Data Cleaning Documentation

# Description of Data Source

The dataset we will be using is the Open Sourcing Mental Illness dataset.

Open Sourcing Mental Illness, LTD (2016a). *OSMI* (Version 1) [Data file]. Retrieved September 7, 2018, from <https://osmihelp.org/research>

In 2016, a voluntary survey was conducted of those who worked in the technology industry (Open Sourcing Mental Illness, LTD, 2016b). The anonymous survey asks questions about how mental health is handled in the workplace as well as what are common mental health disorders in the technology industry (Open Sourcing Mental Illness, LTD, 2016b). A total of 1,433 complete responses are included in the dataset (Open Sourcing Mental Illness, LTD, 2016b). The survey contains responses from workers from across the world, although almost 60 percent of those surveyed worked in the United States at the time of the survey (Open Sourcing Mental Illness, LTD, 2016a).

[insert text here – needs to be 1-2 paragraphs, provide information about how much, where, and what it contains. There needs to be a citation of the data source]

# Description of Policy Constraints

The dataset has a Creative Commons Attribution-ShareAlike 4.0 International license (Open Sourcing Mental Illness, LTD, 2016b). Under this license, users can copy and distribute the material with no limitations on medium or format, as well as transform the dataset for any purpose, as long as it is distributed under the same license (Creative Commons, n.d.[a]). However, users must follow the terms of the license, which requires users to provide credit and a link to the license (Creative Commons, n.d.[a]). Credit is defined as providing the name of the dataset creator, a notice referring to the copyright, a notice referring to the license, a notice of the disclaimer provided in the license, and a link to the material (Creative Commons, n.d.[b]). However, the license indicates that these conditions may be satisfied “in any reasonable manner” based on the context the dataset is shared (the example provided is giving a link to a resource that provides this information) (Creative Commons, n.d.[b]). Since we are using this for a class project, it would be appropriate to provide a link to the source of the data and the data license, as this has the necessary information.

In addition, those who use the data must indicate if any changes were made to the dataset and cannot imply that the organization holding the license is endorsing the users or their product (Creative Commons, n.d.[a]). Furthermore, those who use the data cannot restrict others from doing things allowed under this license (Creative Commons, n.d.[a]).

[insert text here – I think this indicates limitations on how the data can be used]

# Description of Metadata

The zip file that comes with the dataset contains a file called “osmi-survey-2016\_1479139902,” which provides the available metadata for this dataset. First, the file contains the total number of responses they got, as well as the total “complete” entries which can be found in the data file (Open Sourcing Mental Illness, LTD, 2016b). The file also contains a paragraph providing information about what the survey will cover, how the survey will be anonymous, how long the survey will take to complete, and how the survey will be released and used (Open Sourcing Mental Illness, LTD, 2016b). This paragraph appears to be the information provided to volunteers that chose to reply to this survey. The file provides a list of all questions in the survey, and the “id” provides a clue about the type of question (e.g., yes/no, list, a number, a drop down) (Open Sourcing Mental Illness, LTD, 2016b). Beneath each question is a “field\_id” and it corresponds to the types of answers that were provided (Open Sourcing Mental Illness, LTD, 2016b).

[insert text here – needs to be 1 paragraph]

# Issues Encountered with Data

One of the issues our group has found with the data, is that the column names are very long, since it is the question asked in the survey. This would make analyzing the data difficult in R. Furthermore, we have noticed odd answers for several of the questions from the survey. Examples include:

* Inconsistent answers in the question about the respondent’s gender (e.g., some are “M” or “Mail”, some are in sentence format) (Open Sourcing Mental Illness, LTD, 2016b).
* In the question “How many employees does your organization have” some have 25-Jun and 5-Jan as the answer (it must have been a formatting error with the .csv file) (Open Sourcing Mental Illness, LTD, 2016b). Based off context clues from the other answers, they appear to be 6-25 and 1-5 (respectively) .

# Description of Steps Taking to Remediate

[insert text here – needs to be 1 paragraph]

# Data Cleaning Process

[insert text here – needs to be a script or step-by-step textual description that documents the data cleaning process]

# Works Cited

1. Open Sourcing Mental Illness, LTD (2016a). *OSMI* (Version 1) [Data file]. Retrieved September 7, 2018, from <https://osmihelp.org/research>
2. Open Sourcing Mental Illness, LTD (2016b). osmi-survey-2016\_1479139902 [Metadata File]. Retrieved September 19, 2018, from <https://osmihelp.org/research>.
3. Creative Commons (n.d.[a]), Attribution-ShareAlike 4.0 International [License]. Retrieved October 28, 2018 from <https://creativecommons.org/licenses/by-sa/4.0/legalcode>.
4. Creative Commons (n.d.[b]), Attribution-ShareAlike 4.0 International [Commons Deed]. Retrieved October 28, 2018 from https://creativecommons.org/licenses/by-sa/4.0/.