# Data Cleaning

## Description of Dataset

The data source we will be working from in our final group project is the Toxic Release Inventory Basic Data Files prepared by the Environmental Analysis Division and Toxics Release Information Branch of the Environmental Protection Agency (EPA). Industrial facilities that meet Toxics Release Inventory (TRI) Program requirements and the TRI Basic Data Files contain the reporting form data elements most frequently requested by TRI data users. This data includes the quantities of toxic chemicals released into the environment on site at facilities, the quantities transferred off site to other facilities, as well as summary data pertaining to the release of, recycling, energy recovery, and treatment of chemicals. There are four different types of TRI Basic Data Files: National Data File, State Data File, Federal Facility Data File, and Tribal Data File, starting with the year 1987 and going through 2014. For the purposes of our project, we will be using the National Data File which contains all the TRI data for the United States for a specific calendar year (2014). This data file includes data for all 50 states and the six U.S. Districts and territories, i.e., American Samoa, District of Columbia, Guam, Northern Mariana Islands, Puerto Rico, and the U.S. Virgin Islands. Each Basic Data File (in .csv format) contains 108 data fields, and generally consists of the follow categories: Year, Facility Name, Address, Latitude and Longitude Coordinates, NAICS code and Industry Sector and Codes, Chemical Identification, Classification Information, On-site Release Quantities, Units of Measure, Parent Companies, Publicly Owned Treatment Works (POTW) Transfer Quantities, Off-site Transfer Quantities for Release/Disposal and Further Waste Management, and Summary Pollution Prevention Quantities. EPA, U., & OIAA. (2016, September 21). TRI basic data files: Calendar years 1987 - 2015. Retrieved November 2, 2016, from US Environmental Protection Agency, https://www.epa.gov/toxics-release-inventory-tri-program/tri-basic-data-files-calendar-years-1987-2015

## Intellectual Property constraints and Licensing

With regards to intellectual policy constraints and other licensing impediments, the EPA, in connection with the U.S. government, stipulates that creative works, including writing, images, and computer code that are usually prepared by officers or employees of the U.S. government as part of their official duties, are generally not subject to copyright in the United States (under 17 U.S.C. § 105). Therefore, there is generally no copyright restriction on reproduction, derivative works, distribution, performance, or display of a government work. Unless the work falls under an exception, anyone may, without restriction under

U.S. copyright laws, reproduce the work in print or digital form, create derivative works, perform and display the work publicly, distribute copies or digitally transfer the work to the public by sale or other transfer of ownership, or by rental, lease, or lending. Information on EPA data licensing can be found here EPA data licensing information. Retrieved November 2, 2016, from https://edg.epa.gov/EPA_Data_License.html

## Description of the metadata

The metadata that is available to us is the TRI Basic Data File Format documentation v15 prepared by the EPA, used as an aid in understanding and interpreting the data in the data set. This field documentation represented as Appendix A: Record Layout and Appendix B: Chemical Classifications explicitly identifies the field names, definitions, sections, maximum length, data type, metals and their categories, for us to make sense of the different fields will be working with in the project. All 108 fields are individually named and defined in accordance with the TRI 2014 Data Set that we will be using. There are seven columns in the metadata layout format. The first column is the sequential field number identifier, identified by heading #. The second column, "Field", is the name of the field as it will appear in the data file. Third and fourth, are the maximum length and data type of field, respectively. For example, the notation "22, 7" in the maximum length field would mean that there are 22 digits in the number in the field, with 7 digits to the right of the decimal point. 'C' represents Character/Text data, while 'N' represents numeric data in the data type field. The fifth and sixth columns under the Form R reference heading indicate the "Part" and "Section" of the Form R and Form A. Lastly, the "Definition" column gives a description of each data element, provides notes about its origin and use, and tells us which data fields are added together to obtain totals in several of the data fields. In addition, this documentation gives useful information on how to account for zeroes in the data set.

## Rationale for the steps taken to remediate data

To answer the research questions for the dataset, sub setting of the dataset is done so that the attributes which are not related to the questions are removed. Attributes like BIA code (Code indicating the tribal land a facility is on), tribe, SIC (standard industrial code), SIC_2, SIC3, SIC4, SIC_5, SIC_6, Primary NAIC (Primary North American Industry Code System (NAICS) code that represents the facility's primary business activity.), Document Control number, CAS # / Compound ID, SRS_ID (The Substance Registry System Identification Number.) and Parent Company number are deleted.

Next the column of latitude and longitude are taken into consideration to find out if any data is misrepresented or is wrongly written. It's a fact that latitudes lie in range of $-90^0\ to\ 90^0$ and longitudes in range of $-180^0\ to\ 180^0$ any location with a different longitude and latitude can be disregarded as they don't represent accuracy of the plant location.

The next attribute to compare and validate are metal and metal category. Column Metal has two values (yes and no) and metal category has four values (category 1, 2, 3, 4). To validate if the data is correct we can compare the values of metal and metal category. If it's not a metal then metal category should correspond to 0, else if it's a metal then metal category will be 1, 2, 3 or 4.

The attribute unit of measurement has two values pounds and grams this leads to a disparity in the records. To avoid it we are converting all the records of those rows which have grams as unit of measurement into pounds. After converting the grams into pounds, we have realized that the effluent discharge is minimalistic hence those records can be ignored.

