Datasheet for OkCupid Profile Data for Introductory Statistics and Data Science Courses

# Motivation for Dataset Creation

## **Why was the dataset created?** (e.g., were there specific tasks in mind, or a specific gap that needed to be filled?)

The OkCupid profiles dataset contains information from 59,946 San Francisco users. This dataset was scraped from active profiles containing at least one photo in June 2012 and made available to the public. This data was collected for the intention of creating a rich and imperfect dataset for undergraduate data science and statistical exploration and analysis.

## **What (other) tasks could the dataset be used for?** Are there obvious tasks for which it should *not* be used?

This data could be used for identifying specific product needs of individuals within the data. The specificity of the variables and direct identifiers enables one to use the data in conjunction with public data to identify a specific individual within this dataset with relative ease. While possible, this is a task that should not be done using this dataset because the data was obtained from scraping active profiles rather than receiving explicit permission to use the data for educational or analytical purposes.

## **Has the dataset been used for any tasks already?** If so, where are the results so others can compare (e.g., links to published papers)?

This data has been used for basic analysis and educational visualization. I have not found any academic publications on this dataset other than short articles such as the following: <https://rstudio-pubs-static.s3.amazonaws.com/209370_b62220c849b946088b463fdbec935848.html>

## **Who funded the creation of the dataset?** If there is an associated grant, provide the grant number.

This data was created by professors Albert Y. Kim of Middlebury College and Adriana Escobedo-Land of Reed College without the need for grant funding.

## **Any other comments?**

This rich dataset was created for undergraduate educational purposes and the literature I have been able to find are completed to the extent of a short exploratory assignment/project without intention of publishing in any respected journal.

# Dataset Composition

## **What are the instances?** (that is, examples; e.g., documents, images, people, countries) Are there multiple types of instances? (e.g., movies, users, ratings; people, interactions between them; nodes, edges)

Instances are an individual’s profile composed of basic information such as gender, sexual orientation, ethnicity, religion, education, and short essays answers to 10 prompts such as "Describe your typical Friday night" and "What is the first thing people notice about you?"

## **Are relationships between instances made explicit in the data** (e.g., social network links, user/movie ratings, etc.)?

There are no known relationships between instances except for the fact that they are all individuals who were located in the San Francisco Bay Area.

## **How many instances of each type are there?**

This dataset consists 59,946 instances of OkCupid profiles information from San Francisco users.

## **What data does each instance consist of?** “Raw” data (e.g., unprocessed text or images)? Features/attributes? Is there a label/target associated with instances? If the instances are related to people, are subpopulations identified (e.g., by age, gender, etc.) and what is their distribution?

Each instance contains information about the user’s body\_type, diet preference, drinking habits, drug habits, education, ethnicity, height, income, job, offspring, sexual orientation, pets, religion, sex, sign, smoking preference, languages spoken, relationship status, and answers to 10 open ended essay prompts listed below:

essay0- My self summary

essay1- What I’m doing with my life

essay2- I’m really good at

essay3- The first thing people usually notice about me

essay4- Favorite books, movies, show, music, and food

essay5- The six things I could never do without

essay6- I spend a lot of time thinking about

essay7- On a typical Friday night I am

essay8- The most private thing I am willing to admit

essay9- You should message me if...

## **Is everything included or does the data rely on external resources?** (e.g., websites, tweets, datasets) If external resources, a) are there guarantees that they will exist, and remain constant, over time; b) is there an official archival version. Are there licenses, fees or rights associated with *any* of the data?

While the data was originally collected from an external source, everything is included in the contained dataset.

## **Are there recommended data splits or evaluation measures?** (e.g., training, development, testing; accuracy/AUC)

There are no recommended data splits or evaluation measures.

## **What experiments were initially run on this dataset?**

## Have a summary of those results and, if available, provide the link to a paper with more information here.

Basic correlations within the data between two variables have been run on this dataset with conclusions such as, “Generally the OkCupid data reflects many trends in the general population, such as drinking habit and income disparities. Other trends, such as ethnicity, diverge from the more general demongraphics.” -https://rstudio-pubs-static.s3.amazonaws.com/209370\_b62220c849b946088b463fdbec935848.html

# Data Collection Process

## **How was the data collected?** (e.g., hardware apparatus/sensor, manual human curation, software pro- gram, software interface/API; how were these con- structs/measures/methods validated?)

Using a Python script, data was scraped from users’ public profiles on June 30, 2012; any non-publicly facing information such as messaging was not accessible.

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

OkCupid users who were living within 25 miles of San Francisco, had active profiles on June 26, 2012, were online in the previous year, and had at least one picture in their profile was included in this dataset. These individuals were not compensated.

## **Over what time-frame was the data collected?** Does the collection time-frame match the creation time-frame?

Although the instances were validated to be active profiles on June 26, 2012, the data was collected on June 30, 2012;

## **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 the latter two, were they validated/verified and if so how?

The data was collected directly from raw text submission from the OkCupid users. This raw data is, however, self-reported and therefore subject to bias and misrepresentation.

## **Does the dataset contain all possible instances?** Or is it, for instance, a sample (not necessarily random) from a larger set of instances?

This dataset does not contain all possible instances of OkCupid profiles within 25 miles of San Francisco. Instances without a profile picture, for instance, were not included in this dataset even though the profile picture was not included in the dataset.

## **If the dataset is a sample, then what is the population?**

## What was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? Is the sample representative of the larger set (e.g., geographic coverage)? If not, why not (e.g., to cover a more diverse range of in- stances)? How does this affect possible uses?

Although the sampling strategy was random, this data set sampling is representative of the larger OkCupid San Francisco users, but not representative of the San Francisco demographics within the same dating age range.

