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.
We commit to supporting older versions in the case where we are in the position to add new instances corresponding to measurements arising from new experiments. If minor error(s) in the measurements were to be found, we will update the existing dataset (i.e. new version of the data file under the same file name on OSF) and datasheet. If major error(s) in the measurements were to be found, we will release a new dataset (i.e. upload the data file under a different file name on OSF) and add an errata to the publication associated with the dataset. The old, erroneous version would be discontinued (i.e. hidden from view on OSF).

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.
There are no mechanisms/process developed for this purpose yet. We welcome suggestions from any potential contributors.

Any other comments?
No.

Acknowledgment

We thank the internal and external reviewers for suggesting many improvements to the datasheet. We also thank Christian Garbin for making a LaTeX template of Datasheet for Datasets freely available.

References

C. H. B. Liu, Â. Cardoso, P. Couturier, E. J. McCoy (2021) Datasets for Online Controlled Experiments. In: *NeurIPS'21 Datasets and Benchmarks.* URL: https://openreview.net/forum?id=79shW3z5Eaq.

T. Gebru, J. Morgenstern, B. Vecchione, J. Wortman Vaughan, H. Wallach, H. Daumé III, K. Crawford (2018) Datasheets for Datasets. *arXiv Preprint arXiv:1803.09010* [cs.DB].

Changelog

- Version 1 (2021-10-29): Initial public version following the acceptance of the paper "Datasets for Online Controlled Experiments" into NeurIPS'21 Datasets and Benchmarks.
- Version 0.2 (2021-09-30): Clarified the privacy policy in force during the data collection phase and the target demographics of ASOS.com following second-round internal review and external review.
- Version 0.1 (2021-09-04): Updates based on first-round internal review.
- Version 0 (2021-08-27): Initial datasheet.