## Issues with the data

Firstly, the zip codes for the facilities have varied number of digits in it. Places with zip codes beginning with 0 are usually not correctly represented in excel until the cell values are formatted as text. For eg Essex County at Massachusetts has zip code of 01929, but until the zip code attribute in the dataset is formatted properly (converted into text) the zip code will be displayed as 1929 which is misleading in nature. Hence zip codes needs to be properly checked throughout the dataset matching it against the state and location.

As mentioned in the metadata file, since this dataset intends to be loaded in different statistical and analysis software they require the numerical fields to be populated with a number instead of blanks or NA hence in some places 0 are put. There are three reasons as to why numeric data fields are blank. Firstly, if a facility's chemical releases and other waste management quantities are below certain thresholds, the facility may submit a reporting Form A, which doesn't require any actual quantity data. For chemicals submitted on the Form A, all the quantity data is represented as zeroes. Secondly when facilities report NA for a quantity on Form R reporting, "NA" means that the release or waste management quantity is not possible for that facility. For example, if a facility is not located near a water body, it will not can release any of the chemical to water. The third case where zeroes appear instead of blanks occurs when facilities do not respond to quantity questions on the Form R, leaving them blank. This occurred most often prior to the TRI Electronic Reporting Rule, when the TRI Program still accepted the submission of paper reporting forms. The Web-based TRI-ME web application, however, doesn’t allow blanks in the reporting of quantity data; instead, the facility is required to enter a number or indicate “NA.” In the main data set it’s difficult to ascertain why a zero value is used.

Another issue is some records in the dataset are in grams while the majorities are in pounds. This creates disparity among records and their unit of measurements.

**A step-by-step description of data cleaning process.**

1. Recognizing redundant and non-required attributes

   We have a total of 108 columns, out of which there are some redundant columns as they provide the same information, and certain columns do not provide any value add to our research questions. We will remove the columns listed below and the reason for removing them will be classified as redundant or not required.

| Column | Reason | Column | Reason |
|---|---|---|---|
| PRIMARY_SIC | Redundant | NAICS_4 | Redundant |
| SIC_2 | Redundant | NAICS_5 | Redundant |
| SIC_3 | Redundant | NAICS_6 | Redundant |
| SIC_4 | Redundant | DOC_CTRL_NUM | Not Required |
| SIC_5 | Redundant | SRS_ID | Not Required |
| NAICS_2 | Redundant | PARENT_COMPANY_DB_NUMBER | Not Required |
| NAICS_3 | Redundant | | |

   (All the SIC_[x] and NAICS_[x] can be relate to the column PRIMARY_NAICS which provides the Industry code to identify the facility type)

2. Removal of rows with FORM_TYPE “A”

Companies which meet the following criteria can submit the shorter version of TRI Form i.e. Form A

a. The chemical being reported is NOT a PBT chemical

b. The chemical has not been manufactured, processed, or otherwise used in excess of 1,000,000 lbs

c. The total annual waste management (i.e., recycling, energy recovery, treatment, and disposal or other releases) of the chemical does not exceed 500 lbs

For such companies, all the columns representing waste generation, recycling, regeneration etc will only have zero values. Hence we can omit such rows. In our given dataset, we can filter out the column FORM_TYPE for value A, this yields 9435 rows which should be omitted. The presence of this data will not affect our result in anyway, since all the values are zero, but these rows provide no value add and are all redundant.

3. Correction of range of Latitude and Longitude

The range of latitudes is from -90 to 90 and for longitude is -180 to 180. Values which do not fit in this range are wrong and need to be omitted or corrected.

First we need to convert the LATITUDE and LONGITUDE columns to Number Values. To do so, select the entire column/s, Right Click and select Format Cells, Select the category ‘Number’ and put 4 in Decimal Places.

Then select the column again, click on Sort and Filter (under Editing Section of Home Toolbar) and select Filter. A filter dropdown button appears in the column header. Click on that and select the option Number filters and select Between… option. In the dialog box that opens, select the option “is greater than” and put the higher extremity value (90 for latitude and 180 for longitude), select the radio button “Or”, select the option “is lesser than” and put the lower extremity value (-90 for latitude and -180 for longitude).

We will obtain 2 such rows for LATITUDE, make the corresponding cells Blank, and we will obtain no rows for LONGITUDE.

The presence or absence of values outside the range bound does not affect our result set, as will not consider it for address based information, because we have information about the County.

4. Zip code validation

   Because of implicit conversion of String to Number, all preceding 0 values in a zip code get truncated. For example, “03501” converts to 3501. We need to identify such values and convert then back to the correct value. Zip codes have a length of 5 or 9 digits, hence we must convert the 3 or 4-digit zip code to a 5-digit zip code by adding preceding zeroes. To do so we need to select the entire column, right click and select the Format cells option. In the dialog box that opens, select Custom category, and enter 00000 in the Type input box and click on OK. This converts the 3 and/or 4-digit zip codes to 5-digit zip codes by adding preceding zeroes.

5. Resolving inconsistencies in Units of Measurement

   We have two units of measurement, Pounds and Grams under the column UNIT_OF_MEASURE. We need to filter the values for the UNIT_OF_MEASURE as Grams. To do so, select the column UNIT_OF_MEASURE and click on Filter option from the Sort and Filter Dropdown in the Editing section of the Home toolbar. Click on the Filter icon which appears in the column header, uncheck all options and check just the “Grams” option.

   We see that we have 1012 rows and the measured values are small. They are less than 0.1, hence they were measured in grams. The total across all columns adds up for random rows a range between 0.1g to 50g almost for the sample, which is equivalent to 0.11 pounds. Since remaining 74K plus records have direct measures in pounds and these records have such a minute value, we delete these rows.

Word Count: 2167