## **Is there information missing from the dataset and why?**

## (this does not include intentionally dropped instances; it might include, e.g., redacted text, withheld documents) Is this data missing because it was unavailable?

Apart from blank answers from variables or essay prompts not being answered by an individual, this data does not include any dropped instances.

## **Are there any known errors, sources of noise, or redundancies in the data?**

Yes, this dataset was not able to remove profiles created by the same user despite known information regarding the existence of some users creating multiple profiles for him/her/themselves. This results in potential redundancies that cannot be removed.

# Data Preprocessing

## **What preprocessing/cleaning was done?** (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values, etc.)

No preprocessing or cleaning was done to this dataset as it was intentionally left messy for students and other users of this dataset to clean.

## **Was the “raw” data saved in addition to the preprocessed/cleaned data?** (e.g., to support unanticipated future uses)

n/a (not preprocessed)

## **Is the preprocessing software available?**

n/a (not preprocessed)

## **Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet?**

Yes, the lack of processing serves the purpose of creating a rich but imperfect dataset for educational and practice purposes.

# Dataset Distribution

## **How is the dataset distributed?** (e.g., website, API, etc.; does the data have a DOI; is it archived redundantly?)

The dataset can be downloaded from <https://cran.rstudio.com/web/packages/okcupiddata/index.html>

## **When will the dataset be released/first distributed?** (Is there a canonical paper/reference for this dataset?)

This dataset was published on August 19th, 2016. The canonical form https://CRAN.R-project.org/package=okcupiddata is used to link to the cran page.

## **What license (if any) is it distributed under?** Are there any copyrights on the data?

This data is licensed under CC0 1.0 Universal.

## **Are there any fees or access/export restrictions? Any other comments?**

None.

# Dataset Maintenance

## **Who is supporting/hosting/maintaining the dataset?**

## How does one contact the owner/curator/manager of the dataset (e.g. email address, or other contact info)?

The dataset is hosted at Middlebury College and all and comments can be sent to: Albert Y. Kim <albert.ys.kim@gmail.com>

## **Will the dataset be updated?** How often and by whom? How will updates/revisions be documented and communicated (e.g., mailing list, GitHub)? Is there an erratum?

This dataset is not planned to be updated. If the dataset is to be updated, the updates can be found via updated commits on <http://github.com/rudeboybert/okcupiddata>

## **If the dataset becomes obsolete how will this be communicated?**

Any changes to the dataset will be found via updated commits on <http://github.com/rudeboybert/okcupiddata>

## **Is there a repository to link to any/all papers/systems that use this dataset?**

There is no working repository to link any/all papers/systems that use this dataset.

## **If others want to extend/augment/build on this dataset, is there a mechanism for them to do so?** If so, is there a process for tracking/assessing the quality of those contributions. What is the process for communicating/distributing these contributions to users?

Unknown. The original publishers provide the python script (now broken due to OkCupid website redesign) originally used to scrape the data from OkCupid suggesting that the python script can be corrected and used to expand the dataset, but no process for assessing the quality of these additional contributions is mentioned.

# Legal and Ethical Considerations

**If the dataset relates to people (e.g., their attributes) or was generated by people, were they informed about the data collection?** (e.g., datasets that collect writing, photos, interactions, transactions, etc.)

No. This dataset was scraped from public OkCupid profiles which directly relates to people, their attributes, and sensitive psychographic information meant for other individuals looking for a relationship to view. The people were, however, never informed that their public data was collected and published for greater exposure than they intended.

**If it relates to other ethically protected subjects, have appropriate obligations been met?** (e.g., medical data might include information collected from animals)

n/a

**If it relates to people, were there any ethical review applications/reviews/approvals?** (e.g. Institutional Review Board applications)

From my research, I did not find any IRB applications.

**If it relates to people, were they told what the dataset would be used for and did they consent? What community norms exist for data collected from human communications?** If consent was obtained, how? Were the people provided with any mechanism to revoke their consent in the future or for certain uses?

No, the people were not informed that their data was collected and therefore did not know what the dataset would be used for or how it would be distributed.

**If it relates to people, could this dataset expose people to harm or legal action?** (e.g., financial social or otherwise) What was done to mitigate or reduce the potential for harm?

Yes, while any blatant identifiers such as profile name or location were omitted to maintain some anonymity, the distinguishing information such as body type, height, education, etc. could be easily used to cross reference with other public data to identify an individual and expose the individual’s potentially sensitive and personal answer submissions to the public.

**If it relates to people, does it unfairly advantage or dis- advantage a particular social group?** In what ways? How was this mitigated?

This dataset represents a higher percentage of ethnically Caucasian individuals as compared to the San Francisco population and therefore could disproportionately expose this population.

**If it relates to people, were they provided with privacy guarantees?** If so, what guarantees and how are these ensured?

No, no privacy guarantees (see answer to Legal and Ethical considerations question1).

**Does the dataset comply with the EU General Data Protection Regulation (GDPR)?** Does it comply with any other standards, such as the US Equal Employment Opportunity Act?

The dataset does not comply with GDPR because subjects were not asked for their consent.

**Does the dataset contain information that might be considered sensitive or confidential?** (e.g., personally identifying information)

Yes, distinguishing information such as body type, height, education, etc. could be easily used to cross reference with other public data to identify an individual and expose the individual’s potentially sensitive and personal answer submissions to the public.

## **Does the dataset contain information that might be considered inappropriate or offensive?**

Yes, answers to some of the open-ended essay prompts can be considered inappropriate or offensive as these answers were not monitored for inappropriate or offensive content.